ENCYCLOPEDIA of SOLID EARTH GEOPHYSICS

Edited by
Harsh K. Gupta

Springer

ENCYCLOPEDIA of EARTH SCIENCES SERIES

ENCYCLOPEDIA *of* SOLID EARTH GEOPHYSICS

Encyclopedia of Earth Sciences Series

ENCYCLOPEDIA OF SOLID EARTH GEOPHYSICS

Aims of the Series
The *Encyclopedia of Earth Sciences Series* provides comprehensive and authoritative coverage of all the main areas in the Earth Sciences. Each volume comprises a focused and carefully chosen collection of contributions from leading names in the subject, with copious illustrations and reference lists.

These books represent one of the world's leading resources for the Earth Sciences community. Previous volumes are being updated and new works published so that the volumes will continue to be essential reading for all professional earth scientists, geologists, geophysicists, climatologists, and oceanographers as well as for teachers and students. See the dustjacket of this volume for a current list of titles in the *Encyclopedia of Earth Sciences Series*. Go to http://www.springerlink.com/reference-works/ to visit the "Earth Sciences Series" online.

About the Series Editor
Professor Charles W. Finkl has edited and/or contributed to more than eight volumes in the Encyclopedia of Earth Sciences Series. For the past 25 years he has been the Executive Director of the Coastal Education & Research Foundation and Editor-in-Chief of the international *Journal of Coastal Research*. In addition to these duties, he is Research Professor at Florida Atlantic University in Boca Raton, Florida, USA. He is a graduate of the University of Western Australia (Perth) and previously worked for a wholly owned Australian subsidiary of the International Nickel Company of Canada (INCO). During his career, he acquired field experience in Australia; the Caribbean; South America; SW Pacific islands; southern Africa; Western Europe; and the Pacific Northwest, Midwest, and Southeast USA.

Founding Series Editor
Professor Rhodes W. Fairbridge (deceased) has edited more than 24 Encyclopedias in the Earth Sciences Series. During his career he has worked as a petroleum geologist in the Middle East, been a WW II intelligence officer in the SW Pacific and led expeditions to the Sahara, Arctic Canada, Arctic Scandinavia, Brazil and New Guinea. He was Emeritus Professor of Geology at Columbia University and was affiliated with the Goddard Institute for Space Studies.

ENCYCLOPEDIA OF EARTH SCIENCES SERIES

ENCYCLOPEDIA *of* SOLID EARTH GEOPHYSICS

edited by

HARSH K. GUPTA
National Geophysical Research Institute
Council of Scientific and Industrial Research
Hyderabad
India

Springer

Library of Congress Control Number: 2011924208

ISBN: 978-90-481-8701-0
This publication is available also as:
Electronic publication under ISBN 978-90-481-8702-7 and
Print and electronic bundle under ISBN 978-90-481-8732-4

Published by Springer
P.O. Box 17, 3300 AA Dordrecht, The Netherlands

The original Encyclopedia of Solid Earth Geophysics was compiled by David E. James, and was first published in the Encyclopedia of Earth Sciences Series in 1989.

Printed on acid-free paper

Cover figure credit: Mineral Physics Institute at Stony Brook University, illustration by Keelin Murphy

Contents

Contributors

Geoffrey A. Abers
Lamont-Doherty Earth Observatory of Columbia University
61 Route 9W, P.O. Box 1000
Palisades, NY 10964
USA
abers@ldeo.columbia.edu

Antal Ádám
Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences
P.O. Box 5
9401 Sopron
Hungary
adam@ggki.hu

Wim Aerts
Department of Reference Systems and Planetology
Royal Observatory of Belgium
Avenue Circulaire 3
1180 Brussels
Belgium

Duncan Carr Agnew
Institute of Geophysics and Planetary Physics
Scripps Institution of Oceanography
University of California San Diego
9500 Gilman Dr
La Jolla, CA 92093-0225
USA
dagnew@ucsd.edu

Imtiaz Ahmed
BP
501 Westlake Park Blvd
Houston, TX
USA
Imtiaz.Ahmed@bp.com

Richard M. Allen
Berkeley Seismological Laboratory
Department of Earth and Planetary Science
University of California
307 McCone Hall
Berkeley, CA 94720
USA
rallen@berkeley.edu

Kusumita Arora
National Geophysical Research Institute
Council of Scientific & Industrial Research (CSIR)
Uppal Road
Hyderabad 500 606
India
kusumita@ngri.res.in
karora_ngri@yahoo.co.uk

Estella Atekwana
Boone Pickens School of Geology
Oklahoma State University
Stillwater, OK 74078
USA
estella.atekwana@okstate.edu

Jonathan M. Aurnou
Department of Earth and Space Sciences
University of California
Los Angeles, CA 90095
USA
aurnou@ucla.edu

E. V. S. S. K. Babu
National Geophysical Research Institute
Council of Scientific and Industrial Research
Uppal Road
Hyderabad 500 007
India
evsskbabu@yahoo.co.in

Ran Bachrach
Geophysics and Planetary Sciences Department
Tel Aviv University
P.O. Box 39040
Tel Aviv 69978
Israel
ran.bachrach@gmail.com

Ralph W. Baird
Baird Petrophysical International
Houston, TX
USA
ralph@bairdusa.com

Daniel N. Baker
Laboratory for Atmospheric and Space Physics
University of Colorado
1234 Innovation Drive
Boulder, CO 80303-7814
USA
daniel.baker@lasp.colorado.edu

Georges Balmino
Centre National d'Etudes Spatiales
18, Avenue Edouard Belin Cedex 4
31401 Toulouse
France
georges.balmino@orange.fr

Paramesh Banerjee
Earth Observatory of Singapore
Nanyang Technological University
50 Nanyang Avenue
639798
Singapore
pbanerjee@ntu.edu.sg

Jean-Pierre Barriot
Observatoire Géodésique de Tahiti
University of French Polynesia
BP 6570
98702 Faaa-Tahiti
French Polynesia
Jean-Pierre.Barriot@upf.pf

C. E. Barton
Australian National University
Canberra ACT 0200
Australia
cebarton@gmail.com

Graeme Beardsmore
Monash University
and
Hot Dry Rocks Pty Ltd
Victoria
Australia
graeme.beardsmore@hotdryrocks.com

Thorsten W. Becker
Department of Earth Sciences
University of Southern California
3651 Trousdale Pkwy, MC0740
Los Angeles, CA 90089-0740
USA
thorstinski@gmail.com

David Bercovici
Department of Geology & Geophysics
Yale University
P.O. Box 208109
New Haven, CT 06520-8109
USA
david.bercovici@yale.edu

Jean Besse
Laboratoire de Paléomagnétisme
Institut de Physique du Globe de Paris
4 place Jussieu
75252 Paris
France
besse@ipgp.fr
besse@ipgp.jussieu.fr

Y. J. Bhaskar Rao
National Geophysical Research Institute
Council of Scientific and Industrial Research
Uppal Road
Hyderabad 500 007
India
yjbhaskarrao@gmail.com

D. A. Bhaskara Rao
National Geophysical Research Institute
Hyderabad 500 007
India

Archana Bhattacharyya
Indian Institute of Geomagnetism
Kalamboli, New Panvel
Navi Mumbai 410218
India
abh@iigs.iigm.res.in

Jörg Bialas
Leibniz-Institut für Meereswissenschaften
University of Kiel
Wischhofstr. 1-3
24148 Kiel
Germany
jbialas@ifm-geomar.de

Richard Biancale
DTP, OMP
18 Avenue Edouard Belin
31401 Toulouse Cedex 9
France
richard.biancale@cnes.fr

Andrew Biggin
Geomagnetism Laboratory
University of Liverpool
Liverpool L69 7ZE
UK
biggin@liv.ac.uk

Joakim O. Blanch
BHP Billiton
1360 Post Oak Blvd. Ste 150
Houston, TX 77056
USA
joakim.blanch@bhpbilliton.com

Paul Bodin
Pacific Northwest Seismic Network (PNSN)
University of Washington
146 ATG Building
Seattle, WA 98195-1310
USA
bodin@uw.edu

Sigrid Böhm
Research Group Advanced Geodesy
Institute of Geodesy and Geophysics
Vienna University of Technology
Gusshausstrasse 27–29
1040 Vienna
Austria
sigrid.boehm@tuwien.ac.at

István Bondár
International Seismological Centre
Pipers Lane
Thatcham, Berkshire RG19 4NS
UK
istvan@isc.ac.uk

Pascal Bonnefond
Observatoire de la Côte d'Azur, Géoazur, GRGS
Avenue Nicolas Copernic
06130 Grasse
France

Peter Bormann
Department 2: Physics of the Earth
Formerly GFZ German Research Centre for Geosciences
Telegrafenberg
14473 Potsdam
Germany
pb65@gmx.net

Carla Braitenberg
Department of Geosciences
University of Trieste
Via Weiss 1
34100 Trieste
Italy
berg@units.it

Laurie Brown
Department of Geosciences
University of Massachusetts
Amherst, MA 01003
USA
lbrown@geo.umass.edu

Carine Bruyninx
Department of Reference Systems and Planetology
Royal Observatory of Belgium
Avenue Circulaire 3
1180 Brussels
Belgium
c.bruyninx@oma.be

Hans-Peter Bunge
Geophysics Section, Department of Earth and
Environmental Sciences
Ludwig Maximilians University of Munich
Theresienstrasse 41
80333 Munich
Germany
bunge@lmu.de

Evgueni Burov
ISTEP – Institute of Earth Sciences of Paris (UMR 7193
UPMC-CNRS)
University of Pierre et Marie Curie
75252 Paris
France
evgenii.burov@upmc.fr

Friedrich H. Busse
Institute of Physics
University of Bayreuth
Building NW II
95440 Bayreuth
Germany
busse@uni-bayreuth.de

Michel Campillo
Observatoire de Grenoble
Université Joseph Fourier and CNRS
BP 53
38041 Grenoble
France
Michel.Campillo@obs.ujf-grenoble.fr

David J. Carlson
British Antarctic Survey
IPY International Programme Office
High Cross, Madingley Road
Cambridge CB3 0ET
UK
ipy.djc@gmail.com

Richard W. Carlson
Department of Terrestrial Magnetism
Carnegie Institution of Washington
5241 Broad Branch Road, NW
Washington, DC 20015
USA
rcarlson@ciw.edu

Anny Cazenave
Laboratoire d'Etudes en Géophysique et Océanographie Spatiales (LEGOS)
Centre National d'Etudes Spatiales
18 Avenue Edouard Belin
31401 Toulouse Cedex 4
France
anny.cazenave@legos.obs-mip.fr

Vlastislav Červený
Department of Geophysics, Mathematics and Physics
Charles University
Ke Karlovu 3 121 16
Praha
Czech Republic
vcerveny@seis.karlov.mff.cuni.cz

R. K. Chadha
National Geophysical Research Institute
Uppal Road
Hyderabad 500007, Andhra Pradesh
India
chadha@ngri.res.in

V. Chakravarthi
National Geophysical Research Institute
Uppal Road
Hyderabad 500 007, Andra Pradesh
India
vcvarthi@rediffmail.com
vcvarthi@ngri.res.in

Jonathan E. Chambers
Geophysical Tomography Team
British Geological Survey, Keyworth
Nottingham NG12 5GG
UK
jecha@bgs.ac.uk

Aude Chambodut
Laboratoire de Géodynamique Globale: Champs de Potentiels et Déformations Actives (CNRS UMR 7516)
Ecole et Observatoire des Sciences de la Terre (UDS)
5 rue René Descartes
67084 Strasbourg Cédex
France
aude@unistra.fr

David S. Chapman
Department of Geology and Geophysics
The University of Utah
382 Frederick A Sutton Building, 115 South 1460 East
Salt Lake City, UT 84112-0102
USA
david.chapman@utah.edu

Nikolas I. Christensen
Department of Earth and Ocean Sciences
University of British Columbia
Vancouver, BC V6T 1Z4
Canada
chris@geology.wisc.edu

Ulrich R. Christensen
Max-Planck Institut für Sonnensystemforschung
Max-Planck Str. 2
37191 Katlenburg-Lindau
Germany
christensen@mps.mpg.de

Christoph Clauser
Institute for Applied Geophysics and Geothermal Energy
E.ON Energy Research Center
RWTH Aachen University
Mathieustr. 6
52074 Aachen
Germany
cclauser@eonerc.rwth-aachen.de

Clarence S. Clay
Marine Geophysics and Wave Propagation
Department of Geoscience
University of Wisconsin-Madison
1215 W. Dayton St.
Madison, WI 53706
USA
clay@geology.wisc.edu

Emily A. CoBabe-Ammann
Emily A. CoBabe and Associates
Boulder, CO
USA
ecobabe@spaceeducation.org

Steven Constable
Institute of Geophysics and Planetary Physics
Scripps Institution of Oceanography
La Jolla, CA 92093-0225
USA
sconstable@ucsd.edu

Michele Cooke
Geosciences Department
UMass – Amherst
611 North Pleasant St
Amherst, MA 01003-9297
USA
cooke@geo.umass.edu

Vernon F. Cormier
Physics Department
University of Connecticut
2152 Hillside Road U-3046
Storrs, CT 06269-3046
USA
vernon.cormier@uconn.edu

David Coulot
Laboratoire de Recherche en Géodésie
Institut Géographique National, GRGS
Ecole Nationale des Sciences Géographiques
6 et 8 Avenue Blaise Pascal
Cité Descartes – Champs-sur-Marne
77455 Marne la Vallée Cedex 2
France
David.Coulot@ign.fr

Vincent Courtillot
Institut de Physique du Globe de Paris
1 Place Jussieu
75005 Paris
France

Stuart Crampin
British Geological Survey
Murchison House
West Mains Road
Edinburgh EH9 3LA
Scotland, UK
scrampin@ed.ac.uk

Shamita Das
Department of Earth Sciences
University of Oxford
South Parks Road
Oxford OX1 3AN
UK
das@earth.ox.ac.uk

Michael G. Davis
Department of Geology and Geophysics
University of Utah
115 South 1460 East
Salt Lake City, UT 84112-0111
USA
michael.g.davis@utah.edu

Earl E. Davis
Pacific Geoscience Centre
Geological Survey of Canada
9860 W. Saanich Rd, P.O. Box 6000
Sidney, BC V8L 4B2
Canada
edavis@nrcan.gc.ca

Paul M. Davis
Earth and Space Sciences
UCLA
405 Hilgard Ave
Los Angeles, CA 90095
USA
pdavisucla@gmail.com

Bryan Davy
GNS Science
1 Fairway Drive Avalon, P.O. Box 30-368
5010 Lower Hutt
New Zealand

Eduardo F. J. de Mulder
NGU, Geological Survey of Norway
7491 Trondheim
Norway
e.demulder@planet.nl

Florent Deleflie
Institut de Mécanique Céleste et de Calcul des Ephémérides, GRGS
77 Avenue Denfert Rochereau
75014 Paris
France
Florent.Deleflie@imcce.fr

Mike Dentith
School of Earth and Environment (M004)
The University of Western Australia
35 Stirling Highway
Crawley, WA 6009
Australia
mdentith@segs.uwa.edu.au

John S. Derr
19 Escena Dr
Tijeras, NM 87059
USA
shakyace@yahoo.com

Domenico Di Giacomo
Department 2: Physics of the Earth
GFZ German Research Center for Geosciences
Telegrafenberg
14473 Potsdam
Germany
domenico@gfz-potsdam.de

Vijay P. Dimri
Research, Innovation & Incubation Center
Gujarat Energy Research and Management Institute
IT Tower-1, Infocity
Gandhinagar, Gujarat
India
dimrivp@yahoo.com
vijay.dimri@germi.res.in

Wolfgang Eder
Department of Earth Sciences
University of Munich
Luisenstrasse 37
80333 Munich
Germany
w.eder-geo@iaag.geo.uni-muenchen.de

Gary Egbert
College of Oceanic and Atmospheric Sciences
Oregon State University
104 COAS Administration Building
Corvallis, OR 97331-5503
USA
egbert@coas.oregonstate.edu

Eric Robert Engdahl
Center for Imaging the Earth's Interior
Department of Physics
University of Colorado at Boulder
Campus Box 390 UCB
Boulder, CO 80309-0390
USA
engdahl@colorado.edu

Pierre Exertier
Observatoire de la Côte d'Azur, Géoazur, GRGS
Avenue Nicolas Copernic
06130 Grasse
France

Cinzia G. Farnetani
Institut de Physique du Globe de Paris
75238 Paris
France
cinzia@ipgp.fr

Andrew T. Fisher
Earth and Planetary Sciences Department
University of California at Santa Cruz
Santa Cruz, CA 95064
USA
afisher@ucsc.edu

Ernst R. Flueh
Leibniz-Institut für Meereswissenschaften
University of Kiel
Wischhofstr. 1-3
24148 Kiel
Germany
eflueh@ifm-geomar.de

Luca Formaggia
MOX, Department of Mathematics
Politecnico di Milano
Piazza Leonardo da Vinci 32
20133 Milano
Italy
luca.formaggia@polimi.it

Donald W. Forsyth
Department of Geological Sciences
Brown University
Providence, RI 02912
USA
Donald_Forsyth@brown.edu

Alessandro M. Forte
GEOTOP Département des Sciences de la Terre
et de l'Atmosphère
Université du Québec à Montréal
CP 8888, succursale Centre-Ville
Montréal, QC H3C 3P8
Canada
forte60@gmail.com

Clive Foss
CSIRO
Locked Bag 10
Clayton, South VIC 3169
Australia
Clive.Foss@csiro.au

Peter Fox
Rensslear Polytechnic Institute
Troy, NY
USA
pfox@rpi.edu

Friedemann T. Freund
NASA Ames Research Center/San Jose State University
MS 242-4
Moffett Field, CA 94035-1000
USA
friedemann.t.freund@nasa.gov

Jeffrey T. Freymueller
Geophysical Institute
University of Alaska Fairbanks
903 Koyukuk Drive
Fairbanks, AK 99775
USA
jeff.freymueller@gi.alaska.edu

Masato Furuya
Department of Natural History Sciences
Hokkaido University
N10 W8, Kita-ku
Sapporo 060-0810
Japan
furuya@mail.sci.hokudai.ac.jp

Vineet Gahalaut
National Geophysical Research Institute (CSIR)
Uppal Road
Hyderabad 500 007
India
vkgahalaut@yahoo.com

Carmen Gaina
Geodynamics Group
Geological Survey of Norway
Trondheim
Norway

Kerry Gallagher
UMR 6118- Géosciences Rennes
Gêosciences, Université de Rennes 1
Campus de Beaulieu
35042 Rennes Cedex
France
kerry.gallagher@univ-rennes1.fr

Mengtan Gao
Institute of Geophysics
China Earthquake Administration
No. 5 Minzudaxue Nanlu
Haidian District, Beijing 100081
China
gaomt@vip.sina.com

Edward J. Garnero
School of Earth and Space Exploration
Arizona State University
Tempe, AZ 85287-1404
USA
garnero@asu.edu

Woldai Ghebreab
Department of Geology and Environmental Science
University of Akron
Akron, OH 44325-4101
USA
wg@uakron.edu

Hans-Jürgen Götze
Abteilung Geophysik
Institut für Geowissenschaften
Christian-Albrechts-Universität zu Kiel
Otto-Hahn-Platz 1
24118 Kiel
Germany
hajo@geophysik.uni-kiel.de

D. Gopala Rao
Geology Department
Osmania University
Hyderabad 500 007
India
drgopalarao@yahoo.com

Samuel H. Gray
CGGVeritas
715 5th Avenue SW
Calgary, AB T2P 5A2
Canada
sam.gray@cggveritas.com

Marianne Greff
Institut de Physique du Globe de Paris
1 Place Jussieu
75005 Paris
France

Richard A. F. Grieve
Earth Sciences Sector
Natural Resources Canada
Ottawa, ON K1A 0E8
Canada
rgrieve@nrcan.gc.ca

Gottfried Grünthal
Helmholtz Centre Potsdam
German Research Centre for Geosciences GFZ
Telegrafenberg
14473 Potsdam
Germany
ggrue@gfz-potsdam.de

Harsh K. Gupta
National Geophysical Research Institute
Council of Scientific & Industrial Research (CSIR)
Uppal Road
Hyderabad 500 606, Andhra Pradesh
India
harshg123@gmail.com
harsheseg@gmail.com

Pravin K. Gupta
Department of Earth Sciences
Indian Institute of Technology Roorkee
Roorkee, Uttarakhand 247 667
India
pkgptfes@iitr.ernet.in

Ron Hackney
Petroleum and Marine Division
Geoscience Australia
GPO Box 378
Canberra, ACT 2601
Australia
ron.hackney@ga.gov.au

Susan L. Halgedahl
Department of Geology and Geophysics
University of Utah
115 South 1460 East, Rm. 383
Salt Lake City, UT 84112-0111
USA
s.halgedahl@utah.edu

David J. Hamilton
Ateeda Limited
24 Canning Street
Edinburgh EH3 8EG
UK

Ulrich Harms
Scientific Drilling
GFZ German Research Centre for Geosciences
Telegrafenberg
14473 Potsdam
Germany
ulrich.harms@gfz-potsdam.de

Robert N. Harris
College of Oceanic and Atmospheric Sciences
Oregon State University
104 COAS Admin Bldg
Corvallis, OR 97331-5503
USA
rharris@coas.oregonstate.edu

Hayo Hase
Geodätisches Observatorium Wettzell
Bundesamt für Kartographie und Geodäsie
Sackenrieder Str. 25
93444 Bad Kötzting
Germany
hayo.hase@bkg.bund.de

Akira Hasegawa
RCPEV, Graduate School of Science
Tohoku University
Sendai, Miyagi 980-8578
Japan
hasegawa@aob.gp.tohoku.ac.jp

Derrick Hasterok
Department of Geology and Geophysics
The University of Utah
382 Frederick A Sutton Building, 115 South 1460 East
Salt Lake City, UT 84112-0102
USA
dhasterok@gmail.com

Stuart A. Henrys
GNS Science
1 Fairway Drive Avalon, P.O. Box 30-368
5010 Lower Hutt
New Zealand

John F. Hermance
Department of Geological Sciences
Brown University
P.O. Box 1846
Providence, RI 02912-1846
USA
John_Hermance@brown.edu

Richard N. Hey
Hawaii Institute of Geophysics and Planetology
School of Ocean and Earth Science and Technology
University of Hawaii
2525 Correa Road
Honolulu, HI 96822
USA
hey@soest.hawaii.edu

Jacques Hinderer
Institut de Physique du Globe de Strasbourg, CNRS
Université de Strasbourg
5 rue René Descartes
67084 Strasbourg
France
Jacques.Hinderer@unistra.fr

Klaus-Günter Hinzen
Earthquake Geology Group, Institute of Geology and Mineralogy
Cologne University
Vinzenz-Pallotti-Str. 26
51429 Bergisch Gladbach
Germany
hinzen@uni-koelnde

Albrecht W. Hofmann
Lamont-Doherty Earth Observatory
Columbia University
Palisades, NY 10964
USA
and
Max-Planck-Institut für Chemie
55020 Mainz
Germany
albrecht.hofmann@mpic.de

Ivan Hrvoic
GEM Systems, Inc
135 Spy Court
Markham, ON L3R 5H6
Canada
hrvoics@sympatico.ca

Andrew Hugill
2 Regal Rd Apt 709
Toronto, ON M6H 2J4
Canada
hugill.andrew@gmail.com

Giampiero Iaffaldano
Research School of Earth Sciences
The Australian National University
Bldg 61 Mills Road
Acton, ACT 2601
Australia
giampiero.iaffaldano@anu.edu.au

Anca Isac
Geological Institute of Romania
1 Caransebes Str., Sector 1
12271 Bucharest
Romania

Alik Ismail-Zadeh
Geophysical Institute
Karlsruhe Institute of Technology
Hertzstr. 16
76187 Karlsruhe
Germany
and
International Institute of Earthquake Prediction Theory
and Mathematical Geophysics
Russian Academy of Sciences
Profsoyuznaya 84/32
Moscow 117997
Russia
and
Institut de Physique du Globe
Paris
France
alik.ismail-zadeh@kit.edu

Ursula Iturrarán-Viveros
Departamento de Matemáticas, Facultad de Ciencias
Universidad Nacional Autónoma de México
Ciudad Universitaria, Circuito Escolar s/n
Coyoacán 04510, México DF
Mexico
uiturraran@yahoo.com

David D. Jackson
Department of Earth and Space Sciences
University of California Los Angeles
595 Charles E. Young Blvd
Los Angeles, CA 90095
USA
David.D.Jackson@UCLA.edu

David E. James
Department of Terrestrial Magnetism
Carnegie Institution of Washington
5241 Broad Branch Rd. N.W.
Washington, DC 20015
USA
james@dtm.ciw.edu

Claude Jaupart
Institut de Physique du Globe
4 Pl Jussieu
75252 Paris
France
claude.jaupart@upmc.fr

Christopher Jekeli
Division of Geodetic Science
School of Earth Sciences
Ohio State University
125 South Oval Mall
Columbus, OH 43210-1398
USA
jekeli.1@osu.edu

Mikhail K. Kaban
Department 1: Geodesy and Remote Sensing
Sektion 1.3: Earth-System Modelling
Deutsches GeoForschungsZentrum Potsdam - GFZ
Telegrafenberg A 17
14473 Potsdam
Germany
kaban@gfz-potsdam.de

Yan Y. Kagan
Department of Earth and Space Sciences
University of California Los Angeles
595 Charles E. Young Blvd
Los Angeles, CA 90095
USA

Masashi Kamogawa
Department of Physics
Tokyo Gakugei University
4-1-1, Nukuikitamachi
Koganei-shi, Tokyo 184-8501
Japan
kamogawa@u-gakugei.ac.jp

Teruyuki Kato
Earthquake Research Institute
The University of Tokyo
1-1, Yayoi 1
Bunkyo-ku, Tokyo 113-0032
Japan
teru@eri.u-tokyo.ac.jp

Hartmut Kern
Institut für Geowissenschaften
Universität Kiel
24098 Kiel
Germany
kern@min.uni-kiel.de

Andrzej Kijko
Aon-Benfield Natural Hazard Centre
University of Pretoria
Room 4-30, Mineral Sciences Building
0002 Pretoria
Republic of South Africa
andrzej.kijko@up.ac.za

Rainer Kind
Helmholtz Centre Potsdam
GFZ German Research Centre for Geosciences
Section 2.4, Seismology
Telegrafenberg A3
14473 Potsdam
Germany
kind@gfz-potsdam.de

Jonathan F. Kirby
Department of Spatial Sciences
Curtin University of Technology
GPO Box U1987
Perth, WA 6845
Australia
j.kirby@curtin.edu.au

Miroslav Kobr
Institute of Hydrogeology
Engineering Geology and Applied Geophysics
Charles University in Prague
Albertov 6
128 43 Prague 2
Czech Republic
kobr@natur.cuni.cz

Monika Korte
Sektion 2.3, Erdmagnetfeld
Helmholtz-Zentrum Potsdam
Deutsches GeoForschungsZentrum – GFZ
Telegrafenberg
14473 Potsdam
Germany
monika@gfz-potsdam.de

Vladimir Kossobokov
International Institute of Earthquake Prediction Theory
and Mathematical Geophysics
Russian Academy of Sciences
Profsoyuznaya 84/32
Moscow 117997
Russia
and
Institut de Physique du Globe
Paris
France
volodya@mitp.ru

K. S. Krishna
Geological Oceanography
National Institute of Oceanography (Council of Scientific
and Industrial Research)
Dona Paula, Goa 403004
India
Krishna@nio.org

Mikhail Kulesh
Information Technology
DIS AG
Rathenaustrasse 29A
15831 Mahlow
Germany
mikhail.kulesh@googlemail.com

Ota Kulhanek
Department of Earth Sciences, Section of Seismology
Uppsala University
Villavägen 16
752-36 Uppsala
Sweden
ota.kulhanek@geo.uu.se

Dhananjay Kumar
Chevron
1400 Smith, Room 42007
Houston, TX
USA
dhananjaykumar@gmail.com

Oliver Kuras
Geophysical Tomography Team
British Geological Survey, Keyworth
Nottingham NG12 5GG
UK
oku@bgs.ac.uk

Cristina La Mura
Department of Geosciences
University of Trieste
Via E. Weiss, 4 Pal. P
34127 Trieste
Italy
clamura@gmail.com

Carlo Laj
Laboratoire des Sciences du Climat
Unité mixte CEA-CNRS-UVSQ
Avenue de la Terrasse
91190 Gif-sur-Yvette
France
Carlo.Laj@lsce.ipsl.fr

Gurbax S. Lakhina
Indian Institute of Geomagnetism
Kalamboli Highway
New Panvel (W), Navi Mumbai
India
gslakhina@gmail.com

Carène Larmat
Geophysics Group, EES-17, MS D443
Los Alamos National Laboratory
Los Alamos, NM 87545
USA
carene@lanl.gov

Olivier Laurain
Observatoire de la Côte d'Azur, Géoazur, GRGS
Avenue Nicolas Copernic
06130 Grasse
France

Thorne Lay
Earth and Planetary Sciences Department
University of California
Earth and Marine Sciences Building
Santa Cruz, CA 95064
USA
tlay@ucsc.edu

William H. K. Lee
U.S. Geological Survey
MS 977, 345 Middlefield Road
Menlo Park, CA 94025
USA
lee@usgs.gov

Juliette Legrand
Department of Reference Systems and Planetology
Royal Observatory of Belgium
Avenue Circulaire 3
1180 Brussels
Belgium

Vincent Lesur
Helmholtz-Zentrum Potsdam
Deutsches GeoForschungsZentrum
14473 Potsdam
Germany

Enru Liu
China University of Mining and Technology
Xuzhou
China
eliu0103@hotmail.com

Mian Liu
Department of Geological Sciences
University of Missouri
Columbia, MO 65211
USA
lium@missouri.edu

Meng Heng Loke
Geotomo Software Sdn. Bhd.
115, Cangkat Minden Jalan 5, Minden Heights
11700 Gelugor, Penang
Malaysia
drmhloke@yahoo.com

William Lowrie
Emeritus Professor of Geophysics, Institute of Geophysics
Swiss Federal Institute of Technology
8092 Zürich
Switzerland
lowrie@mag.ig.erdw.ethz.ch

Susan Macmillan
British Geological Survey Murchison House
West Mains Road
Edinburgh EH9 3LA
Scotland, UK
smac@bgs.ac.uk

Raul Madariaga
Laboratoire de Géologie
Ecole Normale Supérieure
24 rue Lhomond
75231 Paris
France
madariag@geologie.ens.fr

Kumarendra Mallick
National Geophysical Research Institute
Uppal Road
Hyderabad 500 606, Andhra Pradesh
India
kumar.muse@yahoo.com

Mioara Mandea
Université Paris Diderot – Institut de Physique du Globe de Paris, Géophysique spatiale et planétaire
Bâtiment Lamarck Case 7,011, 5 rue Thomas Mann
75205 Paris Cedex 13
France
mioara@ipgp.fr

Ajay Manglik
National Geophysical Research Institute
Council of Scientific and Industrial Research (CSIR)
Uppal Road
Hyderabad 500 606
India
ajay@ngri.res.in

Maria-Daphne Mangriotis
Applied Geophysics
Institute of Engineering Seismology and Earthquake Engineering (ITSAK)
46 G. Scholis
5355102 Thessaloniki
Greece
mdmangriotis@gmail.com

Jean-Claude Mareschal
Centre GEOTOP-UQAM
University of Québec at Montréal
P.O. Box 8888
Montréal, QC H3C 3P8
Canada
mareschal.jean-claude@uqam.ca

Ludovic Margerin
Institut de Recherche en Astrophysique et Planétologie
Observatoire Midi-Pyrénées/C.N.R.S.
14, Avenue Edouard Belin
31400 Toulouse
France
margerin@dtp.obs-mip.fr

Yoshihisa Maruyama
Graduate School of Engineering
Chiba University
1-33 Yayoi-cho
Inage-ku, Chiba 263-8522
Japan
ymaruyam@tu.chiba-u.ac.jp

Robert McCaffrey
Department of Geology
Portland State University
Portland, OR 97207
USA
mccafr@gmail.com

James McClain
Department of Geology
University of California
One Shields Avenue
Davis, CA 95616-8605
USA
mcclain@geology.ucdavis.edu

Suzanne McEnroe
Norwegian Geological Survey
7491 Trondheim
Norway
Suzanne.McEnroe@ngu.no

Art McGarr
U.S. Geological Survey
345 Middlefield Road MS 977
Menlo Park, CA 94025-3591
USA
mcgarr@usgs.gov

Allen K. McNamara
School of Earth and Space Exploration
Arizona State University
Tempe, AZ 85287-1404
USA
allen.mcnamara@asu.edu

Rolf Meissner
Institut für Geowissenschaften
Universität Kiel
24098 Kiel
Germany
rmeissner@email.uni-kiel.de

Stelios P. Mertikas
Laboratory of Geodesy and Geomatics Engineering
Exploration and Positioning Division
Department of Mineral Resources Engineering
Technical University of Crete
73100 Chania Crete
Greece
mertikas@mred.tuc.gr
s_mertikas@yahoo.com

Uwe Meyer
Sub-Department Geophysical Reconnaissance – Resources and Near Surface Processes
Federal Agency for Geosciences and Mineral Resources
Stilleweg 2
30655 Hannover
Germany
Uwe.Meyer@bgr.de

Andrew J. Michael
U.S. Geological Survey
345 Middlefield Road, MS 977
Menlo Park, CA 94025
USA
michael@usgs.gov

Dinesh Chandra Mishra
National Geophysical Research Institute
Uppal Road
Hyderabad (CSIR) 500 606
India
dcm_ngri@yahoo.co.in

Jean-Paul Montagner
Seismological Laboratory, Institut de Physique du Globe
University Paris-Diderot, UMR CNRS/7154
1 rue Jussieu
75238 Paris
France
jpm@ipgp.fr

Paul Morgan
Colorado Geological Survey
1313 Sherman Street, Rm 715
Denver, CO 80203
USA
paul.morgan@state.co.us

Wim A. Mulder
Department of Geotechnology
Delft University of Technology
P.O. Box 5048
2600 GA Delft
The Netherlands
W.A.Mulder@tudelft.nl

Roger M. W. Musson
Earthquake Seismology
British Geological Survey
West Mains Road
Edinburgh EH9 3LA
Scotland, UK
rmwm@bgs.ac.uk

Toshiyasu Nagao
Earthquake Prediction Research Center
Institute of Oceanic Research and Development
Tokai University
3-20-1 Orido
Shimizu-ku, Shizuoka-shi, Shizuoka 424-8610
Japan
nagao@scc.u-tokai.ac.jp

Nandini Nagarajan
National Geophysical Research Institute
Uppal Road
Hyderabad 500007
India
nandini_nn@yahoo.com
nandini_ngri@yahoo.com

Shailesh R. Nayak
Ministry of Earth Sciences
Government of India
New Delhi 110003
India
secretary@moes.gov.in
shaileshnayak@hotmail.com

J. W. Neuberg
Institute of Geophysics and Tectonics
School of Earth and Environment
University of Leeds
Leeds LS2 9JT
UK
J.Neuberg@leeds.ac.uk

Guust Nolet
Geosciences Azur
250, Rue Albert Einstein
6560 Sophia Antipolis
France
nolet@geoazur.unice.fr

Axel Nothnagel
VLBI Research Group
Institute of Geodesy and Geoinformation
University of Bonn
Nussallee 17
53115 Bonn
Germany
nothnagel@uni-bonn.de

Emile A. Okal
Department of Earth & Planetary Sciences
Northwestern University
Evanston, IL 60208
USA
emile@earth.northwestern.edu

Gordon R. Osinski
Department of Earth Sciences/Physics and Astronomy
University of Western Ontario
London, ON N6A 5B7
Canada
gosinski@uwo.ca

Harald von der Osten-Woldenburg
National Heritage Department
Regional Government of Baden-Wuerttemberg
Berliner Strasse 12
73728 Esslingen am Neckar
Germany
harald.vonderosten@rps.bwl.de
hvdosten@gmail.com

Giuliano F. Panza
Department of Geosciences
University of Trieste
Via E. Weiss, 4 Pal. P
34127 Trieste
Italy
and
Earth System Physics Section/Sand Group
The Abdus Salam International Centre for Theoretical Physics
Trieste
Italy
panza@units.it

M. Parsons
National Snow and Ice Data Center
University of Colorado
Boulder, CO
USA
parsonsm@nsidc.org

Vincenzo Pasquale
Dipartimento per lo Studio del Territorio e delle sue Risorse, Settore di Geofisica
Università degli Studi di Genova
Viale Benedetto XV 5
16132 Genova
Italy
pasquale@dipteris.unige.it

Nikolaos K. Pavlis
Geodesy and Geophysics Basic and Applied Research
National Geospatial-Intelligence Agency (NGA)
12310 Sunrise Valley Drive, Mail Stop P-126
Reston, VA 20191-3449
USA
Nikolaos.K.Pavlis@nga.mil

Ingo A. Pecher
GNS Science
1 Fairway Drive, P.O. Box 30368
5040 Lower Hutt
New Zealand
and
Institute of Petroleum Engineering
Heriot-Watt University
Edinburgh EH14 4AS
UK
i.pecher@gns.cri.nz

W. R. Peltier
Department of Physics
University of Toronto
60 St. George Street
Toronto, ON M5S 1A7
Canada
peltier@atmosp.physics.utoronto.ca

Leif Persson
Department of Underwater Research
FOI
Gullfossgatan 6
164 90 Stockholm
Sweden
leifp@foi.se

William K. Peterson
Laboratory for Atmospheric and Space Physics
University of Colorado
1234 Innovation Drive
Boulder, CO 80303-7814
USA
pete@lasp.colorado.edu
Bill.Peterson@lasp.colorado.edu

J. D. A. Piper
Geomagnetism Laboratory
University of Liverpool
Liverpool L69 3BX
UK
sg04@liverpool.ac.uk

Donald R. Prothero
Department of Geology
Occidental College
1600 Campus Rd
Los Angeles, CA 90041-4858
USA
Prothero@oxy.edu

Ivan Pšenčík
Institute of Geophysics
Academy of Sciences of Czech Republic
Boční II 141 31
Praha
Czech Republic
ip@ig.cas.cz

Alfio Quarteroni
MOX, Department of Mathematics
Politecnico di Milano
Piazza Leonardo da Vinci 32
20133 Milano
Italy
and
CMCS–MATHICSE
EPFL
1015 Lausanne
Switzerland
alfio.quarteroni@epfl.ch

Tallavajhala Radhakrishna
Geoscience Division
Centre for Earth Science Studies
P.O. no. 7250, Akkulam, Thuruvikkal
Trivandrum 695031
India
tradha@rediffmail.com

V. Rama Murthy
Department of Earth and Planetary Sciences
Institute of Meteoritics, University of New Mexico
313 A Northrop Hall 1
Albuquerque, NM 87131-1126
USA
vrmurthy@unm.edu

G. Ramillien
DTP, OMP
18 Avenue Edouard Belin
31401 Toulouse Cedex 9
France
ramillien@dtp.obs-mip.fr

Sunita Rani
Department of Mathematics
Guru Jambheshwar University of Science and Technology
Hisar 125001
India
s_b_rani@rediffmail.com

Hojjatollah Ranjbar
Department of Mining Engineering
Shahid Bahonar University of Kerman
Jomhouri Eslami Blvd.
76135-133 Kerman
Iran
h.ranjbar@mail.uk.ac.ir

Dhananjay Ravat
Department of Earth & Environmental Sciences
University of Kentucky
101 Slone Building
Lexington, KY 40506-0053
USA
dhananjay.ravat@uky.edu

James W. Rector, III
Department of Civil and Environmental Engineering
University of California at Berkeley
412 Davis Hall
Berkeley, CA 94709
USA
jwrector@lbl.gov

J. N. Reddy
Department of Mechanical Engineering
Texas A&M University
3123 TAMU
College Station, TX 77843-3123
USA
jnreddytamu.edu
jn_reddy@yahoo.com

David A. Rhoades
GNS Science
1 Fairway Drive, Avalon
P.O. Box 30-368
5010 Lower Hutt
New Zealand
d.rhoades@gns.cri.nz

Paul G. Richards
Lamont-Doherty Earth Observatory
Columbia University
61 Route 9W, P.O. Box 1000
Palisades, NY 10964
USA
richards@LDEO.columbia.edu

Paul H. Roberts
Institute of Geophysics and Planetary Physics
University of California
Los Angeles, CA 90095-1567
USA
roberts@math.ucla.edu

Johan O. A. Robertsson
Schlumberger Cambridge Research
High Cross, Madingley Road
Cambridge CB3 0EL
UK
jrobertsson@slb.com

Fabio Romanelli
Department of Geosciences
University of Trieste
Via E. Weiss, 4 Pal. P
34127 Trieste
Italy
romanel@units.it

Barbara Romanowicz
Berkeley Seismological Laboratory
215 McCone Hall
Berkeley, CA 94720
USA
barbara.romanowicz@gmail.com

Philippe Roux
Observatoire de Grenoble
Université Joseph Fourier and CNRS
BP 53
38041 Grenoble
France

Kabir Roy Chowdhury
Department of Earth Sciences
Utrecht University
Budapestlaan 4
3584 Utrecht
The Netherlands
kabir@geo.uu.nl

Ladislaus Rybach
Geowatt AG Zurich
Dohlenweg 28
8050 Zurich
Switzerland
rybach@geowatt.ch

Leonardo Sagnotti
Istituto Nazionale di Geofisica e Vulcanologia
Via di Vigna Murata 605
00143 Rome
Italy
leonardo.sagnotti@ingv.it

Kalachand Sain
National Geophysical Research Institute
Council of Scientific & Industrial Research (CSIR)
Uppal Road
Hyderabad 500 606
India
kalachandsain@yahoo.com

Bertrand de Saint-Jean
Observatoire de la Côte d'Azur, Géoazur, GRGS
Avenue Nicolas Copernic
06130 Grasse
France

Mahmoud Mohamed Selim Saleh
Department of Nature and Applied Sciences
Al-Aflaj Community College
AL-Kharj University
710-11912 Al-Aflaj, Riyadh
Saudi Arabia
msalim@ksu.edu.sa
selim23@yahoo.com

Malcolm Sambridge
Seismology and Mathematical Geophysics
Research School of Earth Sciences
The Australian National University
Canberra, ACT 0200
Australia
Malcolm.Sambridge@anu.edu.au

Francisco J. Sánchez-Sesma
Instituto de Ingeniería
Universidad Nacional Autónoma de México
Ciudad Universitaria, Circuito Escolar s/n
Coyoacán 04510, México DF
Mexico
sesma@servidor.unam.mx

William A. Sandham
Scotsig
40 Westbourne Gardens
Glasgow G12 9PF
UK
w.sandham@scotsig.co.uk

John H. Sass
Great Lakes Geothermal Services
P.O. Box 610208
Port Huron, MI 48061
USA
jhsass@hay.net

Magdalena Scheck-Wenderoth
Sect. 4.4: Basin Analysis
Helmholtz Centre Potsdam
GFZ German Research Centre for Geosciences
Telegrafenberg C4
14473 Potsdam
Germany
leni@gfz-potsdam.de

Harald Schuh
Research Group Advanced Geodesy
Institute of Geodesy and Geophysics
Vienna University of Technology
Gusshausstrasse 27–29
1040 Vienna
Austria
harald.schuh@tuwien.ac.at

Gerard T. Schuster
Division of Environmental and Earth Sciences
King Abdullah University of Science and Technology
Thule
Saudi Arabia
gerard.schuster@kaust.edu.sa

Martha Schwartz
Department of Earth Sciences
University of Southern California
3651 Trousdale Parkway ZHS 117
Los Angeles, CA 90089
USA
mischwar@usc.edu

Johannes Schweitzer
NORSAR
P.O. Box 53
2027 Kjeller
Norway
johannes@norsar.no

Mrinal K. Sen
Department of Geological Sciences
UT Institute for Geophysics, Jackson School of Geosciences
The University of Texas at Austin
Pickle Research Campus, 10100 Burnet Road, Bldg 196
Austin, TX 78758
USA
Mrinal@ig.utexas.edu

A. M. Celâl Şengör
Eurasia Earth Science Institute
Istanbul Technical University
34469 Istanbul
Turkey
sengor@itu.edu.tr

Nikolai M. Shapiro
Institut de Physique du Globe de Paris
4 place Jussieu
75252 Paris
France

Krishna Kant Sharma
Department of Applied Geology
University of Madras
Chennai 600 025, Tamilnadu
India
kksharma25@hotmail.com

John Shaw
Geomagnetism Laboratory
University of Liverpool
Liverpool L69 7ZE
UK
shaw@liv.ac.uk

Lydie Sichoix
Observatoire Géodésique de Tahiti
University of French Polynesia
BP 6570
98702 Faaa-Tahiti
French Polynesia
lydie.sichoix@upf.pf

Michael G. Sideris
Department of Geomatics Engineering
University of Calgary
2500 University Drive NW
Calgary, AB T2N 1N4
Canada
sideris@ucalgary.ca

Kumar Hemant Singh
NASA Goddard Space Flight Center
Planetary Geodynamics Laboratory
Greenbelt, MD 20771
USA
and
Goddard Earth Sciences and Technology Center
University of Maryland Baltimore County
5523 Research Park Drive, Suite 320
Baltimore, MD 21228
USA
Kumar.H.Singh@nasa.gov

Sarva Jit Singh
Indian National Science Academy
New Delhi 110002
India
s_j_singh@yahoo.com

Satish C. Singh
Laboratoire de Géoscience Marines
Institut de Physique du Globe de Paris
1 rue Jussieu, Case 89
75238 Paris
France
singh@ipgp.fr

Lee Slater
Department of Earth and Environmental Sciences
Rutgers University
101 Warren St., Smith 136
Newark, NJ 07102
USA
lslater@andromeda.rutgers.edu

Alan G. Smith
Department of Earth Sciences
University of Cambridge
Sedgwick Museum, Downing Street
Cambridge CB2 3EQ, CB1 2AX
UK
ags1@cam.ac.uk

Gennady Sobolev
Institute of Physics of the Earth
Russian Academy of Sciences
B. Gruzinskaya ul., 10
Moscow 123810
Russia
sobolev@ifz.ru

Semere Solomon
Det Norske Veritas
DNV Research and Innovation
1322 Høvik
Norway
Semere.Solomon.Foto@dnv.com

Xiaodong Song
Department of Geology
University of Illinois at Urbana-Champaign
1301 W. Green St.
Urbana, IL 61801
USA
xsong@illinois.edu

Ravi P. Srivastava
Fractals in Geophysics
National Geophysical Research Institute
Uppal Road
Hyderabad 500 007
India
ravi_prakash@ngri.res.in

Frank D. Stacey
Division of Exploration and Mining
CSIRO
P.O. Box 883
Kenmore 4069
Australia
Frank.Stacey@csiro.au

Vaughan Stagpoole
GNS Science
1 Fairway Drive Avalon, P.O. Box 30-368
5010 Lower Hutt
New Zealand

Seth Stein
Department of Earth and Planetary Sciences
Northwestern University
Evanston, IL 60208
USA
seth@earth.northwestern.edu

France St-Laurent
125, 68e Avenue
LaSalle, QC H8P 3H1
Canada
france.st-laurent@sympatico.ca

Paul L. Stoffa
Department of Geological Sciences, UT Institute for Geophysics, Jackson School of Geosciences
The University of Texas at Austin
Pickle Research Campus, 10100 Burnet Road, Bldg 196
Austin, TX 78758
USA

Dmitry A. Storchak
International Seismological Centre (ISC)
Pipers Lane
Thatcham, Berkshire RG19 4NS
UK
dmitry@isc.ac.uk

Peter Suhadolc
Department of Geosciences
University of Trieste
Via Weiss, 2
Trieste
Italy
suhadolc@units.it

Neil Suttie
Geomagnetism Laboratory
University of Liverpool
Liverpool L69 7ZE
UK
suttie@liv.ac.uk

Christopher J. Swain
Department of Spatial Sciences
Curtin University of Technology
GPO Box U1987
Perth, WA 6845
Australia
c_swain@wt.com.au

László Szarka
Geodetic and Geophysical Research
Institute of the Hungarian Academy of Sciences
P.O. Box 5
9401 Sopron
Hungary

Donald H. Tarling
School of Earth, Ocean and Environmental Sciences
University of Plymouth
Drake Circus
Plymouth PL4 8AA
UK
D.Tarling@Plymouth.ac.uk

Erwan Thébault
Équipe de Géomagnétisme
Institut de Physique du Globe de Paris
1 rue Jussieu
75005 Paris Cedex
France
ethebault@ipgp.fr

Robert Thériault
Quebec Ministry of Natural Resources and Wildlife
5700, 4ème Avenue Ouest, Bureau A-401
Québec, QC G1H 6R1
Canada
robert.theriault@mrnf.gouv.qc.ca

Clifford H. Thurber
Department of Geoscience
University of Wisconsin-Madison
1215 W. Dayton St.
Madison, WI 53706-1672
USA
thurber@geology.wisc.edu

Virendra M. Tiwari
Gravity and Magnetic Studies Group
National Geophysical Research Institute (CSIR)
Uppal Road
Hyderabad 500 606, Andhra Pradesh
India
vmtiwari@yahoo.com
vmtiwari@ngri.res.in

Harold J. Tobin
Department of Geoscience
University of Wisconsin-Madison
1215 W. Dayton St.
Madison, WI 53706
USA
htobin@wisc.edu

Shinji Toda
Disaster Prevention Research Institute
Kyoto University
Gokasho
Uji, Kyoto 611-0011
Japan
toda@rcep.dpri.kyoto-u.ac.jp

Tamás Tóth
Geomega Limited
Mester u. 4
1095 Budapest
Hungary
info@geomega.hu

Bruce T. Tsurutani
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA
USA
Bruce.T.Tsurutani@jpl.nasa.gov

Srinivasa Kumar Tummala
Indian National Centre for Ocean Information Services
Hyderabad 500055
India
srinivas@incois.gov.in

James A. Tyburczy
School of Earth and Space Exploration
Arizona State University
Tempe, AZ 85287-1404
USA
jim.tyburczy@asu.edu

Seiya Uyeda
Japan Academy
7-32, Ueno Park
Taito-ku, Tokyo 110-0007
Japan
suyeda@st.rim.or.jp

Franco Vaccari
Department of Geosciences
University of Trieste
Via E. Weiss, 4 Pal. P
34127 Trieste
Italy
vaccari@units.it

Petr Vaníček
Department of Geodesy and Geomatics Engineering
University of New Brunswick
P.O. Box 4400
Fredericton E3B 5A3
Canada
vanicek@unb.ca

Anthwar Vasanthi
National Geophysical Research Institute
Uppal Road
Hyderabad 500 606, Andhra Pradesh
India
vasanti_kulkarni@yahoo.com

Nimisha Vedanti
Fractals in Geophysics
National Geophysical Research Institute
Uppal Road
Hyderabad 500 007
India
nimisha@ngri.res.in

Lev P. Vinnik
Institute of Physics of the Earth
Bol. Gruzinskaya str., 10
GSP Moscow 123995
Russia
vinnik@ifz.ru

John M. Wahr
Department of Physics and Cooperative Institute for Research in Environmental Sciences
University of Colorado
Campus Box 390
Boulder, CO 80309-0390
USA
john.wahr@gmail.com

David J. Wald
U.S. Geological Survey
Denver Federal Center
P.O. Box 25046
MS966 Lakewood, CO 80225
USA
wald@usgs.gov

Yanghua Wang
Department of Earth Science and Engineering
Centre for Reservoir Geophysics
Imperial College London
London SW7 2BP
UK
yanghua.wang@imperial.ac.uk

Steven N. Ward
Institute of Geophysics and Planetary Physics
University of California at Santa Cruz
Santa Cruz, CA 95064
USA
ward@es.ucsc.edu
wardsn@ucsc.du

Anthony B. Watts
Department of Earth Sciences
University of Oxford
Parks Road
Oxford OX1 3PR
UK
Tony.Watts@earth.ox.ac.uk

Lianxing Wen
Department of Geosciences
State University of New York, at Stony Brook
Stony Brook, NY 11794
USA
Lianxing.Wen@sunysb.edu

Helmut Wiesemeyer
Millimeter and Submillimeter Astronomy
Max-Planck-Institute for Radio Astronomy
Auf dem Hügel 61
53121 Bonn
Germany
hwiese@mpifr-bonn.mpg.de

Ray Wood
GNS Science
1 Fairway Drive Avalon, P.O. Box 30-368
5010 Lower Hutt
New Zealand
r.wood@gns.cri.nz

Ian Wright
National Oceanography Centre
European Way
Southampton SO14 3ZH
UK

Guangyin Xu
Institute of Geophysics, China Earthquake Administration
No. 5 Minzudaxue Nanlu
Haidian District, Beijing 100081
China
xugy1971@163.com

Fumio Yamazaki
Graduate School of Engineering
Chiba University
1-33 Yayoi-cho
Inage-ku, Chiba 263-8522
Japan
yamazaki@tu.chiba-u.ac.jp

Takashi Yoshino
Institute for Study of the Earth's Interior
Okayama University
827 Yamada
Misasa, Tottori 682-0193
Japan
tyoshino@misasa.okayama-u.ac.jp

Yanxiang Yu
Institute of Geophysics, China Earthquake Administration
No. 5 Minzudaxue Nanlu
Haidian District, Beijing 100081
China
yuyx@cea-igp.ac.cn

Xiaohui Yuan
Section Seismology
GFZ German Research Centre for Geosciences
Telegrafenberg
14473 Potsdam
Germany
yuan@gfz-potsdam.de

Wu Zhongliang
Institute of Geophysics
China Earthquake Administration
No. 5, Minzudaxuenanlu
Beijing 100081
China
wuzl@cea-igp.ac.cn

Colin A. Zelt
Department of Earth Science
Rice University
MS-126, P.O. Box 1892
Houston, TX 77251-1892
USA
czelt@rice.edu

Mark A. Zumberge
Institute of Geophysics and Planetary Physics
Scripps Institution of Oceanography
University of California
La Jolla, San Diego, CA 92093-0225
USA
mzumberge@ucsd.edu

Preface

All information about the Earth's interior comes from field observations and measurements made within the top few kilometers of the surface, from laboratory experiments and from the powers of human deduction, relying on complex numerical modeling. Solid Earth Geophysics encompasses all these endeavors and aspires to define and quantify the internal structure and processes of the Earth in terms of the principles of physics, corresponding mathematical formulations and computational procedures. The role of Solid Earth Geophysics has gained prominence with increasing recognition of the fact that knowledge and understanding of Earth processes are central to the continued well being of the global community. Apart from persistent search for natural resources, this research line is linked to basic investigations regarding the mutual relationships between climate and tectonics and on the effects of global change in terms of a wide spectrum of natural hazards. Consequently, the pursuit of this science has seen spectacular progress all over the world in recent decades, both in fundamental and applied aspects, necessarily aided by advancements in allied fields of science and technology.

The ***Encyclopedia of Solid Earth Geophysics***, aims to serve as a comprehensive compendium of information on important topics of Solid Earth Geophysics and provide a systematic and up-to-date coverage of its important aspects including primary concepts as well as key topics of interest. It, however, does not claim to chronicle each and every niche area that in reality is a part of this multidisciplinary and multi-faceted science. Neither does it attempt to describe the basic physics of matter and energy systems, which comprise the underlying tenets of geophysical research. The first edition of this Encyclopedia, edited by Prof. David James, was published in 1989 by the Van Nostrand Reinhold publishing company. The extraordinary growth and diversification of this science over the last twenty years called for a complete revision. This is realized by identifying the necessary topics and bringing together over 200 articles covering established and new concepts of Geophysics across the sub-disciplines such as Gravity, Geodesy, Geoelectricity, Geomagnetism, Seismology, Seismics, Deep Earth Interior and Processes, Plate Tectonics, Geothermics, Computational Methods, etc. in a consistent format. Exceptional Exploration Geophysics and Geotechnical Engineering topics are included for the sake of completeness. Topics pertaining to near Earth environs, other than the classical 'Solid Earth', are not within the scope of this volume as it is felt that the growth of knowledge in these fields justify a dedicated volume to cover them.

Articles written by leading experts intend to provide a holistic treatment of Solid Earth Geophysics and guide researchers to more detailed sources of knowledge should they require them. As basic understanding and application of Solid Earth Geophysics is essential for professionals of many allied disciplines such as Civil Engineering; Environmental Sciences; Mining, Exploration and software industries; NGOs working on large scale social agenda; etc., it would be useful to them to have access to a ready and up-to-date source of knowledge on key topics of Solid Earth Geophysics. Hopefully, this Encyclopedia would prove to be an authoritative and current reference source with extraordinary width of scope, drawing its unique strength from the expert contributions of editors and authors across the globe.

I am grateful to Anny Cazenave, Kusumita Arora, Bob Engdahl, Seiya Uyeda, Rainer Kind, Ajay Manglik, Kalachand Sain and Sukanta Roy, members of the Editorial Board for their constant advice and guidance in developing the framework of this Encyclopedia and help with the editorial work. I am equally grateful to all the authors who readily agreed to contribute and honoured the guidelines and time schedule.

Petra van Steenbergen, Sylvia Blago, Simone Giesler and D. Nishantini from Springer were very co-operative. It has been a pleasure working with Springer. Ms M. Uma Anuradha provided extraordinary assistance in the preparation of this volume. My wife Manju and daughters Nidhi & Benu supported me through the entire project.

Harsh K. Gupta
(Editor-in-Chief)

Acknowledgments

The manuscripts went through several iterations during their development and at each stage the manuscripts were carefully reviewed by well known experts in their respective fields. The following have graciously spent their valuable time in reviewing one or more articles:

Rachel Abercrombie, USA; Nick Ambraseys, UK; John Anderson, USA; Kusumita Arora, India; Vladislav Babuska, Czech Republic; Georges Balmino, France; Paramesh Banerjee, Singapore; Francois Barlier, France; Jean-Pierre Barriot, France; David Bercovici, USA; Eric Bergman, USA; Richard Biancale, France; Francesca Bianco, Italy; Sue Bilek, USA; Thomas Bleier, USA; Chaiwoot Boonyasiriwat, USA; Michael Bostock, Canada; C. Boucher, France; Michel Bouchon, France; Carla Braitenberg, Italy; Kevin Burke, USA; Rhett Butler, USA; Anny Cazenave, France; David S. Chapman, USA; George Choy, USA; Christoph Clauser, Germany; Steve Constable, USA; Chris Cramer, USA; Shamita Das, USA; James Dewey, USA; Michel Diament, France; Mathieu Dumberry, USA; Utpal Dutta, USA; Eric Robert Engdahl, USA; Cinzia G. Farnetani, France; Donald W. Forsyth, USA; Alessandro Forte, Canada; C. Mary R. Fowler, UK; Olivier Francis, Luxembourg; Bernhard Friedrich, Germany; Yoshio Fukao, Japan; Vineet Gahalaut, India; Fuchun Gao, USA; Wolfram Geissler, Germany; Louis Geli, France; Harsh K.Gupta, India; Pravin K. Gupta, India; Winfried Hanka, Germany; Yasushi Harada, Japan; Bob Hardage, USA; Robert N. Harris, USA; Derrick Hasterok, USA; Stuart Henrys, New Zealand; I.V. Radhakrishnamurthy, India; Heiner Igel, Germany; Kojiro Irikura, Japan; Alik Ismail-Zadeh, Germany; Andrew Jackson, Switzerland; Sridevi Jade, India; Thomas Jahr, Germany; Masashi Kamogawa, Japan; Brian Kennett, Australia; Rainer Kind, Germany; Shuichi Kodaira, Japan; Naoji Koizumi, Japan; Monika Korte, Germany; Ivan Koulakov, Russia; H. Kumagai, Japan; Prakash Kumar, India; Kanani Lee, USA; Joachim Linthe, Germany; William Lowrie, Switzerland; M. Venkateshwarlu, India; M. Ravi Kumar, India; Kumarendra Mallick, India; Subhashish Mallick, USA; Mioara Mandea, France; Ajay Manglik, India; Jean-Claude Mareschal, France; Didier Massonet, France; T. Mtsuzawa, Japan; Walter D. Mooney, USA; Daniel Moos, USA; Roger Musson, UK; Steve Myers, USA; N. Purnachandra Rao, India; Toshi Nagao, Japan; Nandini Nagarajan, India; S.W.A. Naqvi, India; Guust Nolet, France; Maheswar Ojha, India; Bob Pawlowski, USA; Mita Rajaram, India; Gilles Ramstein, France; M.V.M.S. Rao, India; V. Vijaya Rao, India; Paul G. Richards, USA; Paul Roberts, USA; Barbara Romanowicz, USA; Sukanta Roy, India; Kalachand Sain, India; Martha Savage, New Zealand; M. M. Selim, Egypt; Peter Shearer, USA; Anne Sheehan, USA; Satish C. Singh, France; Vipin K. Srivastava, India; Peter Suhadolc, Italy; Kiyoshi Suyehiro, Japan; Matthieu Sylvander, France; Pradeep Talwani, USA; John Tarduno, USA; Virendra M. Tiwari, India; R.K. Tiwari, India; C.C. Tscherning, Denmark; Shin'ya Tsukada, Japan; Seiya Uyeda, Japan; Chakravarthi V.C., India; David Wald, USA; John T. Weaver, USA; Maximilian Werner, Switzerland; Rudolf Widmer-Schnidrig, Germany; Max Wyss, USA; Robert S. Yeats, USA; Yongxian Zhang, China.

A

ABSOLUTE AGE DETERMINATIONS: RADIOMETRIC

Richard W. Carlson
Department of Terrestrial Magnetism, Carnegie Institution of Washington, Washington, DC, USA

Definition

Radiometric dating uses the decay of naturally occurring radioactive elements to determine the absolute age of geologic events.

Introduction

Time is an essential datum in geology. Some geological processes, particularly those responsible for generating natural hazards, occur over time intervals that can be measured with conventional clocks. Most geologic processes, however, occur on timescales beyond human experience – thousands, millions, and even billions of years (see *Continental Drift*; *Continental Rifts*; *Geodynamics*; *Lithosphere, Continental*; *Paleomagnetism, Magnetostratigraphy*; *Paleoseismology*; *Plate Tectonics, Precambrian*). For these, the chronometer of choice is radiometric dating, where the decay of naturally occurring radioactive elements is translated into time (Dickin, 2005; Faure and Mensing, 2005). In its application to geochronology, the most important aspect of radioactive decay is that it occurs at a known, constant rate, independent of environmental factors such as changing temperature and pressure, at least within the ranges of these parameters found outside the interior of stars.

Only 10 years after the discovery of radioactivity, the renowned physicist Ernest Rutherford measured the amount of helium in a uranium ore and derived an age of 500 million years for the ore (Rutherford, 1906). By 1913, radioactive dating of rocks had made important inroads into geologic investigations of Earth history (Holmes, 1913). The discovery of isotopes, also in 1913, and the improvement of the instruments used to measure their abundance (mass spectrometers) over the next decades, allowed radioactive dating to be applied to an increasingly broad range of natural materials, and hence, geologic processes. This, in turn, facilitated the transition of isotopic measurements from physics to geology departments and established radiometric geochronology as a tool of widespread use in the geosciences in the years following World War II.

Radioactivity and the systematics of its use as a chronometer

Radiometric dating is based on the principle of nuclear transmutation, common to alchemy. Some elements have isotopes whose nuclei contain an unstable number of neutrons and protons. The instability is most often remedied by the ejection of material from the nucleus. Alpha decay involves the ejection of two protons and two neutrons (a 4-helium (^{4}He) nucleus or alpha particle). Beta decay occurs through loss of an electron from the nucleus, turning a neutron into a proton. Another form of radioactive decay involves the capture of an electron from those surrounding the nucleus, turning a proton into a neutron. All forms of radioactive decay also release energy, and are thus a heat source in Earth's interior (see *Radiogenic Heat Production of Rocks*; *Energy Budget of the Earth*). The result of radioactive decay is that an isotope of one element is transformed into an isotope of another element. While the path from lead to gold does not occur naturally, if one waits long enough, radioactive decay will eventually transform all uranium and thorium into lead.

Table 1 lists the major radioactive nuclides that have been used for dating various geologic events. Naturally

Harsh K. Gupta (ed.), *Encyclopedia of Solid Earth Geophysics*, DOI 10.1007/978-90-481-8702-7,

Absolute Age Determinations: Radiometric, Table 1 Radioactive elements commonly used for absolute age determinations

Parent isotope	Production mechanism	Daughter isotope	Half-Life (Million years)
7-Berylium (^{7}Be)	Cosmogenic	7-Lithium (^{7}Li)	53 days
210-Lead (^{210}Pb)	Uranium decay	210-Bismuth (^{210}Bi)	22.3 years
226-Radium (^{226}Ra)	Uranium decay	222-Radon (^{222}Rn), ^{4}He	1,622 years
14-Carbon (^{14}C)	Cosmogenic, Bomb	14-Nitrogen (^{14}N)	5,730 years
231-Protactinium (^{231} Pa)	Uranium Decay	227-Thorium (^{227}Th), ^{4}He	0.033
234-Uranium (^{234}U)	Uranium Decay	230-Thorium (^{230}Th), ^{4}He	0.25
36-Chlorine (^{36}Cl)	Cosmogenic, Bomb	36-Argon (^{36}Ar)	0.31
26-Aluminum (^{26}Al)	Cosmogenic, Stellar	26-Magnesium (^{26} Mg)	0.73
230-Thorium (^{230}Th)	Uranium Decay	226-Radium (^{226}Ra), ^{4}He	0.75
60-Iron (^{60}Fe)	Stellar	60-Nickel (^{60}Ni)	1.5
10-Berylium (^{10}Be)	Cosmogenic, Stellar	10-Boron (^{10}B)	1.6
53-Manganese (^{53}Mn)	Cosmogenic, Stellar	53-Chromium (^{53}Cr)	3.7
107-Paladium (^{107}Pd)	Stellar	107-Silver (^{107}Ag)	6.5
182-Hafnium (^{182}Hf)	Stellar	182-Tungsten (^{182} W)	9
129-Iodine (^{129}I)	Stellar, Cosmogenic	129-Xenon (^{129}Xe)	15.7
244-Plutonium (^{244}Pu)	Stellar	Various fission products	80
146-Samarium (^{146}Sm)	Stellar	142-Neodymium (^{142}Nd)	103
235-Uranium (^{235}U)	Stellar	207-Lead (^{207}Pb), ^{4}He	704
40-Potassium (^{40} K)	Stellar	40-Argon (^{40}Ar) 40-Calcium (^{40}Ca)	1,270
238-Uranium (^{238}U)	Stellar	206-Lead (^{206}Pb), ^{4}He	4,469
232-Thorium (^{232}Th)	Stellar	208-Lead (^{208}Pb), ^{4}He	14,010
176-Lutetium (^{176}Lu)	Stellar	176-Hafnium (^{176}Hf)	35,700
187-Rhenium (^{187}Re)	Stellar	187-Osmium (^{187}Os)	41,600
87-Rubidium (^{87}Rb)	Stellar	87-Strontium (^{87}Sr)	48,800
147-Samarium (^{147}Sm)	Stellar	143-Neodymium (^{143}Nd), ^{4}He	106,000
190-Platinum (^{190}Pt)	Stellar	186-Osmium (^{186}Os)	450,000

occurring radioactive isotopes are produced through four mechanisms:

1. Stellar nucleosynthesis where the very high temperatures and pressures present in stellar interiors fuse nuclei, creating new elements (Truran and Heger, 2005).
2. Some radioactive elements decay into other radioactive elements. For example, 238-uranium requires eight alpha decays and six beta decays before it reaches stable 206-lead. Along the way, the decay steps include some isotopes with long enough decay lives to be useful in geochronology. These include ^{234}U, ^{230}Th, ^{226}Ra, ^{222}Rn, and ^{210}Pb.
3. High-energy cosmic ray particles collide with atoms in Earth's atmosphere or surface rocks with enough energy to cause nuclear reactions. For example, 14-carbon is created when 14-nitrogen in the atmosphere captures a neutron released by a cosmic ray interaction with some other atom. Other cosmic-ray-produced nuclides include ^{10}Be and ^{26}Al that are made by spallation, which occurs when an energetic cosmic ray proton simply breaks off a fragment from the nucleus of an atom with which it collides. Radioactive isotopes produced in these reactions are known as cosmogenic isotopes.
4. Above-ground detonations of nuclear bombs introduced into the environment substantial quantities of a number of radioactive species that have been used to investigate a number of atmosphere – Earth's surface exchange processes, and for tracing ocean water circulation.

The probability that a radioactive isotope will decay over any time interval is described by the equation:

$$-dN/dt = \lambda N \quad (1)$$

where N is the number of atoms of the radioactive species (parent isotope), t is time, and λ is the rate constant of the decay. Rate constants often are converted to "half-life"; the time needed for half of the radioactive species to decay. The half-life of a radioactive species is equal to ln(2)/λ. Integrating Equation 1 gives:

$$N = N_0 e^{-\lambda t} \quad (2)$$

where N_0 is the number of atoms present when t = 0. This equation can be used to determine ages if one has independent information on the initial abundance of the parent isotope. For example, in ^{14}C dating, if one assumes that the ratio of ^{14}C to stable ^{12}C is constant in the atmosphere, then one need only measure the present day ^{14}C/^{12}C ratio in a material that obtained its carbon from the atmosphere and use Equation 2 in order to determine the material's age. The assumption of a constant atmospheric ^{14}C/^{12}C

ratio is now known to be invalid because the production rate of ^{14}C in the atmosphere depends on issues as diverse as sunspot activity and the varying strength of Earth's magnetic field (see *Paleomagnetic Field Intensity*). Various schemes have been used to correct ^{14}C chronology for variations in the atmospheric ^{14}C production rate including comparison with the carbon in tree rings dated simply by counting the annual growth rings (Friedrich et al., 1999) or, for longer time intervals, to growth rings in corals dated with ^{230}Th and ^{234}U (Fairbanks et al., 2005). Although the variations in ^{14}C production rate complicate ^{14}C dating, they potentially provide information on the variability of Earth's magnetic field, or sunspot activity, and how these parameters may affect Earth's climate (Bard and Frank, 2006).

Because the initial abundance of the parent isotope generally is not known, most radioactive dating schemes measure the increase in the abundance of the decay product (daughter). If no atoms are lost, the abundance of the daughter isotope (D) increases with time at the rate described by:

$$D = N_0 - N = N\left(e^{\lambda t} - 1\right) \tag{3}$$

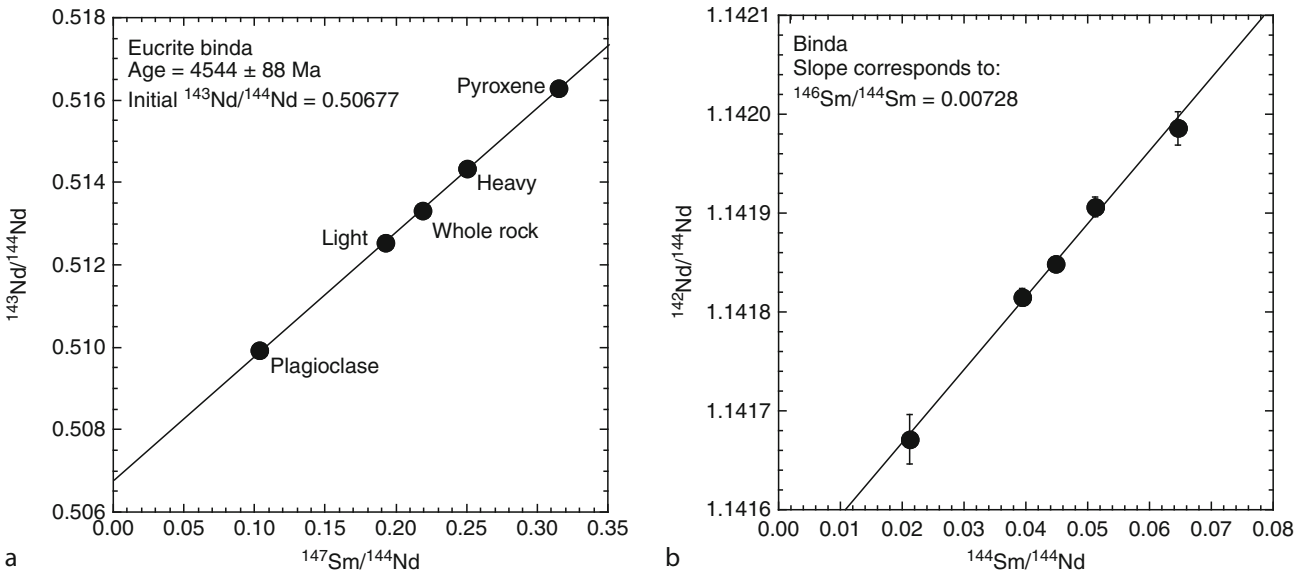
Thus, in an ideal situation where, for example, a mineral forms that contains radioactive ^{40}K, but no ^{40}Ar, the age of the mineral can be determined simply by measuring the abundance of ^{40}K (N) and ^{40}Ar (D) in the mineral after some decay interval has passed. In nature, perfect separation of parent from daughter isotope is rare, so if some atoms of the decay product are present initially (D_0), then:

$$D = D_0 + N\left(e^{\lambda t} - 1\right) \tag{4}$$

Because mass spectrometers can measure isotopic ratios of some elements to precisions of 0.0005% whereas elemental abundances can only be determined to about 0.1–1% precision, Equation 4, using the ^{238}U-^{206}Pb decay scheme as an example, can be expressed as:

$$\left(^{206}Pb/^{204}Pb\right)_m = \left(^{206}Pb/^{204}Pb\right)_0 + \left(^{238}U/^{204}Pb\right)_m\left(e^{\lambda t} - 1\right) \tag{5}$$

where "m" stands for the measured ratio and λ is the decay constant of ^{238}U. Pb-204 is a stable isotope of lead. If two or more samples with different U/Pb ratios are formed at the same time with the same initial Pb isotope composition ($(^{206}Pb/^{204}Pb)_0$), after some time has passed, plotting the measured $^{206}Pb/^{204}Pb$ for each sample against its $^{238}U/^{204}Pb$ ratio will produce a line, called an isochron, whose slope and y-intercept correspond to $(e^{\lambda t} - 1)$, and the initial $^{206}Pb/^{204}Pb$ ratio, respectively (Figure 1a). Thus, the slope of an isochron provides the time when the samples on the line had the same Pb isotope composition. The U–Pb system is unique in that it contains two isotopes of U that decay into two isotopes of Pb.



Absolute Age Determinations: Radiometric, Figure 1 Isochron diagrams for the long-lived ^{147}Sm–^{143}Nd and short-lived, now extinct, ^{146}Sm–^{144}Nd systems. The data shown are for minerals separated from the Binda eucrite, a type of meteorite that is a volcanic rock erupted on some small planetesimal. The meteorite consists primarily of the minerals plagioclase and pyroxene, which have different ratios of Sm to Nd when they form, but the same initial Nd isotopic composition. As Sm decays, it changes the isotopic composition of Nd. Using an equation similar to 5, the slope on the ^{147}Sm-^{143}Nd diagram directly provides the age of the sample, but on the ^{146}Sm-^{142}Nd diagram the slope provides the initial abundance of ^{146}Sm (actually the $^{146}Sm/^{144}Sm$ ratio as illustrated in an equation analogous to Equation 8). Data from Boyet et al. (2010).

Thus, an analogous Equation 5 can be written for the decay of ^{235}U to ^{207}Pb. More importantly, the two equations can be combined to give:

$$\frac{\left(^{207}Pb/^{204}Pb\right)_m - \left(^{207}Pb/^{204}Pb\right)_0}{\left(^{206}Pb/^{204}Pb\right)_m - \left(^{206}Pb/^{204}Pb\right)_0} = \frac{\left(^{235}U/^{238}U\right)\left(e^{\lambda_5 t} - 1\right)}{\left(e^{\lambda_8 t} - 1\right)} \quad (6)$$

where λ_5 and λ_8 are the decay constants for ^{235}U and ^{238}U, respectively. Because the $^{235}U/^{238}U$ ratio is nearly constant in nature at any given time, this equation allows an age to be determined from measurements of lead isotopic composition alone, without the need for concentration determinations of U or Pb. Such ages, called Pb–Pb ages, are obtained by plotting $(^{207}Pb/^{204}Pb)_m$ versus $(^{206}Pb/^{204}Pb)_m$, which will produce a line for two or more samples whose slope is equal to the right hand side of Equation 6. This equation can then be solved numerically to determine "t", the time when the samples formed.

A very active application of radiometric chronology uses the variety of short-lived (half-lives of 10^5–10^8 year) radionuclides that were present when the solar system formed, but have long since decayed away (Carlson and Boyet, 2009). For these systems, the "t" in Equation 2 cannot be referenced to today because $e^{-\lambda t}$ would always be close to zero. Instead, the "t" for these extinct systems must be a time interval referenced to a time when the parent isotope still existed. In this case, using the decay of ^{26}Al to ^{26}Mg as an example, Equation 5 can be rewritten as:

$$\left(^{26}Mg/^{24}Mg\right)_{\Delta t} = \left(^{26}Mg/^{24}Mg\right)_0 + \left(^{26}Al/^{27}Al\right)_0 \left(^{27}Al/^{24}Mg\right)_m e^{-\lambda \Delta t} \quad (7)$$

Plotting the two measurable parameters of this equation ($^{27}Al/^{24}Mg$, $^{26}Mg/^{24}Mg$) for two or more samples will produce a line whose slope gives not the age, but the $(^{26}Al/^{27}Al)$ at the time when the samples formed (Figure 1b). The time interval between the reference time and the time of sample formation is then:

$$\Delta t = -\left(\ln\left[\left(^{26}Al/^{27}Al\right)_{\Delta t} \Big/ \left(^{26}Al/^{27}Al\right)_0\right]\right)/\lambda \quad (8)$$

To convert these relative ages into absolute ages requires cross-calibration with a chronometer that provides absolute ages. For example, a certain type of mineral inclusion within the Allende meteorite provided a Pb–Pb age of 4567.6 ± 0.4 Ma and an Al-Mg isochron whose slope corresponds to $^{26}Al/^{27}Al = 4.96\ (\pm\ 0.25) \times 10^{-5}$ (Jacobsen et al., 2008). This provides a "pinning point" for the $^{26}Al/^{27}Al$ ratio present in the solar system at a known time. Thus, if analyses of some other sample produced an Al–Mg isochron whose slope corresponded to, for example, $^{26}Al/^{27}Al = 2.5 \times 10^{-5}$, this sample would be one half-life of ^{26}Al, or 750,000 years, younger than the other. Extinct radioactive dating systems thus provide not absolute, but relative, ages. If these relative ages are anchored to an absolute age, as done in the example above, then the extinct system can be used to indicate that the hypothetical sample mentioned above is 4566.8 million years old. A good example of the use of extinct radionuclides is the attempt to date the oldest volcanism in the solar system. Several types of meteorites are interpreted to be lavas produced on small planetesimals. One of these, the D'Orbigny meteorite, has provided Al–Mg, Mn–Cr, Hf–W, and Pb–Pb ages of 4562.8 ± 0.5 Ma, 4564.0 ± 0.6 Ma, 4562.8 ± 1.5 Ma, and 4564.42 ± 0.12 Ma, respectively. These results offer the promise of temporal precisions of 0.002% for the earliest events in solar system history.

The application of radiometric dating

The relatively simple equations given in the previous section provide the basis by which measurements of radioactive decay products can be translated into absolute ages. Applying these chronometers to natural processes requires matching the distinct properties of the many radioactive systems listed in Table 1 to the problem being investigated. Systems with short half-lives can be used to determine precise ages of geologic events only a few to a few thousand years old. Many of these short-lived nuclides are created either in the atmosphere or upper meter of Earth's surface, which makes them particularly amenable to dating a variety of near-surface processes. The concentrations of ^{10}Be or ^{26}Al in quartz in rocks or sediments can be used to date the time when the rock was exposed to cosmic rays, which can provide information on both uplift and erosion rates (von Blanckenburg, 2005) (see *Paleoseismology*). The rates of water movement from surface through subsurface aquifers can be traced with a number of radiometric chronometers (Bethke and Johnson, 2008). The presence of ^{10}Be in the lavas erupted at convergent margins tracks the subduction of surficial sediments to the 100 km + depths of magma generation (Morris, 1991) (see *Subduction Zones*). Carbon-14 is used to date a huge variety of events from the eruption of young volcanoes, movement along faults (see *Earthquakes and Crustal Deformation*; *Archaeoseismology*; *Paleoseismology*), the rates of ocean water circulation and ocean uptake of atmospheric CO_2, and the death of a living organism with its applications in archeology and paleoecology (Broecker, 2005). The longer-lived radionuclides can be applied to a wide variety of issues concerning the evolution of the solid earth, from dating the formation of crustal rocks (see also *Paleomagnetism, Magnetostratigraphy*), to determining the time of continental assembly and breakup (see *Continental Drift*; *Plates and Paleoreconstructions*), to defining the age of the Earth.

The key point in the application and proper interpretation of radiometric dating is understanding exactly what

event is being dated. What "starts" a radiometric clock is some event that changes the composition of a sample and fixes this new compositional state until the sample is dated in the laboratory. For example, ^{14}C often is used to determine the age of young volcanic eruptions. This system works in this application not by dating the rock itself, but by dating charcoal formed when the lava flowed over, and burnt, a bush or tree. What the ^{14}C age actually dates is the time when the plant stopped exchanging carbon with Earth's atmosphere. In most cases, this happened because the plant was killed by the volcanic eruption, in which case, the ^{14}C age accurately dates the eruption. However, if a 100-year-old lava flow were to flow over a 10,000-year-old bristlecone pine tree trunk lying long dead on the ground, the ^{14}C date of the charcoal would be 10,000 years, not 100 years.

The precision of a radiometric age depends on the amount radioactive decay has changed the isotopic composition of the parent or daughter element. This depends primarily on the time passed relative to the half-life of the system, and in systems where measurement of the daughter element is involved, on the degree to which parent and daughter element are fractionated from one another. Thus, one would use ^{14}C with its 5,730 year half-life to date an organic remain from ancient Rome, but it would be useless to date a million year old limestone because the ^{14}C "clock" stopped measuring time when the ^{14}C decayed away after a few half-lives into the history of the limestone. Similarly, ^{147}Sm, with a 106 billion year half-life, is used to date a variety of materials ranging in age from millions to billions of years old, but would be useless to date very young processes because only 7×10^{-15} g of ^{143}Nd is produced every thousand years per ppm of Sm, and this amount simply is too small to measure accurately.

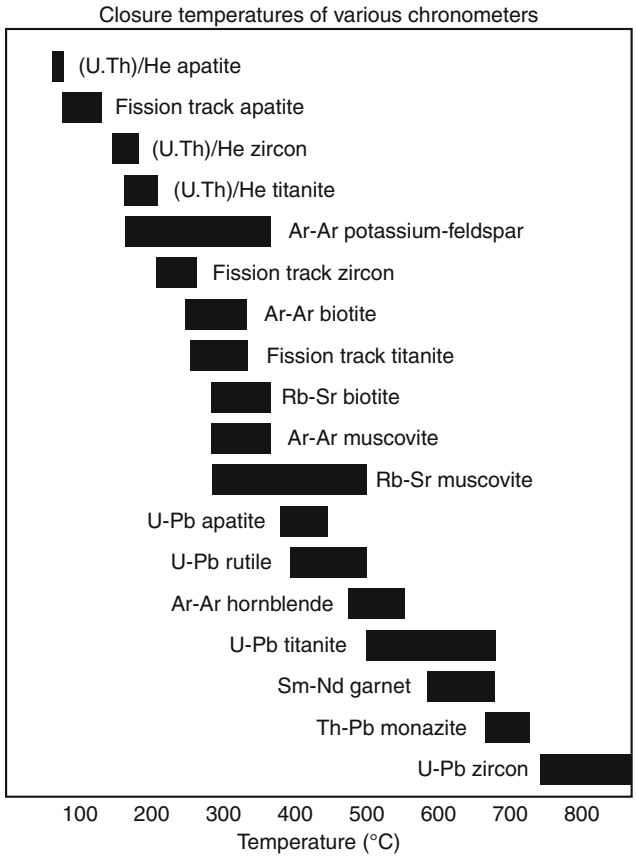
Besides choosing the proper radiometric system for the application, understanding whether a radiometric age is providing an accurate age for the event of interest demands an understanding of the processes that set, and reset, a radiometric clock. For Equation 5 to provide an accurate age demands that the samples used to construct the isochron: (1) all formed at exactly the same time, (2) all had exactly the same initial daughter isotope composition, (3) experienced no change in parent/daughter element ratio between formation of the sample and its measurement in the laboratory, and (4) nothing other than the decay of the parent modified the daughter isotopic composition. In nature, all of these requirements have the potential to fail, but, in some cases, these "failures" can still provide useful information.

A good candidate to meet the requirements above is a volcanic rock. Under conditions ideal for an accurate age, a well-mixed magma will crystallize, over a short time interval after eruption, a variety of minerals with different parent/daughter elemental ratios, but the same initial daughter isotopic composition that they inherit from the magma. Over time, if the minerals experience no chemical exchange with their surroundings, the radioactive decay products will build up allowing their measurement to provide the accurate age of eruption of this volcanic rock. The Apollo 15 basalt 15386 provides evidence that when these criteria are met, ages from many radioactive systems can agree. For this sample, the age (in billions of years) obtained from different radiometric systems are: K-Ar 3.89 ± 0.04, Rb-Sr 3.88 ± 0.01, Sm-Nd 3.85 ± 0.08, and the U–Pb age for zircon from a compositionally similar rock is 3.89.

For volcanic rocks this old, the assumption that the minerals all formed at the same time, at least relative to the ancient age of the rock, likely will be accurate. When one moves to younger volcanic systems, and the use of chronometers with better temporal resolution, the issue of the duration of crystallization becomes important. Several cases have now been presented where different minerals in the same volcanic rock give ages that range from the age of eruption to significantly older. One example is the Coso Volcanic field in California where zircons in one of the rhyolites range in age from the same as a K-Ar date on sanidine from the rhyolite, interpreted as the eruption age, to as much as 200,000 years older (Miller and Wooden, 2004). This result suggests that emplacement and crystallization of distinct magma batches in the crustal magma chamber that eventually fed this eruption is a prolonged process occurring over many thousands of years. Similar results have now been reported in many young volcanic rocks (Cooper and Reid, 2008). Although the age range of the minerals mentioned above is interpreted as reflecting a prolonged magmatic event, another way that an old crystal can get into a young magma is if a crystal from an older rock underlying the volcano is picked up by the magma during its ascent. Finding such "xenocrystic" material, particularly within more viscous, lower melting temperature, silica-rich magmas, is so common that it can be used to detect and even determine the age of older crust buried beneath a volcano, in some cases enabling the identification of older basement where none was previously known to exist (Smyth et al., 2007).

This brings up what is possibly the most insidious cause of an inaccurate radiometric age: the generally poorly-known temperature history of the sample. Because the rate of chemical diffusion within solids increases with temperature, elevated temperatures facilitate chemical changes. Heating a mineral can cause gases, such as He and Ar that are present in a crystal lattice as a result of their production through radioactive decay, to diffuse out of the crystal and escape along grain boundaries. At a temperature of 1,000°C, diffusion will cause a Sr atom in a mineral to move about 2 mm in a million years. If a small grain of a high Rb/Sr ratio mineral, such as mica, is next to a grain of a low Rb/Sr ratio, and high Sr content, mineral like plagioclase, the ingrowth of radiogenic ^{87}Sr in the mica will be diluted by diffusional exchange with the unradiogenic Sr in the plagioclase. Diffusion thus makes all radioactive chronometers "thermochronometers" that date the time when the material of interest reached a low-enough temperature to stop diffusional resetting of

the radiogenic ingrowth. Because rates of diffusion are a function both of the element that is diffusing and the structure of the crystal through which it is diffusing, different radiometric systems, and different minerals, can have different sensitivities to temperature excursions. This gives rise to the concept of "closure" or "blocking" temperature, above which the diffusion rate is too high to allow for the buildup of the radioactive decay product. In other words, the radiometric clock does not start until the mineral falls below its closure temperature. In the Coso example above, the K–Ar age obtained for the sanidine is interpreted as the eruption age because even if this mineral formed well prior to eruption, the 200–400°C closure temperature of the K–Ar system in sanidine (Figure 2) is so low that the radiogenic ^{40}Ar being produced in the sanidine was diffusing out as fast as it was being produced because of magma temperatures in the range of ~750°C.



Absolute Age Determinations: Radiometric, Figure 2 Closure temperatures for various radiometric systems in different minerals. The fission track technique counts the density of damage tracks made in crystals due to the fission of ^{238}U. Coupled with the measurement of U concentration in the crystal, the density of fission tracks can be converted into a radiometric age. Figure modified from (Pollard, 2002 – http://pangea.stanford.edu/~dpollard/NSF/main.html) which was adapted from a figure produced by P. Fitzgerald, S. Baldwin, G. Gehrels, P. Reiners, and M. Ducea.

In contrast, the closure temperature of the U–Pb system in zircon is higher than the magma temperature, so as soon as the zircon first crystallized, its U–Pb clock started to record time. Temperature-induced diffusional resetting of radiometric systems may make interpretations of ages a bit more difficult, but it creates a new field – thermochronometry – that constructs time-temperature histories for various geologic processes (Reiners and Ehlers, 2005). Though many radiometric systems can be used for thermochronometry, the oldest radiometric dating technique, the decay of U and Th to He, has been reenergized (Farley, 2002) because the low closure temperature of the U,Th – He system can be used to address geological processes such as the rate of mountain uplift or surface erosion by telling the time when a rock arrived close enough to Earth's surface to cool below the closure temperature of the radiometric clock (Reiners and Brandon, 2006).

While the appreciation and use of thermochronometry has been the primary growth field in radiometric geochronology in the last decade, other efforts are directed at finding ways to avoid reset ages. The traditional way to do this is to move to sample sizes much larger than diffusion lengths. This is the basis for "whole rock isochrons" where kilogram rock samples are homogenized and measured as an individual point on an isochron. Although this approach has produced some accurate ages, it suffers two main weaknesses. First, deciding whether all the rocks included on an isochron are the same age is not always easy, particularly for ancient, highly deformed, gneisses where this approach often is applied. Second, in order to obtain a sufficient range in parent/daughter ratio to precisely determine an isochron, compositionally distinct rocks are usually included on the isochron. Judging from modern igneous rocks, large compositional distinctions often are associated with variable initial isotope composition, which would violate the requirement that all the samples used to construct an isochron have the same initial isotope composition. Where whole rock isochrons can be useful is in detecting when much older components are involved in the genesis of younger rocks. A good example of this are the Sm-Nd data for lunar basalts where mineral isochrons of individual samples provide eruption ages generally less than 4 billion years, but the whole rock data scatter about a line corresponding to an age near 4.45 Ga (Carlson and Boyet, 2009). The precision of this older age is debatable, as is its interpretation, but it could imply that the younger basalts are made by melting source materials in the lunar interior that formed from some global differentiation event on the Moon that occurred some time near 4.45 Ga. Memory of this event was erased from the minerals in the basalts, that know only of their crystallization from the magma, but the magmas themselves may "remember" the events that led to the formation of their source materials in the lunar interior.

The other approach to avoiding thermally reset ages is to use minerals whose crystal structure allows only very slow diffusion. One curious example is diamond.

Diamond cannot be dated directly because it contains insufficient amounts of the radioactive elements, but it often contains as inclusions minerals that can be dated. Even though these minerals were formed, and have been stored, at temperatures above 1,000°C and depths greater than 150 km in the mantle (see *Lithosphere, Continental: Thermal Structure*), the encapsulation by the diamond does not allow them to communicate through diffusion with their surroundings. As a result, dating of inclusions in diamond show that diamond formation in the mantle has occurred over a good part of Earth history, with the oldest diamonds being approximately 3.5 Ga (Pearson and Shirey, 1999) (see *Lithosphere, Continental*).

By far, the most commonly used and most versatile way to avoid reset ages is with the application of the U–Pb system in zircon (Hanchar and Hoskins, 2003). The strengths of U–Pb zircon dating include: (a) very high U–Pb closure temperature, (b) resistance to chemical modification through alteration and metamorphism, (c) high U concentrations, but initially vanishingly small Pb contents, (d) the U–Pb system includes two independent decay schemes that allow tests to determine whether or not the age has been disturbed (Wetherill, 1956), (e) among the long-lived radioisotopes, the relatively short half-life of ^{235}U allows highly precise ages to be obtained for ancient rocks, (f) both the trace element concentrations and Hf isotope composition in zircon provide information on the nature of the source rock from which the zircon was derived, and (g) U–Pb ages can be obtained in samples as small as fragments of individual grains or even spots, tens of microns in diameter, ablated with either lasers or ion beams. Perhaps the only weakness of zircon is that it is not found in all rock types, occurring primarily in felsic igneous rocks. These characteristics explain the many varied applications of zircon U–Pb dating that include the discovery of the oldest dated material on Earth, 4.36 Ga zircons from quartzites in western Australia (Wilde et al., 2001), to high-precision calibration of the geologic timescale (Bowring and Schmitz, 2003), to the previously described use of zircons to determine the duration of crystallization in the magma chambers of young volcanic rocks.

Summary and conclusions

Continually improving analytical capabilities coupled with expanding appreciation of the physical controls on what sets, and resets, radiometric clocks are allowing ever-increasing expansion of this now century-old approach. The variety of radioactive species now in use provide the ability to determine absolute ages for geologic processes that occur over timescales that range from days to the age of the Earth and solar system.

Bibliography

Bard, E., and Frank, M., 2006. Climate change or solar variability: What's new under the sun? *Earth and Planetary Science Letters*, **248**, 1–14.

Bethke, C. M., and Johnson, T. M., 2008. Groundwater age and groundwater age dating. *Annual Review of Earth and Planetary Sciences*, **36**, 121–152.

Bowring, S. A., and Schmitz, M. D., 2003. High-precision U-Pb zircon geochronology and the statigraphic record. *Reviews in Mineralogy and Geochemistry*, **53**, 305–326.

Boyet, M., Carlson, R. W., and Horan, M., 2010. Old Sm-Nd ages for cumulate eucrites and redetermination of the solar system initial ^{146}Sm/^{144}Nd ratio. *Earth and Planetary Science Letters*, **291**, 172–181.

Broecker, W. S., 2005. Radiocarbon. In Keeling, R. K. (ed.), *Treatise on Geochemistry*. Amsterdam: Elsevier, pp. 1–18.

Carlson, R. W., and Boyet, M., 2009. Short-lived radionuclides as monitors of early crust-mantle differentiation on the terrestrial planets. *Earth and Planetary Science Letters*, **279**, 147–156.

Cooper, K. M., and Reid, M. R., 2008. Uranium-series crystal ages. *Reviews in Mineralogy and Geochemistry*, **69**, 479–544.

Dickin, A. P., 2005. *Radiogenic Isotope Geology*, 2nd edn. Cambridge: Cambridge University Press.

Fairbanks, R. G., Mortlock, R. A., Chiu, T.-Z., Cao, L., Kaplan, A., Guilderson, T. P., Fairbanks, T. W., Bloom, A. L., Grootes, P. M., and Nadeau, M.-J., 2005. Radiocarbon calibration curve spanning 0–50, 000 years BP based on paired 230Th/234U/238U and 14C dates on pristine corals. *Quaternary Science Reviews*, **24**, 1781–1796.

Farley, K. A., 2002. (U-Th)/He dating: techniques, calibrations, and applications. In Porcelli, D., Ballentine, C. J., and Wieler, R. (eds.), *Reviews in Mineralogy and Geochemistry*: Noble gases in geochemistry and cosmochemistry, pp. 819–844.

Faure, G., and Mensing, T. M., 2005. *Isotopes: Principles and Applications*, 3rd edn. Hoboken: John Wiley & Sons.

Friedrich, M., Kromer, B., Spurk, M., Hofmann, J., and Kaiser, K. F., 1999. Paleo-environment and radiocarbon calibration as derived from Late Glacial/Early Holocene tree-ring chronologies. *Quaternary International*, **61**, 27–39.

Hanchar, J. M., and Hoskins, P. W. O., 2003. Zircon, *Reviews of Mineralogy and Geochemistry*, **53**, 500.

Holmes, A., 1913. *The Age of the Earth*. London: Harper and Brothers.

Jacobsen, B., Yin, Q., Moynier, F., Amelin, Y., Krot, A. N., Nagashima, K., Hutcheon, I. D., and Palme, H., 2008. ^{26}Al-^{26}Mg and ^{207}Pb-^{206}Pb systematics of Allende CAIs: canonical solar initial ^{26}Al/^{27}Al reinstated. *Earth and Planetary Science Letters*, **272**, 353–364.

Miller, J. S., and Wooden, J. L., 2004. Residence, resorption and recycling of zircons in Devils Kitchen rhyolite, Coso Volcanic field, California. *Journal of Petrology*, **45**, 2155–2170.

Morris, J. D., 1991. Applications of cosmogenic ^{10}Be to problems in the Earth sciences. *Annual Review of Earth and Planetary Sciences*, **19**, 313–350.

Pearson, D. G., and Shirey, S. B., 1999. Isotopic dating of diamonds. In Lambert, D. D., and Ruiz, J. (eds.), *Application of Radiogenic Isotopes to Ore Deposit Research and Exploration*. Boulder: Society of Economic Geologists, pp. 143–172.

Reiners, P. W., and Ehlers, T. A., 2005. Low-temperature thermochronology: Techniques, interpretation, and applications, *Reviews in Mineralogy & Geochemistry*, **58**, 622.

Reiners, P. W., and Brandon, M. T., 2006. Using thermochronology to understand orogenic erosion. *Annual Review of Earth and Planetary Sciences*, **34**, 419–466.

Rutherford, E., 1906. *Radioactive Transformations*. New York: Scribner's.

Smyth, H. R., Hamilton, P. J., Hall, R., and Kinny, P. D., 2007. The deep crust beneath island arcs: inherited zircons reveal a Gondwana continental fragment beneath East Java, Indonesia. *Earth and Planetary Science Letters*, **258**, 269–282.

Truran, J. W. J., and Heger, A., 2005. Origin of the elements. In Davis, A. M. (ed.), *Treatise on Geochemistry*. Amsterdam: Elsevier, pp. 1–16.
von Blanckenburg, F., 2005. The control mechanisms of erosion and weathering at basin scale from cosmogenic nuclides in river sediments. *Earth and Planetary Science Letters*, **237**, 462–479.
Wetherill, G. W., 1956. Discordant U-Pb ages. 1. *Transactions on American Geophysical Union*, **37**, 320–326.
Wilde, S. A., Valley, J. W., Peck, W. H., and Graham, C. M., 2001. Evidence from detrital zircons for the existence of continental crust and oceans on Earth 4.4 Gyr ago. *Nature*, **409**, 175–178.

Cross-references

Archaeoseismology
Continental Drift
Continental Rifts
Energy Budget of the Earth
Geodynamics
Lithosphere, Continental
Lithosphere, Continental: Thermal Structure
Paleomagnetic Field Intensity
Paleomagnetism, Magnetostratigraphy
Paleoseismology
Plate Tectonics, Precambrian
Radiogenic Heat Production of Rocks
Subduction Zones

ARCHAEOMAGNETISM

Donald H. Tarling
School of Earth, Ocean and Environmental Sciences, University of Plymouth, Plymouth, UK

Synonyms

Magnetic dating, archaeological; Remanence dating

Definition

Archaeomagnetism is the study of all remanent magnetization associated with materials found in, or associated with, an archaeological context. In practice, it is the application of paleomagnetic techniques to archaeological materials, predominantly in terms of their uses in dating or "sourcing" such materials. This generally means that magnetic surveys are not usually considered "archaeomagnetic," other than when used to provide information on the actual directions and intensities of the remanent magnetizations of the materials causing the anomalies.

As most applications of archaeomagnetic studies relate to magnetic dating, these aspects are described first.

Basic features

Most materials within archaeomagnetic contexts contain iron oxide minerals that are capable of carrying a magnetic remanence acquired at some past time. More rarely, iron sulphides and hydroxides can carry such remanences. Even more rarely, a remanence can be carried by pure iron, nickel and cobalt and their alloys, and in very specific conditions, iron sulphates. The commonness of magnetite (Fe_3O_4) and hematite (Fe_2O_3) makes these almost always the dominant minerals involved. Such ferromagnetic (*sensu lato*) minerals can acquire a *thermal remanence* in the Earth's magnetic field (or of nearby magnetic objects) while cooling from higher than ambient temperatures, *chemical remanences* as a result of crystalline or compositional changes. Some archaeological materials can comprise iron objects, but these are generally poor recorders of early remanences, being generally magnetically "soft." All archaeological materials can become comminuted and eventually deposited as sediments (*v.i.* Archaeological sediments) in which the individual magnetic grains can be magnetically aligned during their deposition in aqueous or aeolian conditions, resulting in a *detrital magnetization* (see *Paleomagnetism, Principles* for details).

In an archaeological environment, it is unlikely that only one form of remanence is present. It can also be presumed that some of the originally acquired remanence will have decayed with time, and other time-dependent magnetizations, *viscous magnetizations*, will have been acquired as the materials lay within a geomagnetic field that gradually changes in both direction and intensity. In order to determine the direction and intensity of the geomagnetic field for some specific time in the past (such as when last heated or deposited), it is necessary to remove such viscous magnetization. Where such viscous magnetizations are not associated with chemical changes, they can be removed by partial demagnetization in alternating magnetic fields or heating in zero-magnetic field in nonmagnetic furnaces. Alternating field demagnetization randomizes the low coercivity components of remanence, that is, the viscous remanences, leaving the high coercivity components unchanged. Partial thermal demagnetization reduces the relaxation time of all contained minerals, that is, the time taken for the direction of individual grain magnetization to relax into the ambient field direction. As the ambient field is zero, the grains with low relaxation times (those that carry the viscous magnetizations) thereby become randomized, leaving the high relaxation grains still in original directions. (For further detail, see *Paleomagnetism, Principles*.) Such partial demagnetization procedures are necessary for magnetic dating using directions or intensity. Statistical analyses, initially Principle Component Analyses, are then made to determine the number, direction, intensity, and precision of the magnetic vectors present in any single sample (see *Paleomagnetism, Measurement Techniques and Instrumentation* for details). The isolated individual sample vectors defined within some 5° that are considered to have been acquired originally are then combined, usually using Fisherian Statistics and giving equal weight to each sample vector, to provide a mean site direction. (Where individual oriented hand samples have been subsampled, these subsamples are combined to obtain the sample values that then yield the site mean value.)

Such treatments normally require different instrumentation for dating based on directions than those for dating based on intensity and are described separately. The theoretical bases are the same.

Archaeological site sampling

Directional analyses

Generally between 7 and 20 individually oriented samples are taken. This enables orientation errors to be averaged, as well as measurement errors. Ideally such samples are distributed evenly throughout the structure being sampled, although concentration is made on materials that are most likely to yield meaningful results and where the physical stability (since acquiring remanence) is better established. Sample orientation in situ is preferably undertaken nonmagnetically (sun compass, gyro-theodolite, or by sightings) as the objects themselves commonly distort magnetic compass readings. It is also desirable that the nature and conditions at the site, with the in situ samples, is recorded so that, when completing the analyses, possible causes from anomalous determinations can be better evaluated (see *Paleomagnetism, Measurement Techniques and Instrumentation*).

Paleo-intensity analyses

Samples for such studies do not need to be oriented while in situ nor need the samples be still in their original positions in which their remanence was acquired. However, such orientations can be useful is assisting assessments of samples that show more than one remanent magnetic vector. Generally, some three to five samples are selected for study from any one archaeological site (see *Paleomagnetic Field Intensity*).

Magnetic dating based on the direction of remanence

This dating method depends on the longtime spatial variations ($>\sim$10 years) of the geomagnetic field (see *Geomagnetic Field, Secular Variation*). Analyses of the present-day geomagnetic variations suggest that, at a medium latitude location, the observed geomagnetic field in a circular area of some 550 km radius ($\sim$1 Mkm^2) can be modeled as that of a geocentric inclined dipole field within an error of $\sim$0.5° (solid angle) in direction. At greater distances, the error increases to $>$1° (solid angle) rendering the model increasingly unreliable for comparing the directions of remanence at more distant sites. Such modeling enables "Master Curves" (Paleo-Secular Variation Curves) of directional change to be constructed using previously studied sites for specific regions, usually for a specific country or groups of small countries. In the UK, most English and Welsh observations were corrected to a central location, Meriden, while French observations were generally corrected to Paris.

Master Curves are now more commonly calculated within a far more sophisticated statistical modeling, notably hierarchical Bayesian modeling (Lanos et al., 2005). In such analyses, the Gaussian errors for the date assignments as well as for directional properties are incorporated in the construction of the Master Curves. Such Bayesian statistics similarly incorporate harmonic analyses. Spherical Cap analyses have the objective to determine the geomagnetic field variations attributable to fields at the Earth's core (Pavón-Carrasco et al., 2008). The latter method is therefore more appropriate for geomagnetic field modeling (see *Geomagnetic Field, Theory*).

Relative dating

Nearby sites (within some 50–80 km) that acquired their remanence at the same time will exhibit identical (within 1°) directions of remanence. Consequently, nearby sites with differing directions of remanence can be considered to have acquired their remanence at different times. Conversely, similar directions can indicate similar ages for their remanence acquisition. However, identical geomagnetic directions can occur in the same location at different times. When such repeat directions occur at time intervals of a few 100 years, the archaeological context is often adequate to distinguish between alternative age assignments.

Chronological dating

This dating method is dependent on the secular variation of the geomagnetic field being known for the region being investigated (see *Geomagnetic Field, Secular Variation*). As the current geomagnetic theories for the origin of the geomagnetic field are still evolving (see *Geomagnetic Field, Theory*), this usually requires previously well-dated (by non-archaeomagnetic means) site values to be available. In newly studied areas, the method is clearly inhibited by the lack of such data, but as more and more observations are obtained, the past directions of the geomagnetic field become increasingly well defined. Archaeomagnetic dating is thus an unusual scientific dating technique as its precision increases as more observations accumulate.

Magnetic polarity and polarity excursion dating

This dating application is based on unusual aspects of the long-term geomagnetic behavior. On scales of a few thousand years, the geomagnetic field can undertake major excursions away from its normal direction, geomagnetic events, or excursions. On timescales of a few 100,000 years, the Earth's magnetic field changes polarity – the North magnetic pole becoming the South magnetic pole, and vice versa for the South magnetic pole. Consequently, some geomagnetic excursions occur on archaeological timescales, and the younger polarity changes occur on archaeological-anthropological timescales (see *Geomagnetic Excursions*; and *Geomagnetic Field, Polarity Reversals* for more detailed considerations).

Magnetic dating based on the intensity of remanence

Relative dating can be based on the determination of paleo-intensities of the ancient field in a similar way as for direction (qv). Most current studies are directed toward defining the changes in geomagnetic intensity during archaeological time although sufficient quantities of high quality paleo-intensity determinations are now becoming

available for specific regions, enabling chronological dating to be undertaken. As for directional studies, when past records of the Earth's magnetic field intensity are available, a Master Curve can be constructed for a similar circular area of ~1 Mkm^2, radius of ~550 km, as for directional studies.

Errors

General errors

These apply to both directional and intensity dating techniques.

1. It is usually assumed that any effects of weathering and other environmental changes are minimal and that the magnetic grains carrying the remanence of interest are essentially unaffected. However, some weathering products, such as goethite and hematite, can carry high-stability remanence. It is desirable that, as far as practicable, unweathered materials are collected.
2. During partial demagnetization, it is assumed that no chemical or structural alterations occur to the existing magnetic minerals and that no new magnetic minerals are formed. Such changes are far more likely during thermal demagnetization than by alternating magnetic field techniques and are usually monitored by repeated measurements of the initial susceptibility. This is, however, a coarse measure and magnetic domain changes (see *Magnetic Domains*) which can occur with no clear change in susceptibility.
3. Both directions and intensities can be affected by magnetic anisotropy (shape and crystalline) that may be associated with the sample shape, the shape of contained minerals and the orientations of their crystallographic axes (see *Magnetic Anisotropy*).
4. Lightning strikes at times subsequent to the original acquisition of remanence, and have electromagnetic effects that can drastically alter the magnetic remanence. Although mostly associated with electrical charges that usually travel through surface waters, these effects can penetrate. Such effects can usually be recognized by their high intensities and tendency to be in random directions.
5. The local geomagnetic field direction and intensity may well be distorted if there were nearby magnetized objects, natural or manufactured, at the time of remanence acquisition.
6. The behavior of the geomagnetic field is still poorly defined and, particularly at high magnetic latitudes, can show major diurnal variations. It also appears that secular variations are not necessarily smooth, but may occur in "jerks" (see *Geomagnetic Field, Secular Variation*) although these appear to be mostly within the current experimental error of the technique.
7. The geomagnetic field over time can return to previous direction and intensity values. When such "crossover" points occur in the Master Curves at sufficiently different times, then the archaeological age constraints on the site may enable the most likely chronological age to be applied. However, most extant Master Curves are likely to still be distorted by some studies for which the archaeological or chronological age has subsequently been modified. However, such "anomalous" sites becoming increasingly recognized as databases increase.

Errors specific to directional dating

In addition to the general errors, directional studies are sensitive to the oriented samples having remained in situ since they acquired their remanences. In a solid structure, such loose materials can be readily identified, but individual parts of some structures can have undergone motions relative to the rest of the structure, for example, kiln-wall "fall-out" or "fall-in." On hillside, it is also possible that the whole structure may have undergone motion away from where the remanence was acquired. Thus, a knowledge of the archaeological controls on possible motions of the site is essential.

Errors specific to intensity dating

In addition to the general errors, paleo-intensity studies are sensitive to the "cooling rate" of the acquired magnetic properties. The naturally acquired remanence will commonly have been acquired over times that can be of the order of several days, while the laboratory studies are necessarily on far shorter times scales. The laboratory redox conditions during cooling are normally very different.

Overall error assessment

While there are many potential sources of error in archaeomagnetic dating, in most archaeological sites where the conditions are clear and the materials largely pristine, the errors arising appear to be only rarely of sufficient magnitude to warrant rejection.

Specialized applications

Murals and plaster

Remarkably, it has been shown that pigments can carry a remanence relating to the time when they were applied. Where such pigments have been applied to walls, their remanence has been shown to relate to the time that the pigments were applied, particularly in tempera paintings – either originally or during subsequent retouching. It seems most likely that this remanence is associated with hematite and was acquired "detritally," that is, by rotation of the grains while still fluidized (Lanza et al., 2009). Similar observations have also been shown for plaster work (Soler-Arechalde et al., 2006).

Object reconstruction

Specific artifacts, such as pottery, when fired, obtain a uniform magnetization in the ambient geomagnetic field as they cool within the furnace. If broken, individual shards retain this original remanence and can be reassembled to assist in defining the original shape of the pot (Burnham and Tarling, 1975).

Archaeological sediments

As sediments are deposited within an archaeological site, any contained already magnetized grains are influenced by the ambient geomagnetic field. As these alignment forces are weak (compared with the effects of gravity and flow motions) only a percentage of the magnetic grains are aligned and such alignments can be gravitational flattening during the deposition process. However, the more magnetically stable grains (usually with single-domain dimensions of ~1 μm) can rotate into full alignment while they are still fluidized immediately after deposition. However, such alignments can be modified by meniscus forces as the water table subsequently rises or falls. Sedimentary minerals also undergo a series of chemical reactions as they adjust to their new environment. Consequently, it is often difficult to establish the true direction of the geomagnetic field at the time of deposition with sufficient certainty to assign ages based on the magnetic determinations. However, there are archaeological sites, such as basal sediments, deposited immediately after a ditch had been cut, that appear to be recording the geomagnetic field direction at that time.

Magnetic "sourcing"

Magnetic sourcing attributes an object to a particular location on the basis of its magnetic properties. For example, an obsidian arrowhead can sometimes be sourced to the obsidian outcrop from which it was originally fashioned.

The physical basis for this is that the compositions and grains sizes of the magnetic minerals within a rock can be specific to a particular outcrop. Each outcrop also tends to have a slightly different cooling history and became magnetized by a specific geomagnetic field strength. Several magnetic properties can be rapidly and cheaply measured, such as its initial (low field) susceptibility, its saturation remanence, and its intensity of natural remanence, are sufficient to identify a specific outcrop. Generally, the range of magnetic properties in most sedimentary rocks is too great to allow such assignments, but obsidians, rapidly cooled volcanic glasses, commonly have a range of magnetic properties that are either unique to that outcrop or to only two or three outcrops.

Summary

The study of the magnetic properties of archaeological materials, archaeomagnetism, is particularly significant in geophysics as it enables secular variations in the direction and strength of the geomagnetic field over timesscale that are far longer than those for direct measurement (~400 years). In comparable magnetic studies of rocks, paleomagnetism, these secular variations are commonly averaged out during the processes by which rocks acquire their remanent magnetization. Conversely, archaeomagnetic records provide an additional dating tool for archaeologists. As a comparative dating tool, archaeomagnetism enables relative dating between nearby samples (100–200 km) within ~10–20 years. As a dating tool, it is mainly dependent on the validity of the ages used to compile the regional Master Curve. It is thus unique, as a scientific dating method, in that as more data are acquired, age "outliers" can be identified and their age reassessed. Consequently, the Master Curves are continually improving and, ultimately, could have the same precision as for relative dating. There are also interesting applications in using such observations in conservation studies and establishing past environmental conditions.

Bibliography

Batt, C. M., and Zanarini, I. (eds.), 2008. Archaeomagnetic applications for the rescue of cultural heritage (AARCH). *Physics and Chemistry of the Earth*, **33**(6–7), 403–608.

Burnham, R. J. P., and Tarling, D. H., 1975. Magnetisation of shards as an assistance to the reconstruction of pottery vessels. *Studies in Conservation*, **20**, 152–158.

Lanos, Ph., Le Goff, M., Kovacheva, M., and Schnepp, E., 2005. Hierarchical modeling of archaeomagnetic data and curve estimation by moving average technique. *Geophysical Journal International*, **160**, 440–476.

Lanza, R., Zanella, E., and Sandino, S., 2009. Magnetic remanence of haematite-bearing minerals. *Geophysical Research Letters*, **36**, L24302.

Pavón-Carrasco, F. J., Osete, M. L., Torta, J. M., Gaya-Piqué, L. R., and Lanos, Ph., 2008. Initial SCHA.DI.00 regional archaeomagnetic model for Europe for the last 2000 years. *Physics and Chemistry of the Earth*, **33**(6–7), 597–608.

Soler-Arechalde, A. M., Sánchez, F., Rodriguez, M., Caballero-Miranda, A., Goguitchaichvil, A., Urrutia-Fufugauchi, J., Manzanilla, L., and Tarling, D. H., 2006. Archaeomagnetic investigation of oriented pre-Columbian lime-plasters at Teotihuacan, Mesoamerica. *Earth Planets Space*, **58**, 1433–1439.

Cross-references

Geomagnetic Excursions
Geomagnetic Field, Polarity Reversals
Geomagnetic Field, Secular Variation
Geomagnetic Field, Theory
Magnetic Anisotropy
Magnetic Domains
Paleomagnetic Field Intensity
Paleomagnetism, Measurement Techniques and Instrumentation
Paleomagnetism, Principles

ARCHAEOSEISMOLOGY

Klaus-Günter Hinzen
Earthquake Geology Group, Institute of Geology and Mineralogy, Cologne University, Bergisch Gladbach, Germany

Synonyms

Earthquake archeology

Definition

The study of pre-instrumental earthquakes that, by affecting locations of human occupation and their environments,

have left their mark in ancient structures uncovered by means of archaeological excavations or pertaining to the monumental heritage (Buck and Stewart, 2000; Galadini et al., 2006).

Introduction

Ever since man-made structures have been erected, earthquakes have left their marks on these constructions. However, damages in archaeologically excavated buildings or continuously preserved monuments are often hard to unravel in terms of the causative effects. The use of archaeological data to investigate unknown or poorly known historical earthquakes and descriptions of earthquake effects recorded in the archaeological heritage started in the nineteenth and early twentieth century (e.g., De Rossi, 1874; Schliemann, 1881; Evans, 1928). Since the 1980s, the increased interest in the subject led to the publication of special volumes and articles in seismological and geological journals (e.g., Guidoboni, 1989; Stiros and Jones, 1996; McGuire et al., 2000; Galadini et al., 2006; Reicherter et al., 2009). While earlier investigations were dominated by the qualitative description of damage records, more recent studies take a quantitative approach. The main questions to be answered by archaeoseismic investigations are (1) how probable are seismic ground motions, or secondary earthquake effects, as the cause of damage observed in man-made structures from the past, (2) when did the damaging ground motion occur, and (3) what can be deduced about the nature of the causing earthquake.

Archaeoseismic observations

The marks in ancient structures relevant for archaeoseismology fall into four main categories: (1) displacements along shear planes directly linked to the earthquake fault plane or side branches of it. In particular, earthquakes with a strike slip mechanism can leave distinctive traces in buildings and lifelines such as aqueducts, roads, and sewer systems. Case studies (e.g., Ellenblum et al., 1998; Meghraoui et al., 2003) show that under favorable conditions the amount of slip and in case of repeated events the slip rate of the faults can be revealed. (2) Off-fault-shaking effects including fractured building elements, tilted walls, shift of building elements, lateral warping, breaking and overthrow of walls, rotations of vertically oriented objects (tomb stones, columns, monuments). For most of these features, a seismogenic origin is not the only possible interpretation. Therefore, alternative causes must be taken into account in the damage analysis. (3) The secondary shaking affects lateral spreading and cyclic mobility as a consequence of liquefaction of the subsurface. Liquefaction always requires a certain level of dynamic excitation. So secondary damages in buildings and monuments due to liquefaction help exclude alternative causes from a damage scenario. (4) Archaeologically detected abandonment of a site and evidence of repair and rebuilding. These observations are mainly of interest in connection with the reasons mentioned before because as a single observation in general they do not give enough evidence for the conclusion of a seismogenic cause (Galadini et al., 2006).

While the first category is limited directly to the linear features of active faults, off-fault shaking affects a much larger area and is more common but harder to prove. Ancient structures show deformations related to seismic shaking similar to those observed in recent earthquakes. Typical earthquake effects on masonry walls are (1) cross fissures, due to stress concentrations often nucleating at corners of doors and windows, driven by shear forces, (2) corner expulsion of walls caused by differential movements in orthogonal directions, (3) horizontal and independent lateral and rotational shift of wall blocks, best visible in uncemeted walls made of rectangular blocks, (5) spall of block corners due to stress concentrations, (6) height reduction by vertical crashing, (7) movement of keystones and rupture of arch piers, (8) rotation of vertically oriented objects, (9) domino-like toppling of structured columns and walls. Examples are shown in Figure 1. Following the term seismite, used for seismically disturbed sediment layers, the above-mentioned deformations may be called archaeoseismites. Several of these deformations may also originate without dynamic earthquake excitation. In these cases, a single piece of evidence or at a single edifice only cannot be considered a conclusive archaeoseismite. If damages are identified as archaeoseismites, intensities can be derived following the classical macroseismic methods.

Methods

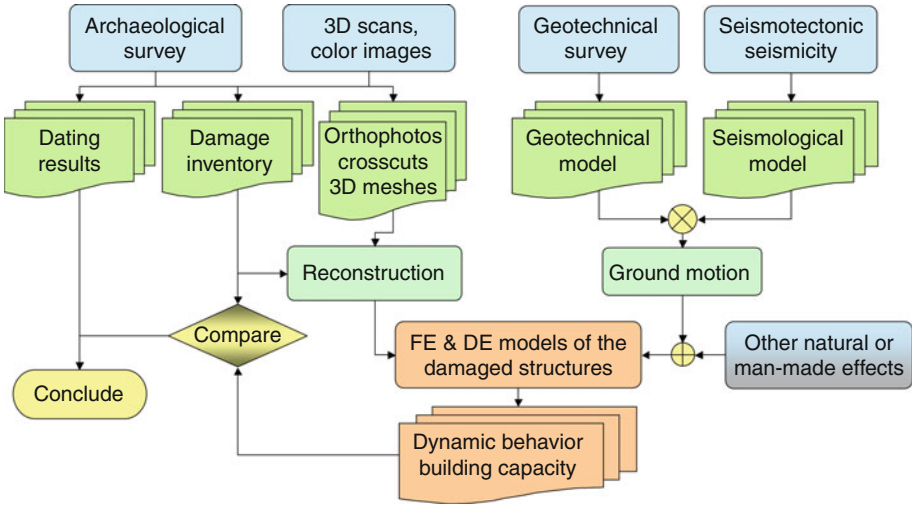
Archaeoseismology requires the integration of several scientific disciplines. Figure 2 shows a scheme of workflow in a quantitative archaeoseismic study. Besides archaeological and seismological methods, input from geodesy, pure and applied geology, and civil engineering is needed (Ambrasyes, 2006). Physical absolute dating techniques including radiometric methods, luminescence methods, and dendrochronology are necessary to determine a time window for the damaging event. In case of multiple events, the rough dates are crucial for seismicity models and hazard analysis.

The first step in studying field cases is the documentation and analysis of damages to the archaeologically excavated buildings or objects. In addition to the important traditional excavation techniques and classical measurements and drawings, 3D imaging techniques have proved their usefulness in archaeoseismic field cases. Photogrammetry and laserscanning are fast in the field and allow detailed analysis and post processing, especially a quantification of deformations, fractures, etc. In cases where building parts are fragile and cannot be conserved in the original finding situation, successive laser scans during the excavation offer the possibility to build virtual 3D models. In ideal cases, the seismological relevant data can be acquired during an ongoing excavation. Important

Archaeoseismology, Figure 1 Examples of deformations and damages which possibly are earthquake effects: (**a**) Horizontally deformed wall of a crusader fortress build on top of the Dead Sea Transform Fault in the Jordan Valley; (**b**) deformed vault of a Roman sewer in Cologne, Germany; (**c**) toppled columns of a Byzantine church in Sussita located above the Sea of Galilee; (**d**) toppled column of the great palace in Patra, Jordan; (**e**) moved block in an arch of the Nimrod fortress in the Golan Heights; (**f**) shifted blocks of an analemma of a Roman theatre in Pınara, SW Turkey; (**g**) moved blocks of a corner wall of a Roman monument in Patara, SW Turkey; (**h**) shifted blocks of a Roman grave house in Pınara, SW Turkey; (**i**) spall of block corners, same object as in (**g**);(**j**) broken and horizontally displaced fortification wall of the Roman Tolbiacum (Zülpich, Germany); (**k**) rotated Lycien sarcophagus in Pınara, SW Turkey.

is the complete documentation of the stratigraphy of debris layers, which can reveal the decay rate of structures (Galadini et al., 2006). Building collapse may be sudden, progressive, or a combination of both. The progressive deterioration of walls results in accumulations having maximum thickness close to the wall feeding the material. The thickness decreases a short distance from the wall. In case of height reduction of walls or buildings due to earthquake effects, the debris is more distributed. Sometimes even complete walls topple to one side; in this case, the impact pattern can help to constrain ground motion amplitudes. However, if the findings of previous excavations are archaeoseismologically analyzed, the current situation and available excavation reports form the database. Possible alterations of the excavated objects since the first excavation and shortcomings in the documentation, especially if earthquake actions were not considered at the time of the original excavation, might hinder the interpretation.

Depending on site conditions, exploration, and modeling of the properties of the local geology at the site is an important second step. Several of the above-mentioned deformations can also be caused by slow, quasi-static deformations of the subsoil, and ground amplification due to shallow soft sediments can significantly amplify seismic shaking. If the latter is not taken into account, the size of the damaging earthquake can easily be overestimated.



Archaeoseismology, Figure 2 Schematic work flow in a quantitative multidisciplinary archaeoseismic study. The dynamic behavior is studied by finite element (FE) or discrete element (DE) methods.

Only on rare occasions, site-specific strong motion records from recent earthquakes will be available and help to prove or disprove proposed damage scenarios. To overcome this lack of information, the calculation of synthetic seismograms based on models of the capable faults and local geology have been used to better ascertain the probability of a seismogenic cause of the damages and name faults on which the damaging earthquake occurred. When other natural forces (flooding, storm, mass movement) or even human action also come into consideration, their action on the excavated objects has to be tested as well.

Reconstructions of the damaged archaeological objects are the basis for determining the dynamic behavior and the vulnerability of the ancient buildings. Common earthquake engineering techniques including finite element and discrete element modeling are used to analyze the reaction of the buildings to the estimated seismic loading functions. Results from such models are compared to the damages observed in situ.

The most challenging task of an archaeoseismic study is the determination of source parameters of the earthquake which caused the analyzed damages. While timing of the earthquake has to be done by archaeological techniques, including classical archaeological stratigraphy and the above-mentioned physical dating, location and strength of the earthquake can only by narrowed down with seismological models. A territorial approach (Galadini et al., 2006), in which the collected information from the surroundings of an archaeological site are taken into account, is the most promising tool to determine earthquake source parameters.

Summary

In the past decades archaeoseismology has evolved as a new branch of seismological sciences. While palaeoseismology focuses on the marks left by earthquakes in the geological record, archaeoseismologists analyze and interpret marks in the archaeological record and in preserved monuments.

The use of quantitative methods in archaeoseismology, including calculation of synthetic site-specific strong motion seismograms, modeling of natural non-earthquake-related forces, anthropogenic forces, and finite or discrete element models of structures, supports conclusive discrimination between potential damage scenarios. However, if model parameters cannot be well constrained, modeling result uncertainties might still be too large to draw definite conclusions. Common sense interpretations of archaeoseismites as solitary evidence are generally too vague to complement earthquake catalogs for a seismic hazard analysis. Recent advances in ground motion simulation methods and computational possibilities promise to refine quantitative archaeoseismological methods and establish them at levels equal to historical and palaeoseismologial methods. Finally, even if an archaeoseismic study does not deliver the often-requested improvement of hazard determination, it can still advance our picture of the past by attempting to answer open archaeological, historical, and geologic questions in a scientific manner. And following Ambrasyes (2006): "Surveying and mapping an archaeological site is an art, verifying the cause of damage is science."

Bibliography

Ambrasyes, N. N., 2006. Earthquakes and archaeology. *Journal of Archaeological Science*, **33**, 108–1016.

Buck, V., and Stewart, I., 2000. A critical reappraisal of classical and archaeological evidence for earthquakes in the Atalanti region, central mainland Greece. In: McGuire, W. G., Griffith, D. R., Handcock, P. L., and Stewart, I. (eds.), *The Archaeology of Geological Catastrophes*, Geological Society, London, Special Publications 171, pp. 33–44.

De Rossi, M. S., 1874. La basilica di Santa Petronilla presso Roma, testé discoperta, caduta per terremoto. *Bulleltino del Vulcanismo Italiano*, **1**, 62–65.
Ellenblum, R., Marco, S., Agnon, A., Rockwell, T., and Boas, A., 1998. Crusader castle torn apart by earthquake at dawn, 20 May 1202. *Geology*, **26**, 303–306.
Evans, A., 1928. The Palace of Minos, part II. MacMillan: London, 844 pp.
Galadini, F., Hinzen, K.-G., and Stiros, S., 2006. Archaeoseismology: methodological issues and procedure. *Journal of Seismology*, **10**, 395–414.
Guidoboni, E. (ed.), 1989. *I terremoti prima del Mille in Italia e nell'area mediterranea. Storia, Archeologia, Sismologia*, ING-SGA, Bologna, (Italy), 765 pp.
McGuire, W. J., Griffith, D. R., Hancock, P. L., and Stewart I. S. (eds.), 2000. *The Archaeology of Geological Catastrophes*, Geological Society Spec. Publ. no. 171, London, 417 pp.
Meghraoui, M., Gomez, F., Sbeinati, R., Van der Woerd, J., Mounty, M., Darkal, A. N., Radwan, Y., Layyous, I., Al-Najjar, H., Darawcheh, R., Hijazi, F., Al-Ghazzi, R., and Barazangi, M., 2003. Evidence for 830 years of seismic quiescence from palaeoseismology, archaeoseismology, and historical seismicity along the Dead Sea Fault in Syria. *Earth and Planetary Science Letters*, **210**, 35–52.
Reicherter, K., Michetti, A. M., and Silva, P. G. (eds.), 2009. *Palaeoseismology: Historical and Prehistorical Records of Earthquake Ground Effects for Seismic Hazard Assessment*. The Geological Society, London, Special Publications, 316, 324 pp.
Schliemann, H., 1881. Autobiographie des Verfassers und Geschichte seiner Arbeiten in Troja. In: *Ilios. Stadt und Land der Trojaner, Forschungen und Entdeckungen in der Troas und besonders auf der Baustelle von Troja*, Leipzig, 78, 26 pp.
Stiros, S. C., and Jones, R. E. (eds.), 1996. *Archaeoseismology*. British School at Athens, Fitch Laboratory Occasional Paper 7, 268 pp.

Cross-references

Absolute Age Determinations: Radiometric
Earthquakes, Intensity
Earthquakes, Strong-Ground Motion
Numerical Methods, Finite Element
Paleoseismology
Seismic Hazard

ARTIFICIAL WATER RESERVOIR TRIGGERED EARTHQUAKES

Harsh K. Gupta
National Geophysical Research Institute, Council of Scientific & Industrial Research (CSIR), Hyderabad, Andhra Pradesh, India

Definition and Introduction

Under certain suitable geological conditions, anthropogenic activity can trigger or induce earthquakes. The triggered/induced earthquakes are known to have occurred due to gold and coal mining, petroleum production, filling of artificial water reservoirs, high-pressure liquid injection into ground, and natural gas production. The largest scientifically accepted triggered earthquake of magnitude 6.3 occurred on December 10, 1967 in the vicinity of Koyna Dam near the west coast of India. It is debated whether the M 7 Gazli earthquakes of May 1976 and March 19, 1984 were induced due to the production of large quantities of gas at the Gazli Oil Field in Uzbekistan. There is an argument that the Sichuan, China, M 7.9 earthquake of May 12, 2008, that claimed over 80,000 human lives was triggered due to filling of the nearby Zipingpu reservoir. It has been also proposed that flooding of a river near San Andreas fault in California caused at least two M $\sim$ 6 earthquakes. A good account of triggered/induced seismicity can be found in a review by McGarr et al. (2002). Hudyma and Potvin (2005) has extensively dealt with mining-induced seismicity. Gupta (2002) has reviewed artificial water reservoir triggered seismicity.

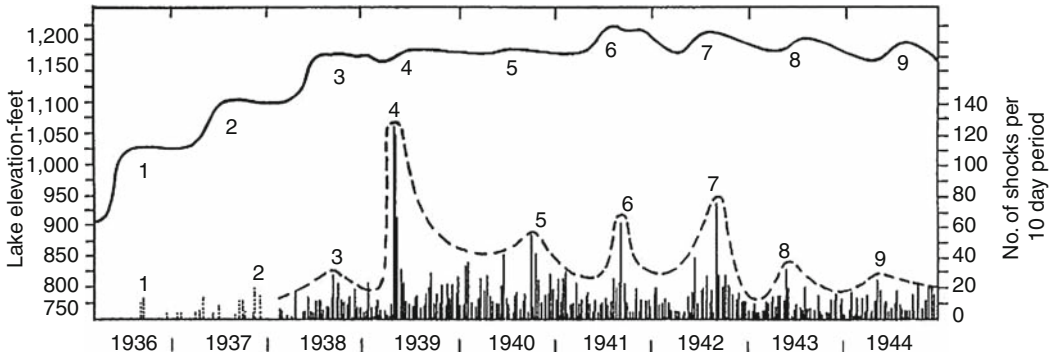
Here we present cases of artificial water reservoir triggered earthquakes that occurred all over the world, with a special emphasis on earthquakes in the Koyna region near the west coast of India. This is a classical case where triggered earthquakes have been occurring since the impoundment of reservoir in 1962.

Triggered vis-à-vis induced earthquakes

For a long time, the adjectives "induced" and "triggered" were used interchangeably whenever one talked of artificially simulated earthquakes. McGarr and Simpson (1997) have addressed this question and suggested that it would be important to draw a distinction between the two. They proposed that the adjective "triggered seismicity" should be used only when a small fraction of stress change or energy associated with earthquakes is accounted for by the causative activity. The term 'induced seismicity' should be used where the causative activity is responsible for a substantial part of the stress change. In case of triggered seismicity, tectonic loading plays an important role. The stress-level changes associated with filling of some of the deepest artificial water reservoirs are only of the order of 1 MPa or so, whereas the stress drop associated with the earthquakes is much larger. Therefore, all cases of earthquakes occurring subsequent to filling of the artificial water reservoirs fall in the category of "triggered earthquakes," and hence it is appropriate to call it "reservoir triggered seismicity" (RTS).

Artificial water reservoir triggered earthquakes

Generation of hydroelectric power, flood control, and irrigation necessitates creation of huge artificial water reservoirs globally. Triggering of earthquakes was for the first time pointed out by Carder (1945) at Lake Mead in USA. Figure 1 depicts Lake Mead water levels and local seismicity for the period 1936 through 1944. The rise in water level and corresponding bursts of seismic activity are numbered. The correspondence is indeed remarkable. Damaging triggered earthquakes exceeding magnitude six occurred at Hsingfengkiang, China (1962); Kariba, Zambia–Zimbabwe border (1963); Kremasta, Greece (1966); and Koyna, India (1967). Koyna earthquake of M 6.3 that occurred on December 10, 1967 is so far the largest scientifically accepted



Artificial Water Reservoir Triggered Earthquakes, Figure 1 Lake Mead water levels and the local seismicity. For 1936 and 1937, only the felt shocks are plotted. The rises in water levels and the corresponding bursts of seismic activity are numbered. General trend of tremor-frequency variation is shown by dotted lines (After Carder, 1945).

triggered earthquake. It claimed over 200 human lives, injured about 1,500, and rendered thousands homeless. The occurrence and potential of triggered earthquakes has caused major modification of civil works and engineering projects. Anticipating a large triggered earthquake, the Hsingfengkiang Dam was strengthened twice before the occurrence of M 6.1 earthquake on March 20, 1962 (Shen et al., 1974). The disposal of waste fluid through injection into the ground at Rocky Mountain Arsenal had to be discontinued due to triggered earthquakes (Evans, 1966). The possibility of high magnitude triggered seismicity was responsible for terminating the Auburn Dam project in California (Allen, 1978). There is a general reluctance on the part of Engineering Community, globally, to accept the significance or even the existence of the phenomenon of triggered seismicity (Allen, 1982; Simpson and Leith, 1988). What Allen (1982) said a quarter century back: "From a purely economic point of view, not to speak of public safety, the problem of reservoir induced earthquakes deserves far more attention than it currently is receiving in most parts of the world", is still true.

Reservoir triggered seismicity (RTS): global distribution

Globally there are about 100 sites where RTS is reported to have occurred (Table 1). These could broadly be grouped in the following categories:

1. Sites where largest earthquake exceeded M 6.0 (4 sites)
2. Sites where the largest earthquake M was 5.0–5.9 (10 sites)
3. Sites where the largest earthquake M was 4.0–4.9 (29 sites)
4. Sites where the largest earthquake M was less than 4.0 (55 sites)

Table 1 gives a global list of artificial water reservoir sites where triggered earthquakes of M $\geq$ 4 are known to have occurred.

Important factors for RTS

Several studies examined the correspondence among possible correlates like rate of loading, highest water level reached, and the duration of retention of high water levels and the occurrence of RTS. The most important correlate is the depth of water column in the reservoir (Baecher and Keeney, 1982). Figure 2 demonstrates that when the water column depth exceeds 200 m, about one in five reservoirs have experienced RTS. A review of recent global examples gives support to this observation. It must also be noted that a reservoir with a water volume exceeding 1 km^3 and/or a water depth exceeding 100 m is called a large reservoir. Globally, there are more than one thousand such reservoirs and only a small percentage has evidenced RTS.

Common characteristics of RTS sequences

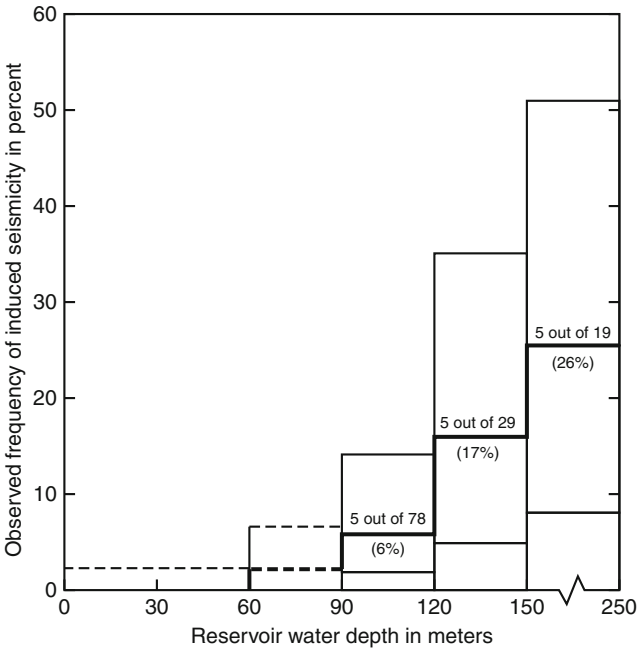
By early 1970s, over a dozen cases of RTS were known. In a couple of detailed studies, Gupta et al. (1972a, b) discovered several common characteristics of RTS sequences, which discriminated them from the normal earthquake sequences occurring in close by regions but not associated with reservoirs. These characteristics are:

1. In the earthquake frequency-magnitude relation ($\log N = A - b\,M$, where N is the number of earthquakes with magnitude $\geq$ M, and A and b are constants), the foreshock and aftershock *b* values of the RTS sequences are higher than the *b* values for natural earthquake sequences in the regions concerned, and the regional *b* values.
2. In addition to high *b* values, the magnitude ratio of the largest aftershock to the main shock is also high.
3. Aftershock activity decays slowly compared to normal earthquake sequences.
4. The foreshock–aftershock sequence pattern belongs to Type II of Mogi's Model (Mogi, 1963), where as the natural earthquake sequence pattern belongs to Type I of Mogi's Model.

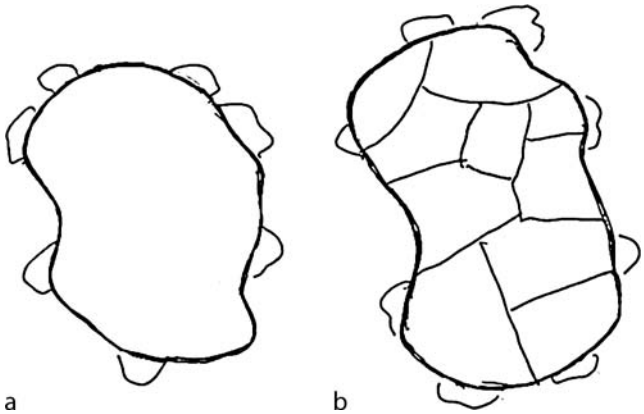
Artificial Water Reservoir Triggered Earthquakes, Table 1 Reported cases of reservoir triggered seismicity (RTS) where M ≥ 4 earthquake occurs

Name of the dam/reservoir	Country	Height of dam (m)	Reservoir volume (10^6 m^3)	Year of impounding	Year of the largest earthquake	Magnitude/ intensity	References
Sites where earthquakes having magnitude ≥ 6.0 were triggered							
Hsinfengkiang	China (PRC)	105	13,896	1959	1962	6.1	1, 2, 3
Kariba Zambia	Zimbabwe	128	175,000	1958	1963	6.2	1, 2, 6
Koyna	India	103	2,780	1962	1967	6.3	1, 2, 4,5
Kremasta	Greece	160	4,750	1965	1966	6.2	1, 2, 4,5
Sites where earthquakes having magnitude between 5.0 and 5.9 were triggered							
Aswan	Egypt	111	1,64,000	1964	1981	5.6	2, 7
Benmore	New Zealand	110	2,040	1964	1966	5.0	1, 2, 8
Charvak	Uzbekistan	148	2,000	1971	1977	5.3	25
Eucumbene	Australia	116	4,761	1957	1959	5.0	2
Geheyan	China	151	3,400	1993	1997	VI	22a
Hoover	USA	221	36,703	1935	1939	5.0	1, 2, 10
Marathon	Greece	67	41	1929	1938	5.7	1, 2, 4,5
Oroville	USA	236	4400	1967	1975	5.7	2, 11
Srinagarind	Thailand	140	11,750	1977	1983	5.9	28
Warna	India	80	1,260	1985	1993	5.0	24
Sites where earthquakes having magnitude between 4.0 and 4.9 were triggered							
Aksombo Main	Ghana	134	148,000	1964	1964	V	2, 9
Bajina Basta	Yugoslavia	90	340	1966	1967	4.5–5.0	2, 5
Bhatsa	India	88	947	1981	1983	4.9	20
Bratsk	Russia	100	169		1996	4.2	22b
Camarillas	Spain	49	37	1960	1964	4.1	2, 4, 5
Canelles	Spain	150	678	1960	1962	4.7	2, 4, 5
Capivari–Cachoeira	Brazil	58	180	1970	1971	VI	17
Clark Hill	USA	60	3517	1952	1974	4.3	2, 12
Dahua	China (PRC)	74.5	420	1982	1993	4.5	26
Danjiangkou	China (PRC)	97	16,000	1967	1973	4.7	16
Foziling	China (PRC)	74	470	1954	1973	4.5	16
Grandwal	France	88	292	1959	1963	V	1, 2, 4,5
Hoa Binh	Vietnam	125		1988	1989	4.9	22c
Kastraki	Greece	96	1000	1968	1969	4.6	2
Kerr	USA	60	1505	1958	1971	4.9	1, 2, 9
Komani	Albania	130	1600	1985	1986	4.2	27
Kurobe	Japan	186	149	1960	1961	4.9	2,13
Lake Baikal	Russia					4–4.8a	23
Lake Pukaki	New Zealand	106	9000	1976	1978	4.6	21
Manicouagan 3	Canada	108	10,423	1975	1975	4.1	2
Marimbondo	Brazil	94	6150	1975	1975	IV	18
Monteynard	France	155	275	1962	1963	4.9	1, 2, 4,5
Nurek	Tadjikistan	317	1000	1972	1972	4.6	1, 2, 14
P. Colombia/ V.Grande	Brazil	40/56	1500/2300	1973–1974	1974	4.2	19
Piastra	Italy	93	13	1965	1966	4.4	2, 4, 5
Pieve de Cadore	Italy	116	69	1949	1950	V	2, 15
Shenwo	China (PRC)	50	540	1972	1974	4.8	16
Vouglans	France	130	605	1968	1971	4.4	2, 4, 5
Karun-III	Iran	185	2970	2005	2005	4.3	29

References: 1 = (Gupta and Rastogi, 1976); 2 = (Packer et al., 1979); 3 = (Shen et al., 1974); 4 = (Rothe, 1970, 1973); 5 = (Bozovic, 1974); 6 = (Gough and Gough, 1970b); 7 = (Toppozada, 1982); 8 = (Adams, 1974); 9 = (Simpson, 1976); 10 = (Carder, 1945); 11 = (Bufe et al., 1976); 12 = (Talwani, 1976); 13 = (Hagiwara and Ohtake, 1972); 14 = (Soboleva and Mamadaliev 1976); 15 = (Caloi, 1970); 16 = (Oike and Ishikawa, 1983); 17 = (Berrocal (personal communication), 1990); 18 = (Veloso et al., 1987); 19 = (Berrocal et al., 1984); 20 = (Rastogi et al., 1986a); 21 = (Reyners, 1988); 22a = (Chen et al., 1996), 22b = (Pavlenov and Sherman, 1996), 22c = (Tung,1996), 23 = (Djadkov, 1997), 24 = (Rastogi et al., 1997b); 25 = (Plotnikova et al., 1992); 26 = (Guang, 1995); 27 = (Muco, 1991b); 28 = (Chung and Liu, 1992); 29 = (Abas Kangi and Nematollah Heidari, 2008) (updated from Gupta 2002)



Artificial Water Reservoir Triggered Earthquakes, Figure 2 Height of water column is the most important correlate (After Stuart-Alexander and Mark, 1976).



Artificial Water Reservoir Triggered Earthquakes, Figure 3 Impoundment of water reservoir increases heterogeneity. (**a**): Homogeneous rock volume before impoundment. (**b**): After impoundment. For **a**, earthquake sequence falls under Type 1 of Mogi's (1963) models: no foreshocks and a few aftershocks. For **b**, earthquake sequence falls under Type 2 of Mogi's (1963) models: considerable foreshock activity. For details see text.

The above-mentioned observations are governed by the mechanical properties of the media, and their deviation from the normal implies changes in these properties consequent to impoundment of the artificial water reservoir. It can be best illustrated by a sketch shown in Figure 3. "A" in this figure is a homogenous media rock volume. When the stress exceeds the strength of the rock, there would be a major event releasing most of the strain, followed by peripheral adjustment aftershocks. In such a sequence, there would not be any foreshocks before the major event. The aftershock activity would be over in a short time, the ratio of the largest aftershock to the main event would be low, and the *b* value would be low. This is typically the situation with the earthquake sequences in stable continental regions not associated with the reservoir loading. Due to filling of the water reservoir, the heterogeneity of the media increases ("B" in Figure 3), and the rock volume gets fragmented, and the accumulated stress is released by smaller rock volumes. In such a situation, the events would start occurring as and when the strength of an individual rock volume is exceeded. The main event would correspond to the largest rock volume and there would be events before it and after it, changing the pattern from Type I of Mogi's Model to Type II. The ratio of the magnitude of the largest aftershock to the main shock would be high, and the *b* value of the foreshock sequence as well as the aftershock sequence would be high. This is what is observed with RTS sequences.

Mechanism of triggered earthquakes

In the following we give a gist of major milestones in comprehending the phenomenon of triggered earthquakes.

The foundation of understanding the phenomenon of triggered earthquakes was laid by the study of the waste-fluid-injection-induced earthquakes in the vicinity of Rocky Mountain Arsenal Well near Denver, Colorado, USA, in early 1960s (Evans, 1966). There are three main effects of reservoir loading relevant to triggering of earthquakes as pointed out by Gough and Gough (1970b) and several others:

1. The elastic stress increase following filling of the reservoir
2. The increase in pore fluid pressure in saturated rocks, basically due to decrease in pore volume due to compaction, in response to increase in elastic stress
3. Pore-pressure changes related to fluid migration

Gough and Gough (1970a, b) provided the first definitive quantitative analysis of the role of the load of the water reservoir in triggering earthquakes at Lake Kariba. The role of reservoir load was also considered by Bell and Nur (1978) and Roeloffs (1988). They pointed out that reservoir-load-induced stresses at seismogenic depths are very small and can only perturb the ambient stress field. Gupta et al. (1972a, b) identified the rate of increase of reservoir water levels, maximum water levels reached, and the duration of retention of high water levels as factors affecting the frequency and magnitude of triggered earthquakes. The influence of pore fluid pressure in inducing earthquakes in very simple 1 D reservoir models was presented by Snow (1972). More sophisticated models were dealt by Withers and Nyland (1976), Bell and Nur

(1978), and Roeloffs (1988) based on Biot's (1941) consolidation theory, which later on is generalized by Rice and Cleary (1976) by recognizing that the fluids too may be compressible. Very interesting results related to modeling of pore-pressure diffusion have been reported for Acu Reservoir in Brazil (Do Nascimento et al., 2004).

Pore-pressure diffusion plays a very important role in the triggering of the earthquakes. However, there are a very few direct measurements of diffusivity, and it is mostly inferred from the temporal migration of seismicity. Talwani et al. (1999) have reported in situ measurements of hydrological diffusivity and Skempton's coefficient at the Bad Creek Reservoir in South Carolina, USA. At an observation well located 250 m away from the reservoir, a change in the water level of the well had a direct correspondence with the changes in the water level of the reservoir, initially with a delay of 98 h, which later stabilized at 72 h. This led to a frequency independent estimate of diffusivity of $\sim$0.076 $m^2\ s^{-1}$ and Skempton's coefficient of 0.66 for an undrained response of the reservoir. Later, Talwani et al. (2007) analyzed more than 90 case histories of triggered seismicity and found diffusivity to vary between 0.1 and 10 $m^2\ s^{-1}$. This range of diffusivity values corresponds to a range of intrinsic permeability between 5×10^{-16} and 5×10^{-14} m^2. Talwani et al. (2007) referred this range of the permeability of fractures as seismogenic permeability. Seismogenic permeability is an intrinsic property of fractures where pore-pressure diffusion is associated with seismicity.

The in situ measurements of physical properties and examination of physical mechanism controlling triggered seismicity at Monticello Reservoir, South Carolina, USA, provided the much needed field verification of the theoretical and model developments of concept of triggered seismicity (Zobak and Hickman, 1982). The effect of changes in lake levels and other derived parameters on triggered earthquakes have been dealt by Simpson and Negmatullaev (1981) for the Nurek Dam, and by Gupta (1983) for the Koyna Dam. The part played by pore-pressure diffusion in triggering earthquakes has been dealt by Talwani and Acree (1984/1985). The effect of inhomogeneities in rock properties on triggering earthquakes was addressed by Simpson and Narasimhan (1990).

Most of the theoretical models discussed the effect of pore fluid pressure in isotropic rocks. Chen and Nur (1992) point out that deviatory effects of pore fluid pressure in anisotropic rocks have wide applications in comprehending triggered seismicity, earthquake precursors, and aftershocks. This approach needs to be applied to a few cases of RTS.

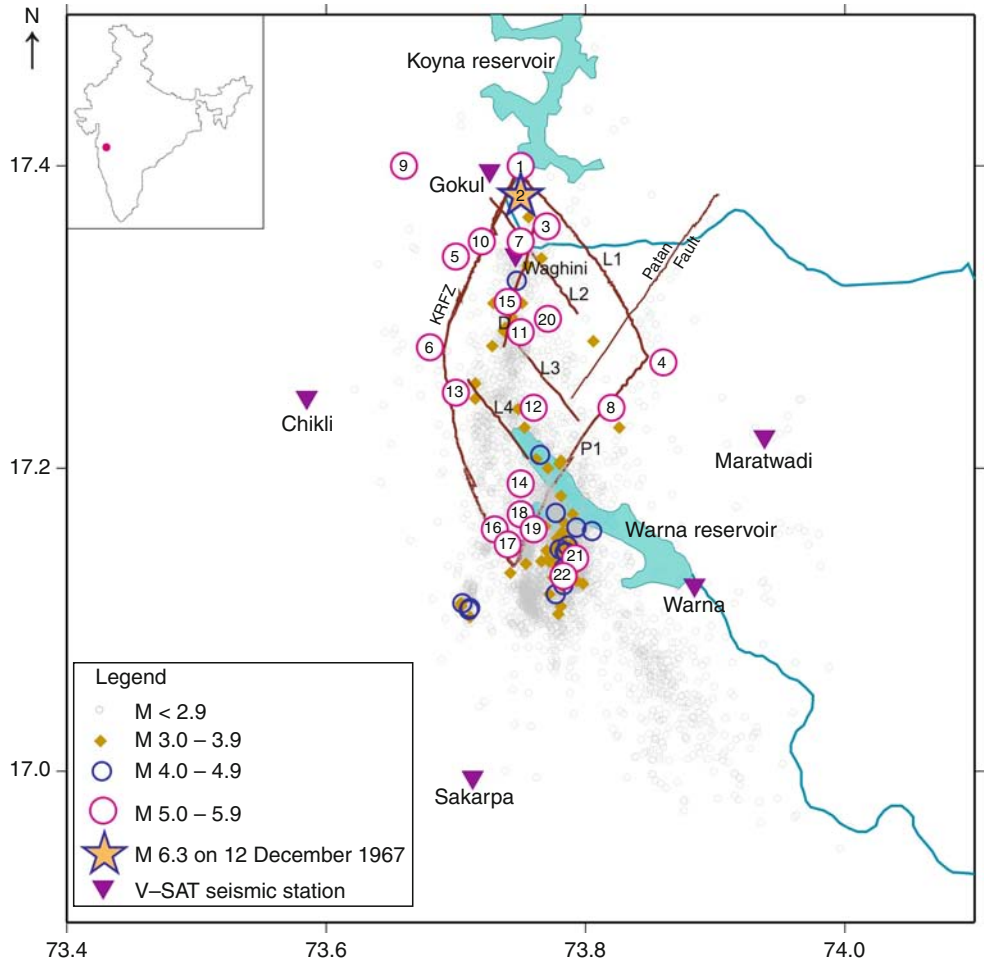
Other interesting studies include the notion of fault stability as a measure of the interplay of frictional stresses mobilized and the resolved shear stresses acting on a fault Chander and Kalpna (1997). Kalpna and Chander (2000) have developed an algorithm for simulation of stresses and pore pressure for more realistic laterally finite 3D models of reservoirs.

Koyna, India

Koyna reservoir located close to the west coast of India continues to be the most significant site of triggered earthquakes. Earthquakes began to occur soon after the impoundment of the Shivaji Sagar Lake created by the Koyna Dam in 1962. So far, globally, the largest triggered earthquake of M 6.3 on December 10, 1967; 22 earthquakes of $M \geq 5$; about 200 earthquakes of $M \geq 4$; and several thousand smaller earthquakes have occurred in the region. Talking to residents in the region in the 1960s revealed that they had not experienced any earthquake in the region in their living memory. There is a seismic station operating at Pune, about 120 km away from Koyna. A scrutiny of the seismograms revealed no earthquakes that could be assigned to the Koyna region. Another water reservoir, Warna, was impounded in 1985. This reservoir is located 35 km SSE of Koyna (Figure 4). After the impoundment of Warna reservoir, the triggered activity got enhanced. For convenience, we shall call the Koyna and the Warna reservoirs region as the Koyna region. Major bursts of seismic activity associated with the Koyna reservoir occurred in 1967, 1993, 1980, 1993–1994, 2005, and 2009.

How long triggered earthquakes will continue at Koyna?

Koyna is indeed a unique site where triggered earthquakes have been occurring since the impoundment of the reservoir in 1962. Gupta et al. (2002) have examined in detail the question as to how long the triggered earthquakes would continue at Koyna. The maximum credible earthquake for the Indian shield region has been estimated to be M 6.8. It is hypothesized that the region between Koyna and Warna was stressed close to critical before the impoundment of the Koyna reservoir, and was capable of generating an M 6.8 earthquake. As demonstrated through the study of b values in earthquake magnitude-frequency relation, foreshock–aftershock patterns, the ratio of the magnitude of the largest aftershock to the main shock and the decay of aftershock activity in earthquake sequences at Koyna, the heterogeneity of the media has increased. None of the $M \sim 5$ earthquake has occurred at the same location. $M \sim 5$ earthquakes occur in Koyna region when the previous water maximum in the reservoir has been exceeded. Long time ago, Kaiser (1953) had reported that acoustic emission, under monotonically increasing stress, shows an appreciable increase after the applied stress exceeds the previously applied stress maxima. This approach has been successfully used by Yoshikawa and Mogi (1981) to estimate crustal stresses from cored samples. For the Nurek Dam also, it was reported that major triggered events occurred when water level in the reservoir reached the previous maximum and exceeded it (Simpson and Negmatullaev, 1981). Gupta et al. (2002) had concluded that as of 2002, about one-half of an M 6.8 earthquake energy has been released in the



Artificial Water Reservoir Triggered Earthquakes, Figure 4 Earthquakes in Koyna region of Western India.

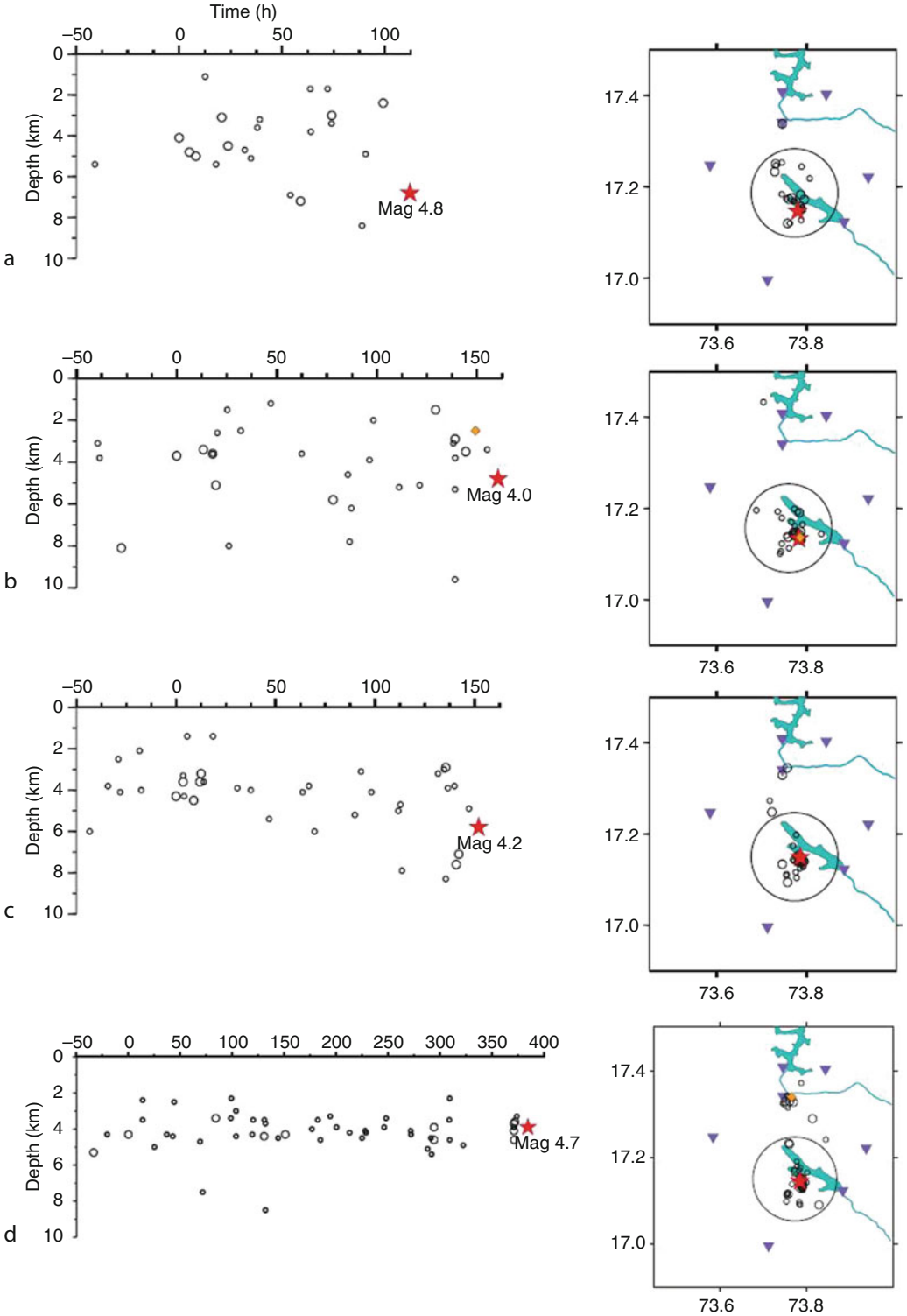
Artificial Water Reservoir Triggered Earthquakes, Table 2 Events preceding M $\geq$ 4.0 earthquakes in Koyna region

Main earthquake	Duration of nucleation period (h)	No. of events before the main earthquake				
		M 1.0–1.9	M 2.0–2.9	M 3.0–3.9	Largest earthquake	50 h prior to beginning of nucleation period
30th August 2005 M 4.8	110	13	8	0	2.6	2
13th November 2005 M 4.0	160	18	8	1	3.0	3
26th December 2005 M 4.2	150	22	8	0	2.9	6
17th April 2006 M 4.7	380	40	10	0	2.7	2

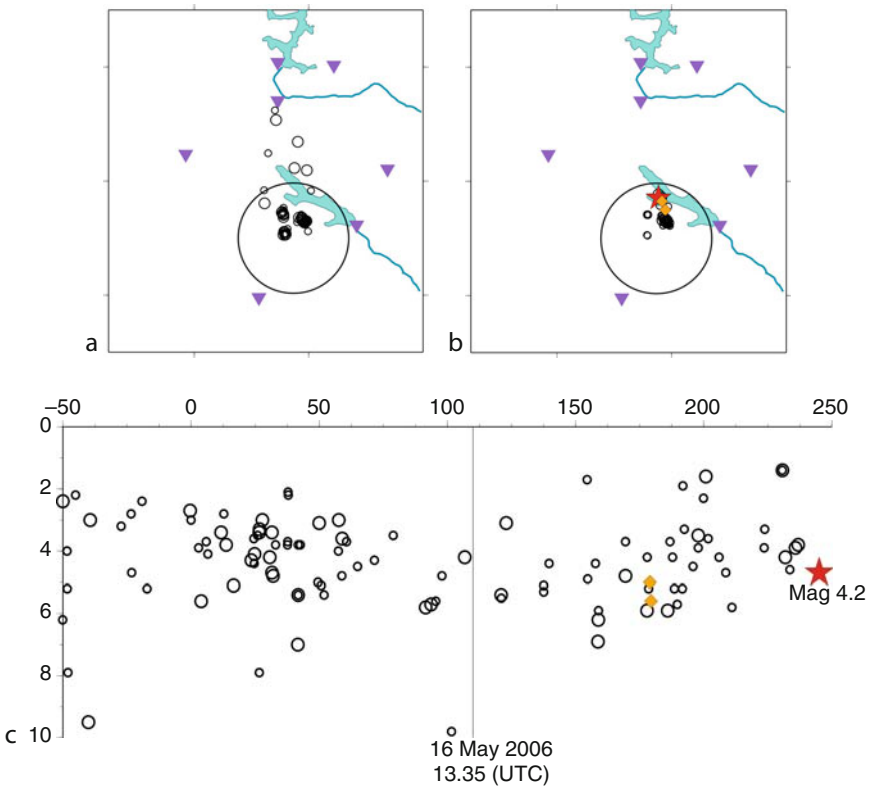
Koyna region, and the activity should continue for another 3–4 decades. Due to increase in heterogeneity, no fault segment long enough to generate an M > 6 earthquake is left intact; so an earthquake like the December 10, 1967, may not occur. The occurrence of M ~ 5 earthquakes would be governed by the fact whether the previous water maximum has been exceeded, and other factors such as the rate of loading, the highest water levels reached, and the duration of the retention of high water levels.

Short-term earthquake forecast at Koyna

A case was made that short-term earthquake forecast may be feasible at Koyna (Gupta, 2001). This was based on the fact that the shallow (depth $\leq$ 10 km) seismic activity in the vicinity of the Koyna is confined to an area of 20 $\times$ 30 km^2, and there is no other seismic source within 50 km radius. Every year, following the rainy season, the reservoirs get filled and there is an enhancement of seismic



Artificial Water Reservoir Triggered Earthquakes, Figure 5 Seismic activity for the period 50 h before the start of nucleation till the occurrence of M $\geq$ 4 earthquakes in Koyna region during August 2005–April 2006. Foreshocks clustering before the main shocks in a region of 10 km radius (*Right*). Temporal distribution of events within the circle with the depth. ◦, ○, ♦ and ★ are events in M 1.0–1.9, 2.0–2.9 and 3.0–3.9 ranges, and main earthquake respectively (*Left*). Onset of identified nucleation period is given "0" time mark. Nucleation period lasted for 110, 160, 150 and 380 h for (**a**) 30th August, (**b**) 13th November and (**c**) 26th December 2005 (**d**) 17th April 2006 earthquakes respectively (Table 2).



Artificial Water Reservoir Triggered Earthquakes, Figure 6 Symbols as in Figure 5. (**a**), Events that occurred during 107 h (11–16 May) preceding identification of nucleation. (**c**), Temporal-depth plot of the events within the 10 km radius circle. Nucleation was inferred to have started 107 h before, where 0 h is put on the time axis. (**b**), Seismic activity from 13:35 UTC on 16 May till the occurrence of the M4.2 earthquake on 21st May 2006, on the right side of (**c**).

activity. From among earthquake precursors, foreshocks are important and have a potential in forecasting earthquakes (Dodge et al., 1996; Helmestetter et al., 2003). Ohnaka (1992) has noted that immediate foreshock activity is a part of the nucleation process leading to main-shock dynamic rupture. Gupta et al. (2007) have reported successful earthquake forecast in Koyna based on identification of nucleation in real time. The seismic activity in the Koyna region is monitored by a closely spaced network of seven modern seismic stations. Table 2 from Gupta et al. (2007) is the learning phase where four earthquakes of M 4.0 to M 4.8 were found to be preceded by well-defined nucleation where several events occurred in a tight cluster of less than 10 km radius (Figure 5a–d). It was realized that if the formation of nucleation could be identified in real time before the occurrence of the main shock, it may be possible to make a short time forecast.

During the middle of May 2006, an interesting situation developed. The events in the Koyna region for the period 11–16th May, 2006, are depicted in Figure 6a (Gupta et al., 2007). By the afternoon of 16th May, 2006, some 50 events of $M \geq 1.0$, the largest being M 2.7, had occurred in the preceding 107 h in a small area, the focal depth being between 2 and 8 km. It was inferred that the region is going through a nucleation phase. Based on the experience of previous nucleation, the following forecast was made at 19:05 IST and communicated to the Editor of *Current Science*, Secretary, Ministry of Earth Sciences of the Government of India, and the President of the Geological Society of India: "On the basis of the data available from seven seismic stations operating in the Koyna region, we have identified a nucleation which started on 12th May 2006. This may lead to the occurrence of an $M \sim 4$ earthquake in the next 15 days. This shallow earthquake (focal depth less than 8 km) will occur within

Artificial Water Reservoir Triggered Earthquakes, Table 3 Forecast of 21st May, 2006, M 4.2 earthquake in Koyna on 16th May, 2006

	Forecast parameters	Occurrence
Epicenter	Within 10 km radius of 17.1°N, 73.8°E	17.171°N, 73.777°E
Magnitude	~4	4.2
Time	Within 15 days of 16th May 2006, i.e., until 31st May 2006	20:29:01 UTC On 21st May 2006
Focal depth	Less than 8 km	4.7 km

a radius of 10 km centered at 17.1°N, 73.8°E. On the basis of our previous experience of studying nucleation-preceding earthquakes in the Koyna region, we expect this earthquake to occur over the next 15 days time (till 31st May, 2006), with a 50% probability." An earthquake of M 4.2 occurred on 21st May. Table 3 gives the comparison of forecasted parameters and that of 21st May, 2006, earthquake. Since 2005, six similar short-term forecasts have been made in the Koyna region and all have come true.

Summary

In this article, we provide a thumbnail review of studies of global occurrence of artificial water reservoir triggered earthquakes, with a special emphasis on Koyna, India. Considering the small changes in the stress regime caused by filling of the deepest reservoirs compared to the stress drop of the associated earthquakes, it is appropriate to replace "reservoir-induced seismicity" by "reservoir triggered seismicity."

Koyna continues to be the most significant sight of RTS globally. The latest $M > 5$ earthquake occurred on December 12, 2009 (M 5.1).

At Koyna, earthquakes of M 4–5 are often preceded by well-defined clusters of foreshocks of $M \leq 3$, referred as nucleation that is found to last typically 100–400 h. Identification of nucleation in real time has led to successful short-term forecasts of $M \sim 4$ earthquakes.

Study of RTS provides exceptionally good input to comprehend physics of earthquakes, finding safer sites of creating artificial water reservoirs and bringing us closer to accurate short-term earthquake forecasts.

Bibliography

Allen, C. R., 1978. Evaluation of seismic hazard at the Auburn damsite. California, U.S. Bureau of Reclamation Report, Denver, p. 10.

Allen, C. R., 1982. Reservoir-induced earthquakes and engineering policy. *California Geology*, **35**, 248–250.

Baecher, B. G., and Keeney, R. L., 1982. Statistical examination of reservoir induced seismicity. *Bulletin. Seismological Society of America*, **72**, 553–569.

Bell, M. L., and Nur, A., 1978. Strength changes due to reservoir-induced pore pressure and stresses and application to Lake Oro-Berrocal. *Journal of Geophysical Research*, **83**, 4469–4483.

Biot, M. A., 1941. General theory of three-dimensional consolidation. *Journal of Applied Physics*, **12**, 155–164.

Carder, D. S., 1945. Seismic investigations in the Boulder Dam area, 1940–1944, and the influence of reservoir loading on earthquake activity. *Bulletin. Seismological Society of America*, **35**, 175–192.

Chander, R., and Kalpna, R., 1997. On categorizing induced and natural tectonic earthquakes near new reservoirs. *Engineering Geology*, 46.

Chen, Q., and Nur, A., 1992. Pore fluid pressure effects in an isotropic Rocks: mechanisms of induced seismicity and weak faults. *Pure and Applied Geophysics*, **139**, 463–480.

Do Nascimento, A. F., Cowie, P. A., Lunn, R. J., and Pearce, R. G., 2004. Spatio-temporal evolution of induced seismicity at Acu reservoir, NE Brazil. *Geophysical Journal International*, **158**, 1041–1052.

Dodge, D. A., Beroza, G. C., and Ellsworth, W. L., 1996. Detailed observations of California foreshock sequences: implications for the earthquake initiation process. *Journal of Geophysical Research*, **101**, 22371–22392.

Evans, M. D., 1966. Man made earthquakes in Denver. *Geotimes*, **10**, 11–17.

Gough, D. I., and Gough, W. I., 1970a. Stress and deflection in the lithosphere near Lake Kariba, 1. *Geophysical Journal*, **21**, 65–78.

Gough, D. I., and Gough, W. I., 1970b. Load induced earthquakes at Kariba. 2. *Geophysical Journal of the Royal Astronomical Society*, **21**, 79–101.

Gupta, H. K., 1983. Induced seismicity hazard mitigation through water level manipulation at Koyna, India: a suggestion. *Bulletin. Seismological Society of America*, **73**, 679–682.

Gupta, H. K., 2001. Short-term earthquake forecasting may be feasible at Koyna, India, Elsevier. *Tectonophysics*, **338**(3–4), 353–357.

Gupta, H. K., 2002. A review of recent studies of triggered earthquakes by artificial water reservoirs with special emphasis on earthquakes in Koyna, India. *Earth Science Reviews*, **58**, 279–310.

Gupta, H. K., Rastogi, B. K., and Narain, H., 1972a. Common features of the reservoir associated seismic activities. *Bulletin. Seismological Society of America*, **62**, 481–492.

Gupta, H. K., Rastogi, B. K., and Narain, H., 1972b. Some discriminatory Characteristics of earthquakes near the Kariba, Kremasta and Koyna artificial lakes. *Bulletin. Seismological Society of America*, **62**, 493–507.

Gupta, H. K., Mandal, P., and Rastogi, B. K., 2002. How long will triggered earthquakes at Koyna, India continue? *Current Science*, **82**(2), 202–210.

Gupta, H. K., et al., 2007. Earthquake forecast appears feasible at Koyna, India. *Current Science*, **93**(6), 843–848.

Helmestetter, A., Sornette, D., and Grasso, J. R., 2003. Mainshocks are aftershocks of conditional foreshocks: How do foreshock statistical properties emerge from aftershock laws? *Journal of Geophysical Research*, **108B**, 2046.

Hudyma, M. R., and Potvin, Y., 2005. Mining-Induced Seismicity in Underground Mechanised Hard rock Mines-Results of a World–Wide Survey. "*Seminar on Advanced Geomechanics in Mines*", 03 August 2005. Perth: Australian Centre for Geomechanics, pp. 47.

Kaiser, J., 1953. Erkenntnisse and Folgerungen aus der Messung von Gerauschen bei Zugbeanspruchung von metallischen Werkstoffen. *Archiv für das Eisenhüttenwesen*, **24**, 43–45 (in German).

Kalpna, and Chander, R., 2000. Green's function based stress diffusion solutions in the porous elastic half space for time varying finite reservoir loads. *Physics of the Earth and Planetary Interiors*, **120**, 93–101.

McGarr, A., and Simpson, D., 1997. Keynote lecture: A broad look at induced and triggered seismicity, "Rockbursts and seismicity in mines." In Gibowicz, S. J., and Lasocki, S. (eds.), *Proceedings of 4th International Symposium on Rockbursts and Seismicity in Mines Poland*, 11–14 Aug 1997. Rotterdam: A. A. Balkema Press, pp. 385–396.

McGarr, A., Simpson, D., and Seeber, L., 2002. Case histories of induced and triggered seismicity. *International Handbook of Earthquake and Engineering Seismology.* Orlando, FL: Academic Press, Vol. 81 A, pp. 647–661.

Mogi, K., 1963. Some discussions on aftershocks, foreshocks and earthquake swarms – The fracture of a semi infinite body caused by an inner stress origin and its relation to earthquake phenomena. *Bulletin. Earthquake Research Institute*, **41**, 615–658.

Ohnaka, M., 1992. Earthquake source nucleation: a physical model for short-term precursors. *Tectonophysics*, **211**, 249–278.

Rice, J. R., and Cleary, M. P., 1976. Some basic stress diffusion solutions for fluid-saturated elastic porous media with compressible constituents. *Reviews of Geophysics and Space Physics*, **14**(2), 227–242.

Roeloffs, E. A., 1988. Fault stability changes induced beneath a reservoir with cyclic variations in water level. *Journal of Geophysical Research*, **93**(B3), 2107–2124.

Shen, C., Chang, C., Chen, H., Li, T., Hueng, L., Wang, T., Yang, C., and Lo, H., 1974. Earthquakes induced by reservoir impounding and their effect on the Hsinfengkiang Dam. *Scientia Sinica*, **17**(2), 239–272.

Simpson, D. W., and Leith, W. S., 1988. Induced seismicity at Toktogul Reservoir, Soviet Central Asia. U.S. Geological Survey, No. 14-08- 0001-G1168, 32 pp.

Simpson, D. W., and Narasimhan, T. N., 1990. Inhomogeneities in rock Properties and their influence on reservoir induced seismicity. *Gerlands Beitrage zur Geophysik*, **99**, 205–219.

Simpson, D. W., and Negmatullaev, S. K., 1981. Induced seismicity at Nurek Reservoir, Tadjikistan, USSR. *Bulletin. Seismological Society of America*, **71**(5), 1561–1586.

Snow, D. T., 1972. Geodynamics of seismic reservoirs. Proceedings of Symposium on Percolation Through Fissured Rocks, Deutsche Gessellschaft fiir Erd- und Grundbau, Stuttgart, Germany, T2-J, pp. 1–19.

Stuart-Alexander, D. E., and Mark, R. K., 1976. Impoundment-induced seismicity associated with large reservoirs, U.S. Geological Survey Open-File Rept. pp. 76–770.

Talwani, P., and Acree, S., 1984/1985. Pore pressure diffusion and the mechanism of reservoir-induced seismicity. *Pure and Applied Geophysics*, **122**, 947–965.

Talwani, P., Cobb, J. S., and Schaeffer, M. F., 1999. In situ measurements of hydraulic properties of a shear zone in northwestern South Carolina. *Journal of Geophysical Research*, **104**(B7), 14993–15003.

Talwani, P., Chen, L., and Gahalaut, K., 2007. Seismogenic permeability, k_s. *Journal of Geophysical Research*, **112**, B07309, doi:10.1029/2006JB004665.

Withers, R. J., and Nyland, E., 1976. Theory for the rapid solution of ground subsidence near reservoirs on layered and porous media. *Engineering Geology*, **10**, 169–185.

Yoshikawa, S., and Mogi, K., 1981. A new method for estimating the crustal stress from cored rock samples; laboratory study in the case of uniaxial compression. *Tectonophysics*, **74**, 323–339.

Zoback, M. D., and Hickman, S., 1982. Physical mechanisms controlling induced seismicity at Monticello Reservoir, South Carolina. *Journal of Geophysical Research*, **87**, 6959–6974.

Cross-references

Earthquake Precursors and Prediction
Earthquake Prediction, M8 Algorithm
Earthquake, Aftershocks
Earthquake, Foreshocks
Earthquakes, Energy
Seismicity, Intraplate
Seismological Networks

B

BIOGEOPHYSICS

Lee Slater[1], Estella Atekwana[2]
[1]Department of Earth and Environmental Sciences, Rutgers University Newark, Newark, NJ, USA
[2]Boone Pickens School of Geology, Oklahoma State University, Stillwater, OK, USA

Definition

Biogeophysics. Sub-discipline of exploration geophysics focusing on the geophysical signatures resulting from microbial interactions with geologic media.

Introduction

Geophysical imaging techniques have the potential to measure not just the subsurface physical and chemical properties, as geophysics is conventionally used, but also microbes, microbial processes, and microbe-mineral interactions. "Biogeophysics" is defined here as a rapidly evolving discipline of exploration geophysics concerned with the geophysical signatures of microbial interactions with geologic media that combines the fields of Microbiology, Biogeoscience, and Geophysics (Atekwana and Slater, 2009) (Figure 1). Within this context, biogeophysics examines the links between dynamic subsurface microbial processes, microbial-induced alterations to geologic materials, and geophysical signatures. We note that the term biogeophysics is also used in other disciplines (a) to describe research into the origins of life, and (b) to describe fluxes of energy, water, and momentum between earth surface and the atmosphere. Our use of the term describes how it has been adopted by the exploration geophysics community in recognition of the geophysical signatures of microbial activity.

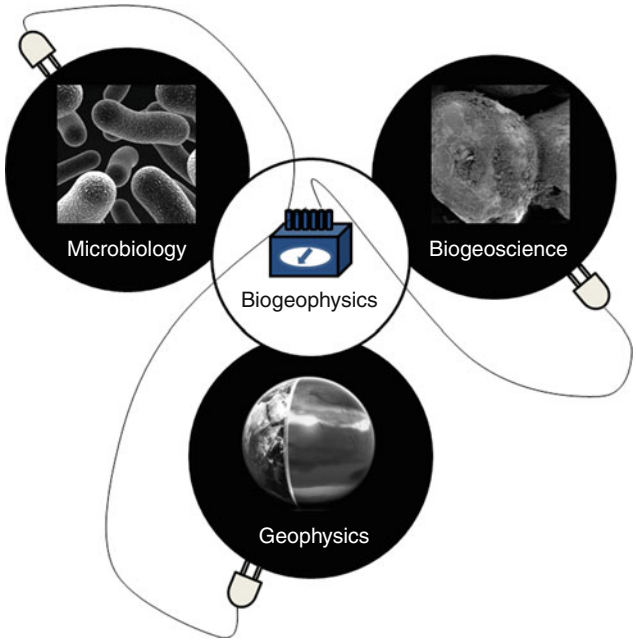
These geophysical signatures may arise from, (1) microbial cells and extracellular structures themselves, (2) growth of microorganisms (production of biomass) and biofilm formation, (3) generation of metabolic by-products and the interactions of these metabolites with the host material, and (4) microbially mediated processes. The geophysical signatures arising from each of these four source mechanisms are described below.

Geophysical detection of cells and biofilms

Microbial cells and biofilms exhibit distinct electrical properties, and certain (magnetotactic) bacteria also display unique magnetic characteristics. The membrane potential (the voltage drop across the membrane due to the negatively charged molecules inside cells) of live cells results in an accumulation of mobile electric-charge carriers at membrane surfaces. When live cells are placed in time-oscillating electric fields, these charges move on the surface of the membrane, giving rise to high polarizations for cellular suspensions that are readily measured with dielectric spectroscopy. As the mobility of these surface charges is relatively small, this effect is manifested at low frequencies, such that the relative dielectric permittivity of live-cell suspensions can be as high as 10^6 (Prodan et al., 2004; Stoy et al., 1982). The outer and inner cell radii, diffusion constants, and membrane potential all have a very distinct effect on broadband dielectric spectroscopy data (Prodan et al., 2008). Polarization enhancement is also observed when microbial cells are present in high concentrations in porous media (Ntarlagiannis et al., 2005), although additional polarization mechanisms associated with mineral surfaces may obscure the signal. This polarization is enhanced when cells are more preferentially adsorbed onto mineral surfaces (Abdel Aal et al., 2009).

The magnetic properties of soils and rock are altered by a diverse group of prokaryotes that exert a significant control over magnetite formation in two ways that differ mechanistically: biologically induced mineralization

Harsh K. Gupta (ed.), *Encyclopedia of Solid Earth Geophysics*, DOI 10.1007/978-90-481-8702-7,



Biogeophysics, Figure 1 Primary disciplinary themes encompassing biogeophysics.

(BIM) and biologically controlled mineralization (BCM) (Jimenez-Lopez et al., 2010). Biologically induced magnetite formation is indistinguishable from magnetite formed by inorganic processes; however, biologically controlled magnetite formation is distinct and produced by magnetotactic bacteria (Jimenez-Lopez et al., 2010). These magnetotactic bacteria biomineralize intracellular, membrane-bound, single-magnetic-domain crystals of the magnetic minerals magnetite and/or greigite called magnetosomes (Bazylinski and Frankel, 2004). Magnetotactic bacteria are typically found in a wide variety of aquatic environments (e.g., fresh water lakes) with the highest numbers occurring mostly at oxic–anoxic interfaces. The presence of magnetotactic bacteria in sediments is often used as a proxy for paleoenvironmental/climatic conditions, and the presence of magnetosomes in sediments and other secondary magnetic minerals produced by microbial activity likely impacts the magnetic properties of the subsurface recorded with magnetic geophysical techniques as has been demonstrated at hydrocarbon contaminated sites (Rijal et al., 2010).

Biofilms (an attached state of cell growth, whereby cells are closely packed and firmly attached to each other) further alter the geophysical properties of soils. Low-frequency electrical measurements respond to self-limiting microbial growth/cell attachment and biofilm formation in porous media, as well as death and lyses of cells (Davis et al., 2006). Electrical polarization may result from the electrical properties of the biofilm itself or from the modification of grain surfaces due to cell attachment. Biofilms also alter the acoustic properties of porous media, with biofilm growth in soil columns resulting in spatial complexity of acoustic amplitudes (Davis et al., 2009). Such variations likely arise from a non-uniform distribution of microbial activity or heterogeneity in the biomass distribution and biofilm morphology (e.g., variations in biofilm thickness, roughness, hydration, etc.).

Geophysical detection of metabolic by-products and mineral weathering

Microbial metabolism enhances the weathering of minerals through the attachment, growth, and colonization of mineral surfaces by microorganisms (Bennett et al., 1996). Metabolic by-products, including biogenic gasses (e.g., CO_2, H_2S, CH_4, etc.), organic acids, and biosurfactants, all affect electrical properties of porous media. Microbial production of organic acids and biosurfactants adds ions to solution, increasing electrolyte concentration of pore fluids (Cassidy et al., 2001). Organic acids enhance mobility of sparingly soluble metals and also increase the number of reaction sites, thus accelerating mineral weathering (Hiebert and Bennett, 1992). Enhanced mineral dissolution catalyzed by increased organic acid concentration can lead to physical changes in grain surface morphology, surface area, surface roughness and the generation of secondary porosity and increased permeability (McMahon et al., 1992). These changes in porosity affect acoustic wave propagation by altering grain contact coupling, and changes in surface area/roughness drive changes in electrical polarization. Biogenic gasses reduce bulk electrical conductivity and enhance attenuation of seismic signal amplitudes.

The effects of metabolic by-products on geophysical properties have been recorded in hydrocarbon contaminated environments (Abdel Aal et al., 2004; Werkema et al., 2003). Such sites are natural bioreactors where excess organic substrates stimulate microbial activity. Enhanced mineral weathering in hydrocarbon contaminated aquifers (Hiebert and Bennett, 1992; McMahon et al., 1995), increases pore fluid conductivity and thereby bulk conductivity sensed with a range of electromagnetic methods (Abdel Aal et al., 2004; Atekwana et al., 2004; Sauck et al., 1998). Elevated bulk conductivity is therefore found where intrinsic bioremediation and enhanced mineral weathering are occurring. These zones of highest bulk electrical conductivity may coincide with the highest percentages of oil degrading microbial populations, with spatial heterogeneity in the microbial community structure and shifts in the microbial community concomitant with vertical changes in bulk electrical conductivity (Allen et al., 2007). Biogenic methane production by archea in anaerobic soils can result in extensive free-phase gas production that reduces dielectric permittivity and electrical conductivity (Comas and Slater, 2007).

Geophysical detection of microbially mediated redox processes

Microbial metabolic activity is a critical driver of redox chemistry because microbes derive energy from

oxidation-reduction reactions (the transfer of electrons from one reactant to the other). Terminal electron acceptors (TEAs) govern nutrient utilization by microbes during the breakdown of organic carbon (Cozzarelli et al., 1999). Microbial respiration consequently results in reduced conditions, and strong redox gradients in the Earth typically develop in the presence of heterotrophic bacteria. Significant changes in Eh and pH result in new mineral stability fields in which some minerals become unstable and are dissolved and mobilized, whereas others may precipitate from solution. As TEAs are consumed, changes in pore fluid chemistry may drive changes in pore fluid conductivity, thereby affecting bulk electrical conductivity.

Strong electrical potentials associated with current sources in the Earth are correlated with redox gradients recorded at sites where microbial degradation of hydrocarbons is occurring (Minsley et al., 2007; Naudet et al., 2003). Such large potentials require geobatteries traditionally invoked to explain very large (~1.0 V) potential gradients due to internal current sources recorded over electronically conductive ore bodies straddling the large redox gradient provided by the water table (Sato and Mooney, 1960). Biogeobatteries may occur in conjunction with a strong redox gradient between highly reducing conditions below the water table within a contaminant plume and an oxidized zone above the water table if microbial activity can generate the required electron bridge (Revil et al., 2010). Possible mechanisms facilitating electron migration include iron oxides, clays, and conductive biological materials (Revil et al., 2010). Metal reducing organisms, such as *Shewanella* and *Geobacter*, produce electrically conductive appendages called bacterial nanowires that may facilitate electron transfer to solid-phase electron acceptors (Reguera et al., 2005). However, the ability of biofilms to facilitate electron transport over the scale of the groundwater interface is unknown although new evidence suggests that such electron transfer can take place at millimeter scales (Nielsen et al., 2010). Nonetheless, more studies are needed to confirm this new finding.

Geophysical detection of microbe-mineral transformations

Biogeochemical processes result in mineral precipitation that alters the physicochemical properties of the grain surfaces. When microbial-induced precipitation is extensive, the volumetric physical properties of porous media are modified. In anaerobic environments, iron-reducing bacteria and sulfate-reducing bacteria can use Fe(III) and sulfate respectively as terminal electron acceptors. Ferrous iron produced by iron reducing bacteria promotes the precipitation of electrically conductive secondary minerals such as siderite ($FeCO_3$), magnetite (Fe_3O_4), and goethite (FeOOH) (Fredrickson et al., 1998). Sulfide produced during microbial sulfate reduction can react with iron (II) produced by iron reducing bacteria to precipitate iron sulfide minerals. Strong seismic and electrical signatures are generated as a result of microbe-induced ZnS and FeS precipitation (Williams et al., 2005). Decreases in seismic wave amplitude result from the development of differential elastic moduli associated with accumulation of metal sulfide-encrusted microbes within pores (Williams et al., 2005). Electrical signals result from the formation, movement and dissolution of electronically conductive biominerals that profoundly enhance the polarization of a porous medium. These electrical signals may be diagnostic of both the concentration and the distribution of the biominerals throughout a porous medium (Slater et al., 2007).

Geophysical signals also result when microbial processes involve the precipitation of semi-metallic or non-metallic minerals, e.g., metabolically induced calcite precipitation by bacterial hydrolysis of urea (Ferris et al., 1995). Relative to metallic minerals, smaller electrical signals result from changes in pore volume/pore tortuosity and/or surface area/surface roughness driven by precipitation of non-metallic minerals (Wu et al., 2010). Calcite precipitation induced by bacteria has been shown to form cements in porous media and affect subsurface fluid flow (Ferris et al., 1995). Such cements profoundly change the elastic properties of soils and rocks, particularly when ureolysis is stimulated to form calcite cement that acts to stiffen the soil matrix (DeJong et al., 2010). Shear waves are well suited to monitoring changes in the particle soil matrix due to precipitation as shear velocity (Vs) is largely unaffected by pore fluid composition and directly dependent on void ratio, coordination number (average number of surrounding particles a given particle is in contact with) and confining stress (DeJong et al., 2010). Large changes in shear-wave velocity accompany stiffening result from initial binding of microbes to the soil matrix, suggesting that relatively small volumes of precipitates can generate large geophysical signals.

Summary

Only a decade ago it seemed inconceivable to suggest that microbial processes could potentially impact geophysical signatures. However, today is it clear that geophysical signatures result from a range of microbial processes covering the scale of an individual cell to the scale of contaminant plumes in the Earth. A pressing question in Biogeophysics is how to quantify the geophysical signatures of microbial processes through the development of appropriate modeling frameworks. Success in this venture will likely involve multidisciplinary research between geophysicists, biogeochemists, and microbiologists.

Bibliography

Abdel Aal, G., Atekwana, E., Radzikowski, S., and Rossbach, S., 2009. Effect of bacterial adsorption on low frequency electrical properties of clean quartz sands and iron-oxide coated sands. *Geophysical Research Letters*, **36**(36), L04403, doi: 10.1029/2008GL036196.

Abdel Aal, G. Z. A., Atekwana, E. A., and Slater, L. D., 2004. Effects of microbial processes on electrolytic and interfacial

electrical properties of unconsolidated sediments. *Geophysical Research Letters*, **31**(12), L12505, doi:10.1029/2004gl020030.

Allen, J. P., Atekwana, E. A., Duris, J. W., Werkema, D. D., and Rossbach, S., 2007. The microbial community structure in petroleum-contaminated sediments corresponds to geophysical signatures. *Applied and Environmental Microbiology*, **73**(9), 2860–2870, doi:10.1128/aem.01752-06.

Atekwana, E. A., Atekwana, E. A., Werkema, D. D., Jr., Allen, J., Smart, L., Duris, J., Cassidy, D. P., Sauck, W. A., and Rossbach, S., 2004. Evidence for microbial enhanced electrical conductivity in hydrocarbon-contaminated sediments. *Geophysical Research Letters*, **31**, L23501, doi: 10.1029/2004GL021359.

Atekwana, E. A., and Slater, L., 2009. Biogeophysics: A new frontier in Earth science research. *Reviews of Geophysics*, **47**, RG4004, doi: 10.1029/2009RG000285.

Bazylinski, D. A., and Frankel, R. B., 2004. Magnetosome formation in prokaryotes. *Nature Reviews. Microbiology*, **2**(3), 217–230.

Bennett, P. C., Hiebert, F. K., and Choi, W. J., 1996. Microbial colonization and weathering of silicates in a petroleum-contaminated groundwater. *Chemical Geology*, **132**(1–4), 45–53.

Cassidy, D. P., Werkema, D. D., Sauck, W. A., Atekwana, E., Rossbach, S., and Duris, J., 2001. The effects of LNAPL biodegradation products on electrical conductivity measurements. *Journal of Environmental and Engineering Geophysics*, **6**(1), 47–52.

Comas, X., and Slater, L., 2007. Evolution of biogenic gases in peat blocks inferred from noninvasive dielectric permittivity measurements. *Water Resources Research*, **43**(5), W05424, doi:10.1029/2006wr005562.

Cozzarelli, I. M., Herman, J. S., Baedecker, M. J., and Fischer, J. M., 1999. Geochemical heterogeneity of a gasoline-contaminated aquifer. *Journal of Contaminant Hydrology*, **40**(3), 261–284.

Davis, C. A., Atekwana, E., Slater, L. D., Rossbach, S., and Mormile, M. R., 2006. Microbial growth and biofilm formation in geologic media is detected with complex conductivity measurements. *Geophysical Research Letters*, **33**(18), L18403. 10.1029/2006gl027312.

Davis, C. A., Pyrak-Nolte, L. J., Atekwana, E. A., Werkema, D. D., and Haugen, M. E., 2009. Microbial-induced heterogeneity in the acoustic properties of porous media. *Geophysical Research Letters*, **36**, L21405, doi:10.1029/2009gl039569.

DeJong, J. T., Fritzges, M. B., and Nusslein, K., 2006. Microbially induced cementation to control sand response to undrained shear. *Journal of Geotechnical and Geoenvironmental Engineering*, **132**(11), 1381–1392, doi:10.1061/(asce)1090-0241 (2006) 132:11(1381).

DeJong, J. T., Mortensen, B. M., Martinez, B. C., and Nelson, D. C., 2010. Bio-mediated soil improvement. *Ecological Engineering*, **36**(2), 197–210, doi: 10.1016/j.ecoleng.2008.12.029.

Ferris, F. G., Fratton, C. M., Gertis, J. P., Schultzelam, S., and Lollar, B. S., 1995. Microbial precipitation of a strontium calcite phase at a groundwater discharge zone near rock-creek, British-Columbia, Canada. *Geomicrobiology Journal*, **13**(1), 57–67.

Fredrickson, J. K., Zachara, J. M., Kennedy, D. W., Dong, H. L., Onstott, T. C., Hinman, N. W., and Li, S. M., 1998. Biogenic iron mineralization accompanying the dissimilatory reduction of hydrous ferric oxide by a groundwater bacterium. *Geochimica et Cosmochimica Acta*, **62**(19–20), 3239–3257.

Hiebert, F. K., and Bennett, P. C., 1992. Microbial control of silicate weathering in organic-rich ground water. *Science*, **258**(5080), 278–281.

Jimenez-Lopez, C., Romanek, C. S., and Bazylinski, D. A., 2010. Magnetite as a prokaryotic biomarker: a review. *Journal of Geophysical Research*, **115**, G00G03, doi: 10.1029/2009JG001152.

McMahon, P. B., Chapelle, F. H., Falls, W. F., and Bradley, P. M., 1992. Role of microbial processes in linking sandstone diagenesis with organic-rich clays. *Journal of Sedimentary Petrology*, **62**(1), 1–10.

McMahon, P. B., Vroblesky, D. A., Bradley, P. M., Chapelle, F. H., and Gullett, C. D., 1995. Evidence for enhanced mineral dissolution in organic acid-rich shallow ground-water. *Ground Water*, **33**(2), 207–216.

Minsley, B., Sogade, J., and Morgan, F. D., 2007. Three-dimensional self-potential inversion for subsurface DNAPL contaminant detection at the Savannah River Site, South Carolina. *Water Resources Research*, **43**, W04429, doi:10.1029/2005WR003996.

Naudet, V., Revil, A., Bottero, J.-Y., and Bégassat, P., 2003. Relationship between self-potential (SP) signals and redox conditions in contaminated groundwater. *Geophysical Research Letters*, **30**(21), 2091.

Nielsen, L. P., Risgaard-Petersen, N., Fossing, H., Christensen, P. B., and Sayama, M., 2010. Electric currents couple spatially separated biogeochemical processes in marine sediment. *Nature*, **463**(7284), 1071–1074.

Ntarlagiannis, D., Williams, K. H., Slater, L., and Hubbard, S., 2005. The low frequency electrical response to microbially induced sulfide precipitation. *Journal of Geophysical Research*, **110**, G02009.

Prodan, C., Mayo, F., Claycomb, J. R., and Miller, J. H. J., 2004. Low-frequency, low-field dielectric spectroscopy of living cell suspensions. *Journal of Applied Physics*, **95**(7), 3754–3756.

Prodan, E., Prodan, C., and Miller, J. H., 2008. The dielectric response of spherical live cells in suspension: an analytic solution. *Biophysical Journal*, **95**(9), 4174–4182, doi:10.1529/biophysj.108.137042.

Reguera, G., McCarthy, K. D., Mehta, T., Nicoll, J. S., Tuominen, M. T., and Lovley, D. R., 2005. Extracellular electron transfer via microbial nanowires. *Nature*, **435**, 1098–1101.

Revil, A., Mendonça, C. A., Atekwana, E. A., Kulessa, B., Hubbard, S. S., and Bohlen, K. J., 2010. Understanding biogeobatteries: Where geophysics meets microbiology. *Journal of Geophysical Research*, **115**, 10.1029/2009jg001065.

Rijal, M. L., Appel, E., Petrovský, E., and Blaha, U., 2010. Change of magnetic properties due to fluctuations of hydrocarbon contaminated groundwater in unconsolidated sediments. *Environmental Pollution*, **158**(5), 1756–1762, doi: 10.1016/j.envpol.2009.11.012.

Sato, M., and Mooney, H. M., 1960. The electrochemical mechanism of sulfide self-potentials. *Geophysics*, **25**(1), 226–249.

Sauck, W. A., Atekwana, E., and Nash, M. S., 1998. High conductivities associated with an LNAPL plume imaged by integrated geophysical techniques. *Journal of Environmental and Engineering Geophysics*, **2**(3), 203–212.

Slater, L., Ntarlagiannis, D., Personna, Y. R., and Hubbard, S., 2007. Pore-scale spectral induced polarization signatures associated with FeS biomineral transformations. *Geophysical Research Letters*, **34**(21), L21404, doi:10.1029/2007gl031840.

Stoy, R. D., Foster, K. R., and Schwan, H. P., 1982. Dielectric properties of mammalian tissues from 0.1 to 100 MHz: a summary of recent data. *Physics in Medicine and Biology*, **27**, 501–513.

Werkema, D. D., Jr., Atekwana, E. A., Endres, A. L., Sauck, W. A., and Cassidy, D. P., 2003. Investigating the geoelectrical response of hydrocarbon contamination undergoing biodegradation. *Geophysical Research Letters*, **30**(12), 1647–1651.

Williams, K. H., Ntarlagiannis, D., Slater, L. D., Dohnalkova, A., Hubbard, S. S., and Banfield, J. F., 2005. Geophysical imaging of stimulated microbial biomineralization. *Environmental Science & Technology*, **39**(19), 7592–7600.

Wu, Y., S. Hubbard, Williams, K. H. and Ajo-Franklin, J., 2010. On the complex conductivity signatures of calcite precipitation. *Journal of Geophysical Research*, **115**, G00G04, doi: 10.1029/2009JG001129.

Cross-references

Electrical Properties of Rocks
Electrical Resistivity Surveys and Data Interpretation
Geomagnetic Field, Measurement Techniques
Geomagnetic Field, Theory
Geophysical Well Logging
Magnetic Anomalies, Interpretation
Magnetic Methods, Principles
Propagation of Elastic Waves: Fundamentals
Remanent Magnetism
Seismic Properties of Rocks
Seismic, Viscoelastic Attenuation

BODY WAVES

Mahmoud Mohamed Selim Saleh
Department of Nature and Applied Sciences, Al-Aflaj Community College, AL-Kharj University, Al-Aflaj, Riyadh, Saudi Arabia

Synonyms

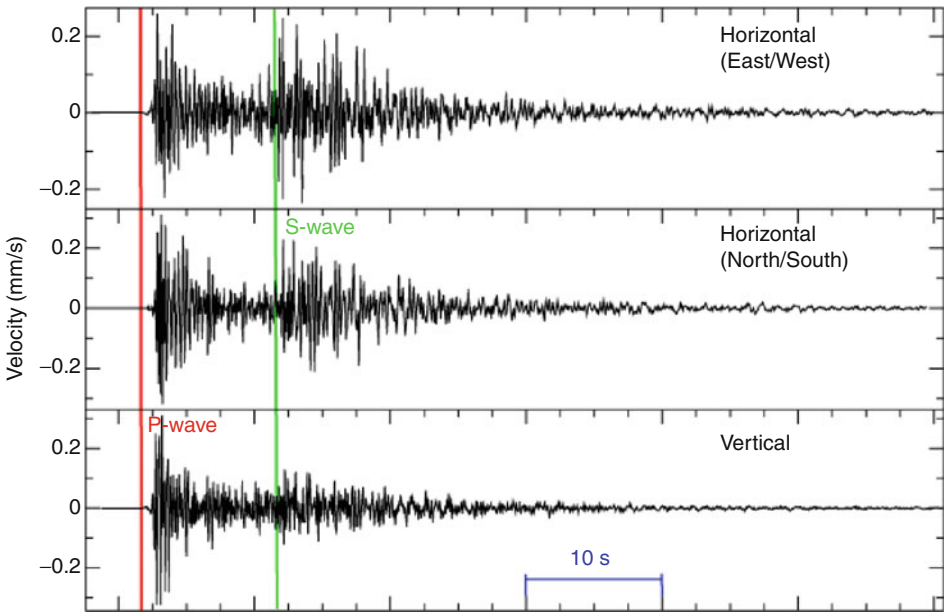
Energy waves generated by an earthquake or an artificial explosion

Definition

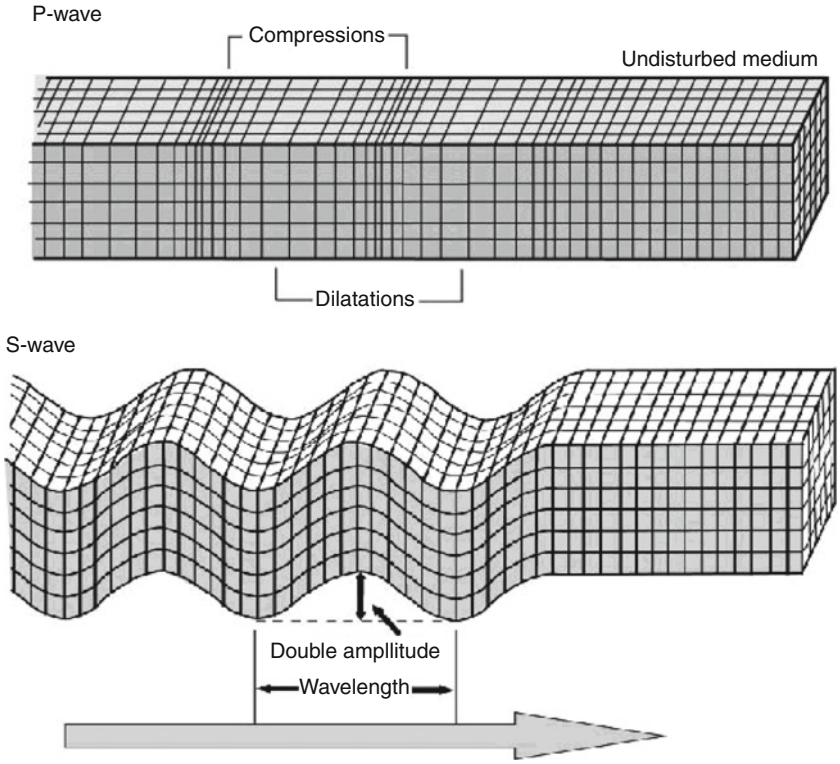
Waves consist of a disturbance in materials (media) that carries energy and propagates. However, the material that the wave propagates in generally does not move with the wave. The movement of the material is generally confined to small motions, called particle motion, of the material as the wave passes. After the wave has passed, the material usually looks just like it did before the wave, and, is in the same location as before the wave. When stress builds up in the Earth to such a level that rapid slip occurs on a fracture (i.e., an earthquake takes place) or when an explosion or mechanical device is used to initiate a seismic disturbance artificially, a complex field of seismic waves is generated. This wave field propagates much like the waves that travel away from the point at which a stone is thrown into a pond. Waves that travel through the interior of the Earth are called body waves. Body waves travel through the interior of the Earth. They follow ray paths bent by the varying density and modulus (stiffness) of the Earth's interior. The density and modulus, in turn, vary according to temperature, composition, and phase. These waves are usually observed at higher frequency.

Types of body waves

There are two main kinds of body waves: P-waves and S-waves, so-named because they are the primary and secondary waves detected by a seismograph (Figure 1). The P-wave is always the first wave to arrive at a seismometer, closely followed by its reflection from the surface and S-waves arrive next. P-waves, or compressional waves, are longitudinal waves (wave motion in the direction the wave is traveling). S-waves are transverse waves or shear waves, involving a back-and-forth shearing motion at right angles to the direction the wave is traveling (Figure 2). Body waves are reflected and transmitted at interfaces where the seismic velocity and/or density changes, obeying Snell's Law. At such an interface, or discontinuity, some of the energy of an incident body wave is reflected as a P-wave, some as an S-wave, some is



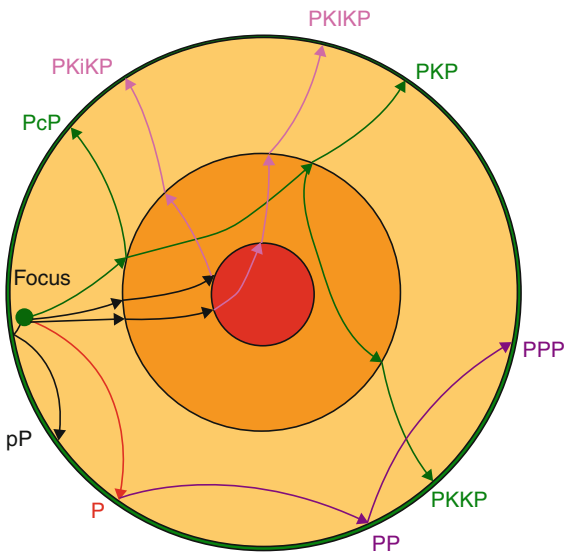
Body Waves, Figure 1 P- and S-wave recorded in a three-component seismograph. P- and S-waves are shown.



Body Waves, Figure 2 P- and S-waves motion. The *arrow* shows the direction that the wave is moving.

transmitted as a P-wave and some as an S-wave. The notation for the various seismic ray paths within the Earth are as follows:

When an earthquake occurs, seismic waves are emitted from the focus (hypocenter); there are several paths that it can take through the Earth before emerging again at the surface. These paths (refer to Figure 3) are symbolized by the letters: p = P-wave arrival from a path that traveled upward from the focus (hypocenter); pP = P-wave arrival from a path that traveled upward from the focus, reflected off the surface of the Earth, then arrived back at the surface; P = P-wave arrival from a path that traveled downward from the focus (hypocenter); PP = P-wave reflected off the surface once; PPP = P-wave reflected off the surface twice; c = a reflection off the outside of the outer surface of the outer core – note that this is the principle cause of multiple arrivals of P- and S-waves right at the epicenter; K = a travel path (refraction) through the outer core; KK = one reflection on the inside outer surface of the outer core; KKK = two reflections off the inside outer surface of the outer core; i = a reflection off the outside of the outer surface of the inner core; I = a travel path (refraction) through the inner core. These letters can be used to indicate the path of a seismic wave through the Earth (Figure 3). For example PKiKP indicates that the wave traveled downward from the focus, refracted through the outer core, reflected off the surface of the inner core, traveled through the outer core, and then traveled through the mantle to arrive at the surface. SKiKS is the same path, but an S-wave. Because liquids have no resistance to shear and cannot sustain a shear wave, S-waves cannot travel through liquid material. The Earth's outer core is believed



Body Waves, Figure 3 P- and S-wave paths in the Earth's interior.

to be liquid because S-waves disappear at the mantle–core boundary, while P-waves do not.

Wave propagation

Seismic body waves propagate through the Earth's interior. Because of the elastic properties of the Earth's materials (rocks) and the presence of the Earth's surface, four main types of seismic waves propagate within the Earth. Compressional (P) and Shear (S) waves propagate through the Earth's interior.

Waves on a seismogram

If we look at a seismogram, we expect to see the first wave to arrive to be a P-wave (the fastest), then the S-wave, and finally, the Love and Rayleigh (the slowest) waves. As you might expect, the difference in wave speed has a profound influence on the nature of seismograms. Since the travel time of a wave is equal to the distance the wave has traveled divided by the average speed the wave moved during the transit, we expect that the fastest waves arrive at a seismometer first. The fact that the waves travel at speeds that depend on the material properties (elastic moduli and density) allows us to use seismic-wave observations to investigate the interior structure of the planet. We can look at the travel times, or the travel times and the amplitudes of waves to infer the existence of features within the planet, and this is an active area of seismological research. To understand how we "see" into Earth using vibrations, we must study how waves interact with the rocks that make up Earth. Several types of interaction between waves and the subsurface geology (i.e., the rocks) are commonly observable on seismograms. We will examine the two simplest types of interaction refraction and reflection.

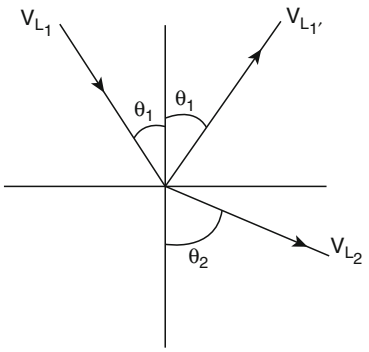
Refraction

As a wave travels through Earth, the path it takes depends on the velocity. Perhaps you recall from high school a principle called Snell's law, which is the mathematical expression that allows us to determine the path a wave takes as it is transmitted from one rock layer into another. Snell's law describes the relationship between the angles and the velocities of the waves. Snell's law equates the ratio of material velocities V_1 and V_2 to the ratio of the sines of incident (θ_1) and refracted (θ_2) angles, as shown in the following equation:

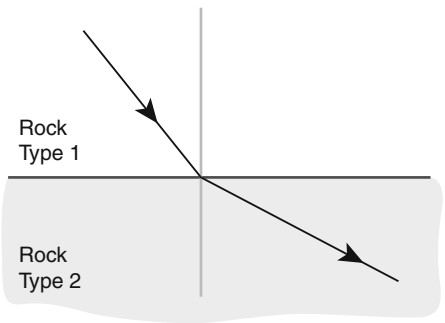
$$\frac{\sin\theta_1}{V_{L_1}} = \frac{\sin\theta_2}{V_{L_2}} \tag{1}$$

where V_{L_1} is the longitudinal wave velocity in material 1 and V_{L_2} is the longitudinal wave velocity in material 2.

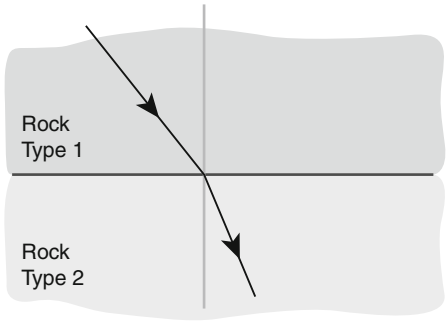
Note that in Figure 4, there is a reflected longitudinal wave ($V_{L_{1'}}$) shown. This wave is reflected at the same angle as the incident wave because the two waves are traveling in the same material, and hence have the same velocities. This reflected wave is unimportant in our explanation of Snell's law, but it should be remembered that some of the wave energy is reflected at the interface. When a longitudinal wave moves from a slower to a faster material, there is an incident angle that makes the angle of refraction for the wave 90°. This is known as the first critical angle. The change in direction depends on the ratio of the wave velocities of the two different rocks. When waves reach a boundary between different rock types, part of the energy is transmitted across the boundary. The actual interaction between a seismic wave and a contrast in rock properties is more complicated because an incident P-wave generates transmitted and reflected P- and S-waves and also an incident S-wave generates transmitted and reflected P- and S-waves, so five waves are involved. The transmitted wave travels in a different direction, which depends on the ratio of velocities of the two rock types (Figures 5 and 6). Part of the energy is also reflected backwards into the region with rock type 1, but we have not shown that on these figures. Refraction has an important effect on waves that travel through Earth. In general, the seismic velocity in Earth increases with depth (there are some important exceptions to this trend) and refraction of waves causes the path followed by body waves to curve upward (Figure 7).



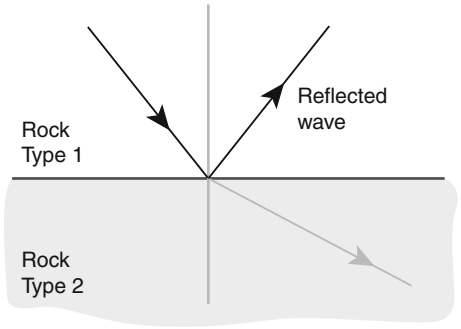
Body Waves, Figure 4 Reflected ($V_{L_{1'}}$) and refracted (V_{L_2}) longitudinal waves.



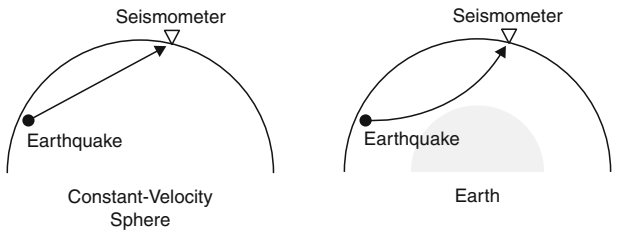
Body Waves, Figure 5 Velocity in rock type 2 is greater than velocity in rock type 1.



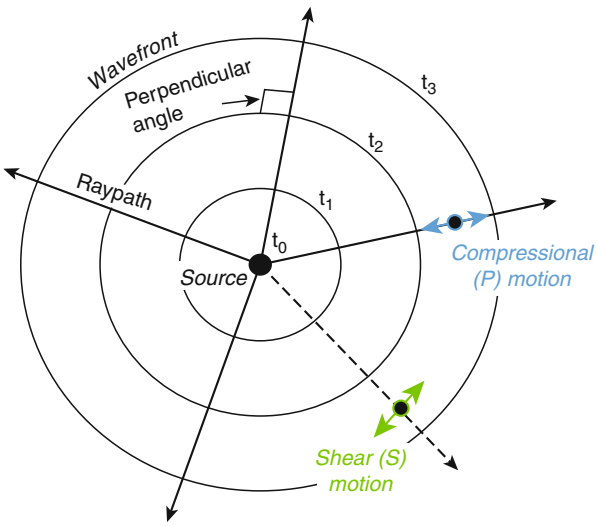
Body Waves, Figure 6 Velocity in rock type 2 is less than velocity in rock type 1.



Body Waves, Figure 8 When a wave encounters a change in material properties (seismic velocities and or density), its energy is split into reflected and refracted waves.



Body Waves, Figure 7 The overall increase in seismic-wave speed with depth into Earth produces an upward curvature to rays that pass through the mantle. A notable exception is caused by the decrease in velocity from the mantle to the core. This speed decrease bends waves backwards and creates a "P-wave shadow zone" between about 100° and 140° distance (1° = 111.19 km).



Body Waves, Figure 9 Wave fronts and ray paths in seismic-wave propagation.

Reflection

The second wave interaction with variations in rock type is reflection. I am sure you are familiar with reflected sound waves: we call them echoes. And your reflection in a mirror or pool of water is composed of reflected light waves. In seismology, reflections are used to prospect for petroleum and investigate Earth's internal structure. In some instances, reflections from the boundary between the mantle and crust may induce strong shaking that causes damage about 100 km from an earthquake (we call that boundary the "Moho" in honor of Mohorovicic, the scientist who discovered it). A seismic reflection occurs when a wave impinges on a change in rock type (which usually is accompanied by a change in seismic-wave speed). A part of the energy carried by the incident wave is transmitted through the material (that is the refracted wave described above), and a part is reflected back into the medium that contained the incident wave (Figure 8). The amplitude of the reflection depends strongly on the angle that the incident wave makes with the boundary and the contrast in material properties across the boundary. For some angles, all the energy can be returned into the medium containing the incident wave. An incident P-wave generates transmitted and reflected P- and S-waves and so five waves are involved. Likewise, when an S-wave interacts with a boundary in rock properties, it too generates reflected and refracted P- and S-waves. Because major boundaries between different rock types within the Earth are normally approximately parallel to the Earth's surface, S-wave particle motion is commonly in the SV (perpendicular to the ray path and vertical) and SH (perpendicular to the ray path and horizontal) directions. P-waves travel faster than S-waves, so there will be separate wave front representations for the P- and S-waves (Figure 9). If the physical properties of the material through which the waves are propagating are constant, the wave fronts will be circular (or spherical in three dimensions). If the physical properties vary in the model, the wave fronts will be more complex shapes. In the transverse or shear wave, the particles oscillate at a right angle or transverse to the direction of propagation. Shear waves

require an acoustically solid material for effective propagation, and therefore, are not effectively propagated in materials such as liquids or gases. Shear waves are relatively weak when compared to longitudinal waves. In fact, shear waves are usually generated in materials using some of the energy from longitudinal waves.

Velocity

Among the properties of waves propagating in isotropic solid materials are wavelength, frequency, and velocity. The wavelength (λ) is directly proportional to the velocity of the wave (C) and inversely proportional to the frequency (f) of the wave. The phase velocity of a seismic wave can be written as:

$$C = \omega/k, \tag{2}$$

where $\omega = 2\pi f$ is the angular frequency and $k = \lambda/2\pi$ is wave number.

Velocity of P- and S-waves

Seismic waves travel fast, of the order of kilometers per second (km/s). The precise speed that a seismic wave travels depends on several factors, the most important of which is the composition of the rock. We are fortunate that the speed depends on the rock type because it allows us to use observations recorded on seismograms to infer the composition or a range of compositions of the planet. But the process is not always simple, because sometimes different rock types have the same seismic-wave velocity, and other factors also affect the speed, particularly temperature and pressure. Temperature tends to lower the speed of seismic waves and pressure tends to increase the speed. Pressure increases with depth in Earth because the weight of the rocks above increases with increasing depth. Usually, the effect of pressure is larger and in regions of uniform composition, the velocity generally increases with depth, despite the fact that the increase of temperature with depth works to lower the wave velocity. When the different seismic-wave types are described below, I will quote ranges of speed to indicate the range of values we observe in common terrestrial rocks. But you should keep in mind that the specific speed throughout Earth will depend on composition, temperature, and pressure.

P-waves

P-waves are the first waves to arrive on a complete record of ground shaking because they travel the fastest (their name derives from this fact –P is an abbreviation for primary, first wave to arrive). They typically travel at speeds between ~1 and ~14 km/s. The slower values corresponds to a P-wave traveling in water, the higher number represents the P-wave speed near the base of Earth's mantle. The velocity of a wave depends on the elastic properties and density of a material. If we let κ represent the bulk modulus of a material, μ the shear modulus, and ρ the density, then the P-wave velocity, which we represent by C_P, is defined by:

$$C_P = \sqrt{\frac{\lambda + 2\mu}{\rho}} \tag{3}$$

A modulus is a measure of how easy or difficult it is to deform a material. For example, the bulk modulus is a measure of how a material changes volume when pressure is applied and is a characteristic of a material. For example, foam rubber has a lower bulk modulus than steel. P-waves are sound waves, it is just that in seismology we are interested in frequencies that are lower than humans' range of hearing (the speed of sound in air is about 0.3 km/s). The vibration caused by P-waves is a volume change, alternating from compression to expansion in the direction that the wave is traveling. P-waves travel through all types of media – solid, liquid, or gas.

S-waves

Secondary, or S-waves, travel slower than P-waves and are also called "shear" waves because they do not change the volume of the material through which they propagate – they shear it. S-waves are transverse waves because they vibrate the ground in a direction "transverse," or perpendicular, to the direction that the wave is traveling.

The S-wave speed, call it C_S, depends on the shear modulus and the density

$$C_S = \sqrt{\frac{\mu}{\rho}} \tag{4}$$

Even though they are slower than P-waves, the S-waves move quickly. Typical S-wave propagation speeds are of the order of 1–8 km/s. The lower value corresponds to the wave speed in loose, unconsolidated sediment, the higher value is near the base of Earth's mantle. An important distinguishing characteristic of an S-wave is its inability to propagate through a fluid or a gas because fluids and gases cannot transmit a shear stress and S-waves are waves that shear the material. The velocity of seismic waves depends on the elastic properties and density of a material in which the wave is traveling. P-waves are sound waves, it is just that in seismology we are interested in frequencies that are lower than humans' range of hearing (the speed of sound in air is about 0.3 km/s). The vibration caused by P-waves is a volume change, alternating from compression to expansion in the direction that the wave is traveling. In general, earthquakes generate larger shear waves than compressional waves and much of the damage close to an earthquake is the result of strong shaking caused by shear waves.

Attenuation

Near the source of a strong disturbance, such as a large explosion or earthquake, the wave-generated deformation

can be large enough to cause permanent deformation, which will be visible as cracks, fault offsets, and displacements of the ground after the disturbance has passed. A source of energy creates the initial disturbance and the resulting waves propagate (travel) out from the disturbance. Because there is finite energy in a confined or short-duration disturbance, the waves generated by such a source will spread out during propagation and become smaller (attenuate) with distance away from the source or with time after the initial source, and thus, will eventually die out (Figure 10). Knowledge of attenuation can be very useful in seismic data processing, as its removal increases resolution.

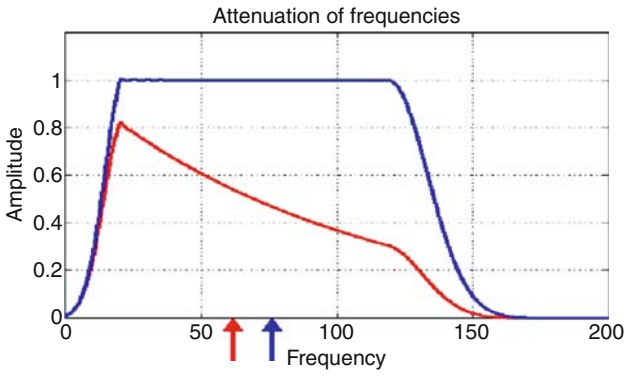
Attenuation often serves as a measurement tool that leads to the formation of theories to explain physical or chemical phenomenon that decreases the body wave's intensity. The amplitude change of a decaying plane wave can be expressed as:

$$A = A_0 \mathbf{e}^{-\alpha z} \tag{5}$$

In this expression A_0 is the unattenuated amplitude of the propagating wave at some location. The amplitude **A** is the reduced amplitude after the wave has traveled a distance **z** from the source. The quantity α is the attenuation coefficient of the wave traveling in the z-direction. The more common unit of the attenuation value is Decibels. Attenuation is generally proportional to the square of the frequency. Quoted values of attenuation are often given for a single frequency, or an attenuation value averaged over many frequencies may be given. The quoted values of attenuation only give a rough indication of the attenuation and should not be automatically trusted. Generally, a reliable value of attenuation can only be obtained by determining the attenuation experimentally for the particular material being used.

Summary: results of regional and global studies

The study of seismic body waves permits us to derive important knowledge about the internal constitution of the Earth. The problems related to propagation of elastic seismic body waves has been discussed in a number of books and in numerous papers (e.g., Bromwich, 1898; Sezawa, 1927; Haskell, 1953; Ewing et al., 1957; Brekhovskikh, 1960; Bullen, 1963; Bath, 1968; Achenbach, 1975 and others). Attenuations of seismic waves have been investigated by many authors (e.g., McDonald et al., 1958; Knopoff and Macdonald, 1958; Kolsky, 1963; White, 1965; Kuster and Toksoz, 1974; Shoenberger and Levin, 1974; Kennett, 1983; Hong and Kennett, 2003 and others). The asymptotic theory of body-wave propagation in anisotropic media is also well developed (e.g., Crampin, 1984; Thomsen, 1986; Tsvankin, 1997; Cerveny, 2001; Chapman, 2004).



Body Waves, Figure 10 Amplitude spectrum in red is for signal after attenuation.

The effect of initial stresses present in the medium is not considered in the above studies. Recently, Dey and Dutta (1998) have been studying the problem of propagation and attenuation of seismic waves, taking the effect of initial stresses present in the medium into account. Biot incremental deformation theory has been used (Biot, 1965) in their study. Selim and Ahmed (2006) have studied the velocities of propagation, damping, and attenuations of seismic body waves in compressible and dissipative medium under the effect of both initial and couple stresses.

Body-wave modeling

Seismological methods for determining Earth structure are often classified as being either active or passive in nature. Passive methods involve waiting for an earthquake to occur and provide the seismic source for recording. In contrast, controlled-source seismology refers to active methods where the experimenter provides the source by an explosion or a mechanical device such as a hammer, a weight that is dropped, or a vibrator. Active methods can, in turn, be divided into two basic classifications. The first is seismic reflection profiling, which is a clearly defined approach in which the goal is to produce an image of the subsurface in which structures can be seen directly, in the way that an X-ray image reveals features inside an object. Other active seismic methods infer seismic velocities and the presence of discontinuities in velocity and structure (such as faults) using a variety of approaches that analyze the arrival times and sometimes the shape of seismic waves traveling along different paths through the Earth. As a consequence of advances in seismic instrumentation and national programs increasing the number of instruments available, there have been many recent developments in these techniques that do not directly image the Earth. A convenient aspect of the theoretical basis for virtually all active-source techniques is that they are largely independent of scale. Thus, detailed studies to address environmental problems and regional studies to determine deep Earth structure employ the same basic techniques of analysis (Hancock and Skinner, 2000). Recently, several efforts have been made to complete our knowledge of the structure of the Earth's upper mantle by detailed observation of seismic body waves, especially by their amplitude–distance curves.

Bibliography

Achenbach, J. D., 1975. *Wave Propagation in Elastic Solids*. Amsterdam: North-Holland, p. 425.

Bath, M. A., 1968. *Mathematical Aspects of Seismology*. Amsterdam: Elsevier, p. 414.

Biot, M. A., 1965. *Mechanics of Incremental Deformation*. New York: Wiley, p. 181.

Brekhovskikh, L. M., 1960. *Waves in Layered Media*. New York: Academic.

Bromwich, T. J. L. A., 1898. On the influence of gravity on the elastic waves and in particular on the vibration of an elastic globe. *Proceedings of the London Mathematical Society*, **30**, 98–120.

Bullen, K. E., 1963. *An Introduction to the Theory of Seismology*. Cambridge: Cambridge University Press.

Cerveny, V., 2001. *Seismic Ray Theory*. Cambridge: Cambridge University Press.

Chapman, C., 2004. *Fundamentals of Seismic Wave Propagation*. Cambridge: Cambridge University Press.

Crampin, S., 1984. An introduction to wave propagation in anisotropic media. *Geophysical Journal of the Royal Astronomical Society*, **76**, 17–28.

Dey, S., and Dutta, D., 1998. Propagation and attenuation of seismic body waves in initially stressed dissipative medium. *Acta Geophysica Polonica*, **XLVI**(3), 351–365.

Ewing, W. M., Jardetzky, W. S., and Press, F., 1957. *Elastic Waves in Layered Media*. New York: McGraw-Hill, p. 380.

Hancock, P., and Skinner, B. J., 2000. Controlled-source seismology. The Oxford Companion to the Earth. *Encyclopedia.com.*

Haskell, N. A., 1953. The dispersion of surface waves in multilayered media. *Bulletin. Seismological Society of America*, **43**, 17–34.

Hong, T. K., and Kennett, B. L. N., 2003. Scattering attenuation of 2D elastic waves: theory and numerical modeling using a wavelet-based method. *Bulletin of the Seismological Society of America*, **93**(2), 922–938.

Kennett, B. L. N., 1983. *Seismic Wave Propagation in Stratified Media*. Cambridge: Cambridge University Press. 342 pp.

Knopoff, L., and Macdonald, J. F., 1958. Attenuation of small amplitude stress waves in solid. *Reviews of Modern Physics*, **30**, 1178.

Kolsky, H., 1963. *Stress Waves in Solid*. New York: Dover.

Kuster, G. T., and Toksoz, M. N., 1974. Velocity and attenuation of seismic waves in two-phase media, Part I: theoretical formulations. *Geophysics*, **39**(5), 587–618.

McDonald, F. J., Angona, F. A., Mills, R. L., Sengbush, R. L., Van Nostrand, R. G., and White, J. E., 1958. Attenuation of shear and compressional waves in Pierre shale. *Geophysics*, **23**(3), 421–439.

Selim, M. M., and Ahmed, M. K., 2006. Propagation and attenuation of seismic body waves in dissipative medium under initial and couple stresses. *Applied Mathematics and Computation*, **182**, 1064–1074.

Sezawa, K., 1927. Dispersion of elastic waves propagated on the surface of stratified bodies and on curves surfaces. *Bulletin of the Earthquake Research Institute (Tokyo)*, **3**, 1–18.

Shoenberger, M., and Levin, F. K., 1974. Apparent attenuation due to intrabed multiples. *Geophysics*, **39**(3), 278–291.

Thomsen, L., 1986. Weak elastic anisotropy. *Geophysics*, **51**, 1954–1966.

Tsvankin, I., 1997. Anisotropic parameters and p-wave velocity for othorhombic media. *Geophysics*, **62**, 1292–1309.

White, J. E., 1965. *Seismic Waves, Radiation, Transformation and Attenuation*. New York: McGraw-Hill, p. 302.

Cross-references

Deep Seismic Reflection and Refraction Profiling
Earth's Structure, Global
Earthquakes and Crustal Deformation
Earthquakes, Energy
Earthquakes, Source Theory
Propagation of Elastic Waves: Fundamentals
Seismic Phase Names: IASPEI Standard
Seismic Signals in Well Observations: Pre, Co, Post
Seismic Tomography
Seismic Velocity-Density Relationships
Seismic Wave Propagation in Real Media: Numerical Modeling Approaches
Seismic Waves, Scattering
Seismic, Ray Theory
Seismic, Viscoelastic Attenuation
Seismic, Waveform Modeling and Tomography
Seismology, Global Earthquake Model
Traveltime Tomography Using Controlled-Source Seismic Data

C

CHARACTERISTIC EARTHQUAKES AND SEISMIC GAPS

David D. Jackson, Yan Y. Kagan
Department of Earth and Space Sciences, University of California Los Angeles, Los Angeles, CA, USA

Definitions

Fault slip. Relative motion, either steady or sudden as in earthquakes, between rock units on either side of a fault.
Rupture. Sudden fault slip in an earthquake.
Elastic rebound. Sudden release, as by an earthquake, of slowly accumulated strain energy.
Segment. A section of a fault or plate interface bounded by features thought to serve as strong barriers to earthquake rupture. Features postulated to form such barriers include changes in fault orientation or in rock type across parts of the fault, and intersections with other faults.
Characteristic earthquake. An earthquake rupturing an entire fault segment. Alternately, one of a sequence of earthquakes rupturing the same area of fault.
Recurrence interval. The time between characteristic earthquakes on a given segment or fault area.
Quasiperiodic. Occurring at approximately equal recurrence intervals.
Seismic cycle. A sequence of events on a segment starting with a large earthquake, followed by aftershocks, then by steady stress accumulation, and culminating with another large earthquake. The term "cycle" is sometimes but not always meant to imply quasiperiodic recurrence.
Seismic gap. A segment for which the time since the previous characteristic earthquake approaches or exceeds the average recurrence interval.

Introduction

The seismic gap hypothesis holds that most long-term geologic slip on faults or plate boundaries is accomplished by characteristic earthquakes on segments. Such quakes are presumed to reduce the stress substantially, necessitating a substantial recurrence time for elastic stress to recover before the next characteristic earthquake. The dates and rupture extent of past earthquakes may be determined by modern seismic networks; by historic reports of faulting, damage, or strong shaking; or by paleoseismic investigation of trenches across faults. The average recurrence time may be determined either from a sequence of dates of past characteristic earthquakes, or by the time required for steady slip on a fault to accumulate the slip experienced in a characteristic earthquake (Working Group on California Earthquake Probabilities, 1988, 1990).

History

Well before plate tectonics became accepted in the 1960s, Gilbert (1884) argued that large earthquakes should be separated by substantial time intervals. Reid (1910, 1911) proposed that faults slowly accumulate energy later released suddenly by earthquakes ("elastic rebound"), and that the time to a future earthquake could be estimated by surveying the strain released by a previous one. Fedotov (1965) noticed that most of the Kamchataka–Kurile trench in the northwest Pacific had been ruptured by earthquakes in the early twentieth century, with the exception of one zone southeast of Hokkaido. He concluded that a future earthquake was likely in that area, a forecast realized in 1973 when the Nemuro-oki earthquake struck with magnitude about 7.5 (Kasahara, 1981, 182).

Plate tectonics now provides a compelling, steady source of the strain energy driving elastic rebound, and

Harsh K. Gupta (ed.), *Encyclopedia of Solid Earth Geophysics*, DOI 10.1007/978-90-481-8702-7,

that fact led many to conclude that occurrence of large earthquakes must be separated by enough time to allow recharge. Scholz (1990, 260) remarked that "A tenet of plate tectonics is that the rates of plate motion must be steady over geologic time periods and must be continuous along plate boundaries. If it is assumed further that a significant portion of this motion must be released seismically, then it follows that segments of plate boundaries that have not ruptured for the longest time are those most likely to rupture in the near future. These places are called seismic gaps." The seismic gap hypothesis was celebrated when Nishenko, 1989 published a list of 13 earthquakes that fit, at least approximately, the descriptions of previously expected earthquakes.

The seismic gap model has been applied to long-term forecasting of earthquakes in many regions (Sykes, 1971; Kelleher, 1972; Kelleher et al., 1973; McCann et al., 1979; Working Group on California Earthquake Probabilities, 1988, Nishenko, 1991). The model and its definition have evolved along with the quality and quantity of data that go into it. Fedotov (1965) defined the Nemuro-oki gap as the last remaining unbroken segment. McCann et al. (1979) used elapsed time and a color code to label segments around the Pacific Rim: more than 100 years had elapsed in red gaps, between 30 and 100 years in orange gaps, and less than 30 years in apparently safe green zones. In the most comprehensive forecast ever using the seismic gap model, Nishenko (1991) actually made testable probabilistic forecasts for about 125 plate boundary segments around the Pacific Rim. In that study he estimated the mean recurrence time and the elapsed time for each segment, and assumed that their ratio obeys a log-normal probability density function. With that information, he calculated the probability that each zone would be ruptured by a characteristic earthquake whose magnitude he also listed for each zone within 5-, 10-, and 30-year periods. A very similar model was used by the Working Group on California Earthquake Probabilities (1988) in its official earthquake probability estimates.

Assumptions

All the applications mentioned above share several important assumptions. First, their authors assume that faults and plate boundaries are segmented and that rupture does not cross segment boundaries. Second, they assume that each characteristic earthquake ruptures to both ends of its segment, reducing the stress to a uniform base level and beginning the process of stress recharge. Third, they assume that the time to the next characteristic earthquake depends almost entirely on the time of the previous one: not on other earthquakes, nonelastic stress redistribution, or other causes. To make useful forecasts, scientists must obviously be able to distinguish characteristic earthquakes from all others in order to know the elapsed time since the last one.

Small characteristic earthquakes

Recent studies in California (e.g., Nadeau and McEvilly, 1999), Japan (e.g., Igarashi et al., 2003; Okada et al., 2003), and elsewhere have identified sequences of small earthquakes fitting the alternative definition of "characteristic earthquake" above. In each sequence, the events are approximately the same size, and rupture approximately the same fault area. They recur at nearly equal time intervals or in some cases variable intervals consistent with variations in their size or in fault slip rate. In most cases, the slipped areas appear to be surrounded by areas where displacement occurs by steady slip rather than earthquakes. Because of that special circumstance, these small repeating earthquakes are not relevant to the discussion of seismic gaps.

Modified seismic gap hypothesis

As time, earthquake experience, and theoretical sophistication have accumulated, earth scientists have modified the seismic gap theory to rely less on the assumptions above. The Working Group on California Earthquake Probabilities (1990) allowed adjustments to account for stresses from earthquakes not on the relevant segment. The 1992 Landers, California earthquake (magnitude about 7) presented a particularly important observation. In that event, rupture jumped segment boundaries and even faults, making use of up to five faults mapped as separate before 1992. In a seismic hazard model produced by the Southern California Earthquake Center (Working Group on California Earthquake Probabilities, 1995), the seismic gap model was modified to allow rupture to jump segment boundaries with a modest probability. Later uses of the seismic gap model in California for official hazard estimates employ increasingly complex assumptions, especially about conditions under which rupture is likely to involve more than one segment (Working Group on California Earthquake Probability, 2002, 2008).

The more complex versions of the model raise interesting questions. What truly constitutes a segment? How effective are weak barriers in stopping rupture, and what controls their strength? Are the boundaries fixed in space, or can they move as stress conditions change? When rupture crosses a boundary, does it consistently continue to the next? If not, does it reset the stress and the clock on the partially ruptured segment? Do the elapsed times on adjacent segments control the probability that rupture will jump the barrier between them? If so, which segment is most important? Modelers must answer these questions, implicitly or explicitly, to forecast using the modified gap models. So far, there is no clear consensus on the answers.

Challenges to the seismic gap model

Despite some reported successes, the seismic gap hypothesis has often been questioned. Critics point to the difficulty of verifying the rather strong assumptions behind the hypothesis, and to its limited success in forecasting earthquakes.

The basic assumption that faults and plate boundaries are segmented has provoked significant debate. Even the few apparent successes (e.g., Loma Prieta, CA, 1989;

Parkfield, CA 2004; Chile, 2010), are equivocal at best. The rupture of the Loma Prieta earthquake was about the length of the nearest segment mapped by the Working Group on California Earthquake Probabilities (1988), but it shifted south and spilled over the southern segment boundary. Moreover, the event occurred near, but not actually on the San Andreas Fault for which the segment was defined. Before 2004, the Parkfield segment was defined in several different ways, so its location at Parkfield does not confirm the segmentation hypothesis (Jackson and Kagan, 2006). The 2010 Chile earthquake went well beyond the segment boundaries specified by Nishenko (1991). In addition, several events have clearly violated preassigned boundaries. The 1992 Landers quake was mentioned above, and the great Sumatra tsunami earthquake of 2004 breached several boundaries along its 1,300-km rupture zone (Nalbant et al., 2005).

The assumption that earthquakes rupture to both ends of segment boundaries also lacks verification. A basic problem is that the locations of segment boundaries are usually estimated inaccurately from the extent of past earthquake ruptures. For earthquakes identified by paleoseismic investigations, rupture can generally be pinpointed at only a few widely spaced locations. For historical earthquakes, rupture extent is typically estimated with great uncertainty from the extent of damage or reported shaking. Even for modern instrumentally recorded earthquakes, the extent of the aftershock zone or fault scarp may not accurately represent the rupture at depth where the elastic rebound occurs. In many cases, the extent of rupture for older earthquakes is assumed to be similar to that of the most recent event, a clear case of circular reasoning.

The connection between plate tectonics and earthquake recurrence referred to in the words of Scholz above depends on the assumption that characteristic earthquakes release most of the slowly accumulated fault slip. However, that assumption fails in many examples. For instance, the cumulative slip of the Parkfield, CA earthquakes since 1857, often regarded as an archetypical characteristic sequence, accounts for only a small fraction of expected slip at the long-term geological rate (Jackson and Kagan, 2006). In such cases, the alternative definition of characteristic earthquakes listed above, and the times of past events, may provide valuable information on the causes of some earthquakes but the direct link to plate tectonics is lost.

Proponents of the seismic gap theory cite many examples in which identified gaps have been filled by earthquakes. The positive examples are appealing but insufficient for two reasons. First, the definitions of gaps and forecasted earthquakes were quite general, making the target easier to hit at random. Second, they included only successes; a fair evaluation needs to consider failures as well.

Kagan and Jackson (1991, 1995), and Rong et al. (2003) applied statistical tests to the seismic gap theory as articulated by McCann et al. (1979) and Nishenko (1991). Earthquakes were actually more frequent in McCann's green zones than in the red ones, opposite to what the gap theory assumes. The 1991 gap model implied far more earthquakes, in different places, than actually occurred. Kagan and Jackson (1995) and Rong et al. (2003) also tested a simple alternative to the gap model, assuming that earthquakes occur randomly in time and near past earthquakes. The alternative model fits the total number and the locations of future earthquakes much better than the 1991 gap model.

These statistical tests necessarily applied to the earlier versions of the gap hypothesis, in which segments were assumed to be independent of one another. Since then, more complex versions of the model have been applied. Most of these applications involve one or a few purported gaps with estimated recurrence times of decades or centuries. Published models generally do not provide probabilities for multi-segment ruptures, and they cannot be effectively evaluated until several seismic cycles have elapsed. To be rigorously testable, such a model must forecast a few tens of well-defined earthquakes. Unfortunately, no systematic, well-specified version of the more complex seismic gap model has been applied broadly enough to be tested.

Conclusions

The intuitively appealing seismic gap model encompasses the virtually unassailable principle that earthquakes release strain energy accumulated over a long time. Although many large events have occurred in previously identified gaps, the same is true of locations outside them. Simple versions of the gap theory, in which characteristic earthquakes occur on independent segments, are no longer tenable. Modified versions of the gap theory have not yet been formulated in a rigorously testable way.

Bibliography

Fedotov, S. A., 1965. Regularities of the distribution of strong earthquakes in Kamchatka, the Kurile Islands and northeastern Japan. *Tr Inst Fiz Zemli, Akad. Nauk SSSR*, **36**(203), 66–93 (in Russian).

Gilbert, G. K., 1884. A theory of the earthquakes of the Great Basin, with a practical application. *American Journal of Science*, **27**(157), 49–54. Ser. 3.

Igarashi, T., Matsuzawa, T., and Hasegawa, A., 2003. Repeating earthquakes and interplate aseismic slip in the northeastern Japan subduction zone. *Journal of Geophysical Research*, **108**(B5), 2249, doi:doi:10.1029/2002JB001920.

Jackson, D. D., and Kagan, Y. Y., 2006. The 2004 parkfield earthquake, the 1985 prediction, and characteristic earthquakes: lessons for the future. *Bulletin. Seismological Society of America*, **96**, S397–S409.

Kagan, Y. Y., and Jackson, D. D., 1991. Seismic gap hypothesis: Ten years after. *Journal of Geophysical Research*, **96**(21), 21419–21431.

Kagan, Y. Y., and Jackson, D. D., 1995. New seismic gap hypothesis: Five years after. *Journal of Geophysical Research*, **100**(B3), 3943–3959.

Kasahara, K., 1981. *Earthquake Mechanics*. Cambridge: Cambridge University Press.

Kelleher, J. A., 1972. Rupture zones of large South American earthquakes and some predictions. *Journal of Geophysical Research*, **77**, 2087–2103.

Kelleher, J. A., Sykes, L. R., and Oliver, J., 1973. Possible criteria for predicting earthquake locations and their applications to

major plate boundaries of the Pacific and Caribbean. *Journal of Geophysical Research*, **78**, 2547–2585.
McCann, W. R., Nishenko, S. P., Sykes, L. R., and Krause, J., 1979. Seismic gaps and plate tectonics: Seismic potential for major boundaries, *Pure and Applied Geophysics*, **117**, 1082–1147.
Nadeau, R. M., and McEvilly, T. V., 1999. Fault slip rates at depth from recurrence intervals of repeating microearthquakes. *Science*, **285**, 718–721.
Nalbant, S., Steacy, S., Sieh, K., Natawidjaja, D., and McCloskey, J., 2005. Seismology: Earthquake risk on the Sunda trench. *Nature*, **435**(7043), 756–757.
Nishenko, S. P., 1989. Earthquakes: hazards and predictions. In James, D. E. (ed.), *The Encyclopedia of Solid Earth Geophysics*. New York: Van Nostrand Reinhold, pp. 260–268.
Nishenko, S. P., 1991. Circum-Pacific seismic potential – 1989–1999. *Pure and Applied Geophysics*, **135**, 169–259.
Okada, T., Matsuzawa, T., and Hasegawa, A., 2003. Comparison of source areas of M4.8 ± 0.1 repeating earthquakes off Kamaishi, NE Japan: Are asperities persistent features? *Earth and Planetary Science Letters*, **213**, 361–374.
Reid, H. F., 1910. *The California Earthquake of April 18, 1906*. Washington: Carnegie Institution of Washington. The Mechanics of the Earthquake, Vol. 2.
Reid, H. F., 1911. The elastic-rebound theory of earthquakes, *University of California Department of Geology Bulletin*, **6**, 413–444.
Rong, Y.-F., Jackson, D. D., and Kagan, Y.Y., 2003. Seismic gaps and earthquakes, Journal of Geophysical Research, 108(B10), 2471, ESE-6, 1–14.
Scholz, C., 1990. *The Mechanics of Earthquakes and Faulting*. Cambridge: Cambridge U. Press.
Sykes, L. R., 1971. Aftershock zones of great earthquakes, seismicity gaps, and earthquake prediction for Alaska and the Aleutians. *Journal of Geophysical Research*, **76**, 8021–8041.
Working Group on California Earthquake Probabilities, 1988. Probabilities of large earthquakes occurring in California, on the San Andreas Fault. U.S. Geol. Surv., Open File Rep., 88–398, 62 ~ pp.
Working Group on California Earthquake Probabilities, 1990. Probabilities of Large Earthquakes in the San Francisco Bay Region, California, USGS Circular 1053.
Working Group on California Earthquake Probabilities, 1995. Seismic hazards in southern California: probable earthquakes, 1994–2024. *Bulletin Seismological Society of America*, **85**, 379–439.
Working Group on California Earthquake Probabilities, 2002. Earthquake probabilities in the San Francisco Bay region: 2002 to 2031, USGS Circular 1189.
Working Group on California Earthquake Probabilities, 2008. The Uniform California Earthquake Rupture Forecast, Version 2 (UCERF 2): U.S. Geological Survey Open-File Report 2007–1437 and California Geological Survey Special Report 203 [http://pubs.usgs.gov/of/2007/1437/].

Cross-references

Earthquake Precursors and Prediction
Earthquake Rupture: Inverse Problem
Earthquakes and Crustal Deformation
Earthquakes, Energy
Earthquake, Location Techniques
GPS, Tectonic Geodesy
Great Earthquakes
Paleoseismology
Seismic Hazard
Seismicity, Subduction Zone
Seismology, Global Earthquake Model
Statistical Seismology

CONTINENTAL DRIFT

Alan G. Smith
Department of Earth Sciences, University of Cambridge, Sedgwick Museum, Cambridge, UK

Definition

Continental drift. The name given to the relative movement between continents.

Introduction

Apparently, continental drift was first postulated by Abraham Ortelius, who made one of the first atlases, in the third edition of his *Thesaurus Geographicus* (Ortelius, 1596) to account for the similarity between the coastlines of western Africa, western Europe, and the eastern Americas. During the next three centuries or so, several other thinkers used this morphological similarity to come to the same conclusion.

Alfred Wegener

Although Ortelius speculated that earthquakes and floods had torn the continents apart, Alfred Wegener, an Austrian meteorologist, was the first to systematically gather the geological evidence for continental drift in his synthesis (Wegener, 1929a), also translated into English (Wegener, 1929b), parts of which had been published as early as 1912 (Wegener, 1912).

Wegener was puzzled in particular by the present-day distribution of former ice-age deposits, or tillites, of Permo-Carboniferous age, now known to be about 300 million years old that are found today in South America, Africa, Antarctica, India, and Australia. Wegener's meteorological background led him to assume that the present-day climatic zones, with cold polar regions and a hot tropical belt, was a fundamental property of the Earth's atmosphere that had been established before these glacial deposits had formed and had persisted to the present-day. He realized that if all the southern continents had been joined together to form a supercontinent lying near the south geographic pole, then the *present-day* distribution of all the Permo-Carboniferous tillites would have a logical explanation. This supercontinent is known as Gondwana. But Wegener went further and postulated that the northern continents had also been joined to form a northern supercontinent known as Laurasia, which, with Gondwana, formed a huge continental area incorporating all the major continents, known as Pangea.

Wegener's solution to the tillite distribution had a compelling elegance about it, partly because it also placed the Carboniferous forests of Europe and North America, (whose compressed remains gave the name to the Carboniferous period) in what would have been the tropical region of that time. However, elegance is not a scientific proof, and Wegener's ideas were rejected by most geophysicists and many geologists, principally

because there were no physical measurements that demonstrated drift and there was also no known *physical* mechanism that could move continents independently of one another [e.g., Jeffreys (1929)].

Wegener (1929a) made the reassembly by positioning the continents at where he presumed the edges of the continents lay, that is, somewhat beyond the present-day coastlines, but his maps were rather schematic. A more convincing Gondwana map, supported by much geological evidence, was made by Du Toit (1937), but the best precomputer map showing the excellence of the fit between the continental edges of South America and Africa was made by Carey (1958). Jeffreys (1964) denied that there was a fit, and it was his denial that in part led Bullard and others to examine how well the continental edges fitted together (Bullard et al., 1965; Everett and Smith, 2008). These ideas were supported by the changes in the Earth's magnetic field preserved in rocks (see *Paleomagnetism, Principles*) determined by workers such as Irving (1964), but the results were *not fully accepted by the geoscience community* until after the plate tectonic revolution.

Conclusions

It is now established that the Earth's lithosphere is divided into a number of rigid *Plate Tectonics, Precambrian* in relative motion. Tracing these motions back in time shows that Gondwana, Laurasia, and Pangea did exist in the past, *vindicating Wegener's idea completely.*

Bibliography

Bullard, E., Everett, J. E., and Smith, A. G., 1965. The fit of the continents around the Atlantic. *Philosophical Transactions of the Royal Society of London*, **A258**, 41–51.

Carey, S. W., 1955. The orocline concept in geotectonics – Part I, Publication No. 28. The Papers and Proceedings of the Royal Society of Tasmania, 89, pp. 255–288.

Carey, S. W., 1958. A tectonic approach to continental drift. In Carey, S. W. (ed.), *Continental Drift: A Symposium*. Hobart: University of Tasmania, pp. 177–355.

Du Toit, A. L., 1937. *Our wandering continents*. Edinburgh: Oliver & Boyd.

Everett, J. E., 1965. The Fit of the Continents around the Atlantic. Ph. D., thesis, Cambridge.

Everett, J. E., and Smith, A. G., 2008. Genesis of a geophysical icon: the Bullard, Everett and Smith reconstruction of the circum-Atlantic continents. *Earth Sciences History*, **27**(1), 1–11. http://pubs.usgs.gov/gip/dynamic/historical.html

Irving, E., 1964. *Palaeomagnetism and its application to geological and geophysical problems*. Wiley: New York, 399 pp.

Jeffreys, H., 1929. *The Earth*. Cambridge: Cambridge University Press.

Jeffreys, H., 1964. How soft is the Earth? *Quarterly Journal of the Royal Astronomical Society*, **5**, 10–22.

Ortelius, A., 1596. Thesaurus geographicus, recognitus et auctus. Antwerp, Ex Officina Plantiniana.

Ortelius, A., 1603. Theatrum orbis terrarum. Abridged edition printed by James Shawe, London.

Smith, A. G., and Hallam, A., 1970. The fit of the southern continents. *Nature*, **225**, 139–144.

Wegener, A. L., 1912. Die Entestehung der kontinente (The origin of continents). *Geologische Rundschau*, **3**, 276–292.

Wegener, A. L., 1929a. Die Entestehung der kontinente und Ozeane. (The origin of continents and oceans). Brunswick, Vieeweg.

Wegener, A. L., 1929b. *The Origin of Continents and Oceans (English translation)*. London: Methuen.

Cross-references

Lithosphere, Continental
Lithosphere, Continental: Thermal Structure
Lithosphere, Mechanical Properties
Lithosphere, Oceanic
Paleomagnetism, Principles
Plate Driving Forces
Plate Motions in Time: Inferences on Driving and Resisting Forces
Plate Tectonics, Precambrian

CONTINENTAL RIFTS

A. M. Celâl Şengör
Eurasia Earth Science Institute, Istanbul Technical University, Istanbul, Turkey

Synonyms

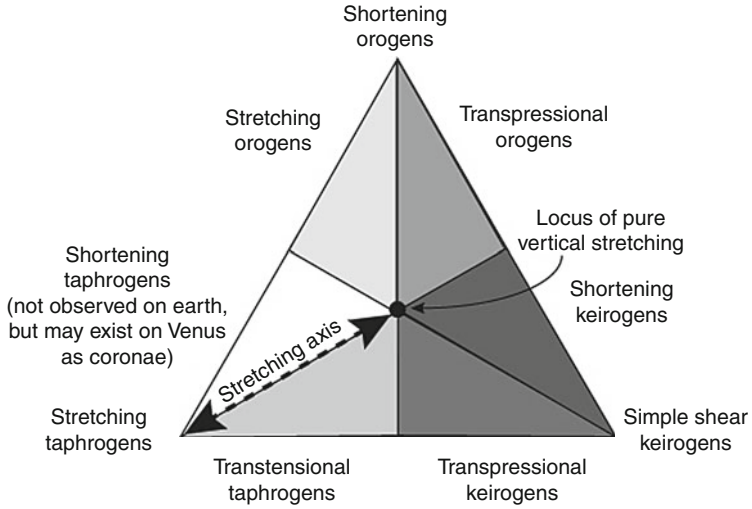
Graben; Taphrogen

Definition

A continental rift (Gregory, 1894; Quennell, 1982, 1985) is a fault-bounded elongate trough under or near which the entire thickness of the lithosphere (see *Lithosphere, Continental*; *Lithosphere, Mechanical Properties*) has been reduced in extension during its formation. Just as old mountain ranges may no longer have any topographic expression because of tectonic and/or erosional events, some, especially old, rifts may no longer appear as troughs for the same reasons, but their original trough shape is recognized by their stratigraphically younger fills, or metamorphically lower grade of their down-dropped central blocks (see *Sedimentary Basins*).

Introduction

Rifts form one of the three main categories of lithospheric-scale structures resulting from differential motion of parts of the lithosphere. Lithospheric shortening creates orogens, simple shear creates keirogens, and stretching creates taphrogens that are collections of rifts (Figure 1). However, at present, only 20.5% of the active plate boundaries show normal convergence and 21% normal divergence. Some 8% show pure dextral strike-slip and some 6% pure sinistral strike-slip. The remaining 58.5% display some deviation from the three end-members, with relative motion vector angles to bountary strikes varying between 0° and 67° (Woodcock, 1986). Plate boundaries must have shown a similar behavior in the past, so only about half of all the rifts a geologist may encounter is likely to show normal extension



Continental Rifts, Figure 1 A ternary diagram of strain showing the three lithospheric megastructure categories resulting from various strains.

(see *Plate Motions in Time: Inferences on Driving and Resisting Forces*; *Plates and Paleoreconstructions*).

Rifts are important structures from the viewpoints of our understanding of the behavior of our planet and our exploitation of its resources. They record evidence of continental fragmentation in diverse tectonic settings, including all three types of plate boundaries and in plate interiors (see *Plate Motions in Time: Inferences on Driving and Resisting Forces*; *Plates and Paleoreconstructions*). Also, at different stages of their evolution, they present opportunities of studying the continental crust (and in extreme cases even the upper mantle) from its surficial sedimentary rocks down to the crust-mantle interface (see *Lithosphere, Continental*). Especially the lacustrine sedimentary sections of rifts are useful for studies on the past climates (e.g., Olson and Kent, 1999; Kravchinsky et al., 2003; Felton et al., 2007) and they have enabled geologists to refine stratigraphic correlations down to thousands of years, as far back as in the early Mesozoic (e.g., Olsen and Kent, 1999).

Rifts house important economic reserves such as hydrocarbons (the Sirt rift in Libya being the most productive with about 45 billion barrels of oil and 33 trillion cubic feet of gas; for the hydrocarbon potential of rifts, see Harding, 1983; MacKenzie and McKenzie, 1983: a classic on hydrocarbon genesis in rifts; Ziegler, 1994a; Lambiase, 1995), hydrothermal and stratiform copper, copper-nickel, molybdenum, giant lead-zinc, and even uranium deposits (for metallogenesis in rifts, see Sawkins, 1990). Most of the geothermal areas in the world are located in active rifts (e.g., Jaffé, 1971; figure 1) (see *Geothermal Heat Pumps*). At present, one fifth of the entire fresh-water reserve of the earth resides in a rift basin, namely Lake Baykal. Rifts are thus significant not only for increasing our geological knowledge, but also for contributing to our well-being. It is perhaps of some interest in this context to remember that our genus and species were born in a rift, namely in the East African.

For good overviews of rifts and rifting, see Burke and Dewey (1973), Burke and Wilson (1976), Coward et al. (1987), McKenzie (1978) (a "must" read for understanding the evolution of rifts), Manspeizer (1988), Ziegler (1992, 1994b), Landon (1994), Şengör (1995), Şengör and Natal in (2001) (a detailed catalog of the world's rifts with tabulated summaries of origin and history and references) and Ring and Wernicke (2009). For the geology of some of the classic rift regions of the world, with increasing age, see Africa as a whole: Kampunzu and Lubala (1991), Burke (1996), and Burke and Gunnell (2008). East Africa: Frostick et al. (1986), Frostick (1997) (a good introduction for beginner), Schlüter (1997) (this book has a bibliography of some 1,300 items), Morley (1999) (has an excellent collection of seismic sections). Basin-and-Range: Wernicke (1990), Beratan (1996), Faulds and Stewart (1998), Snow and Wernicke (2000), Dickinson (2002) (best introduction for an outsider) and McQuarrie and Wernicke (2005) (best kinematic reconstruction of any rift I know). Rio Grande rift, USA: Ingersoll (2001). The Upper Rhine rift: see the special issue on it in the *International Journal of Earth Sciences (Geologische Rundschau)*, vol. 94 (2005); also: Hüttner (1991), Schumacher (2002), Rotstein et al. (2006). The Lake Baykal rift: Crane et al. (1991), Kuz'min et al. (2001), ten Brink and Taylor (2002), Kravchinsky et al. (2003). The West Siberian Basin: Surkov et al. (1994). The North Sea Basin with the North German (Lower Saxony) Basin: Blundell and Gibbs (1990), Evans et al. (2003) (this lavishly illustrated and richly documented mammoth book makes all earlier publications on the subject either obsolete or redundant; also available as a two CD-ROM set), Wilson et al. (2004), Littke et al. (2008). The Mediterranean region: Durand

et al. (1999), Corti et al. (2006) (an excellent paper about a little-known rift cluster).

Rift, graben, and taphrogen

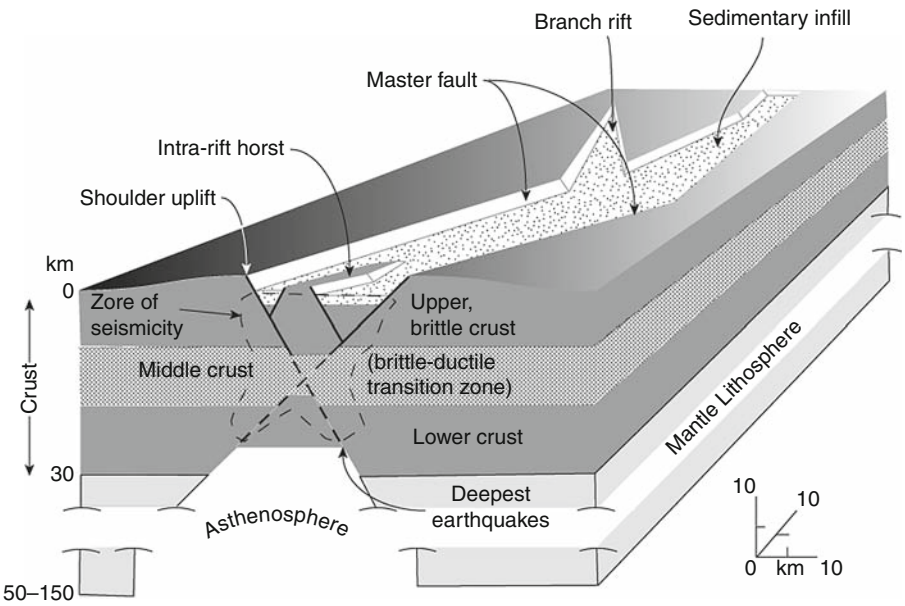
"Rift" and "graben" are two words used interchangeably in the international literature designating elongate troughs bounded by normal faults. However, not all such troughs are parts of structures that rupture, or even thin by rock creep, the entire lithosphere. Some normal-fault-bounded troughs result from sliding and stretching of surficial rocks over lubricating sedimentary horizons such as evaporites or shales. The normal-fault-bounded basinets that formed during a landslide associated with the March 27, 1964, Prince William Sound earthquake in Alaska or the similar troughs in the Canyonlands National Park in Utah formed by sliding and stretching the sedimentary section above the Upper Carboniferous evaporites are examples. Such structures have nothing to do with lithospheric stretching and thus they do not fit the definition of rift given above. Şengör (1995) therefore suggested that the word rift be confined to those structures that actually disrupt or otherwise (by creep) thin the lithosphere and graben be reserved for those that do not. His usage is what is recommended and followed in this entry. In rifts, when the extension factor β, the ratio of extended width to unextended width (McKenzie, 1978), goes beyond 3 the continental lithosphere generally ruptures completely, and ocean generation begins by seafloor spreading (Le Pichon and Sibuet, 1981) (see *Seafloor Spreading*) either by dyke injection or by cold extrusion of subcontinental mantle in the form of ultramafic "core complexes" as seen in such places as along the boundary between the Upper Penninic and the Lower Austroalpine units of the Alps (Bernoulli et al., 2003 and the references therein) or along the South Galicia Bank (Boillot et al., 1987). For obvious reasons, this cannot happen to grabens, however much they may be extended.

Large regions of localized extension exist on our planet where many rifts and grabens occur together. It is inappropriate to call such regions simply rifts. It would be like calling whole orogens simply "thrust wedges." These large regions of extension are the extensional counterparts of orogens (Figures 1 and 9) and therefore it is appropriate to use for them the term "taphrogen," derived from Krenkel's term "taphrogeny."

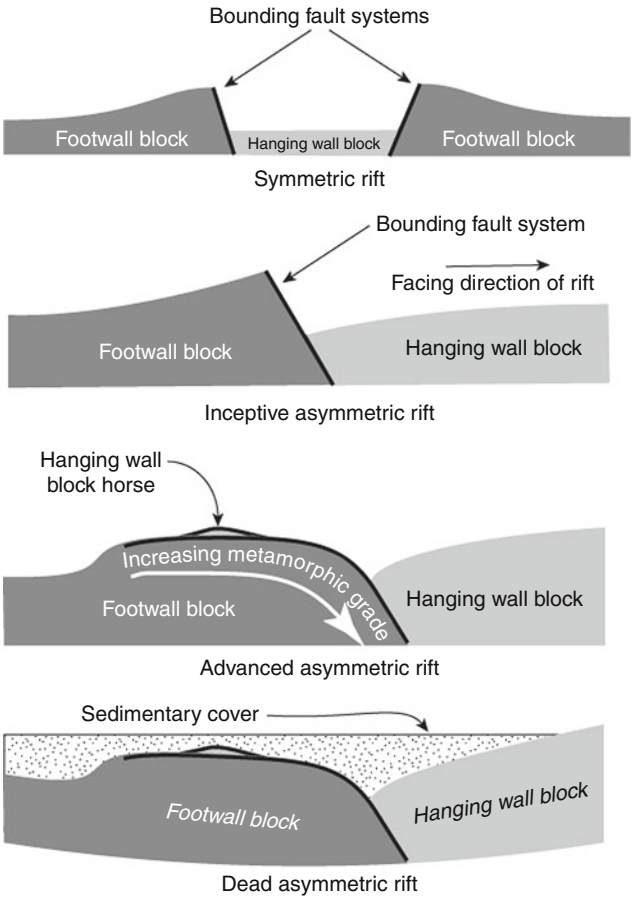
General properties of rifts

Structure of rifts

Figure 2 is an idealized and highly simplified block diagram of a rift. Rifts have two fundamental structural entities: a main hanging wall block and one or two footwall blocks depending on whether the rift is symmetric (as in Figure 2) or asymmetric (see Figure 3). In symmetric rifts, the extension represented by the rift is almost entirely localized between two master bounding normal (or oblique extensional) fault systems, and it is these two systems that take up much of the extension, although the hanging wall block delimited by them is also intensely broken up by smaller normal faults with strikes largely parallel with those of the major bounding fault families (Figure 4). The hanging wall is commonly dropped down because of crustal thinning affecting it and in the basin thus formed sediments accumulate whose thicknesses can be in excess of 15 km in large rifts such as the Dneper-Donetz rift in the Ukraine (Ulmishek et al., 1994), which is very close to the maximum possible



Continental Rifts, Figure 2 A highly schematic block diagram of an "ideal" rift drawn to emphasize the main architectural elements of rifts. The straight marginal faults, e.g., do not mean that all rifts must have straight faults or that they are single faults instead of fault families.



Continental Rifts, Figure 3 Some properties of symmetric and asymmetric rifts.

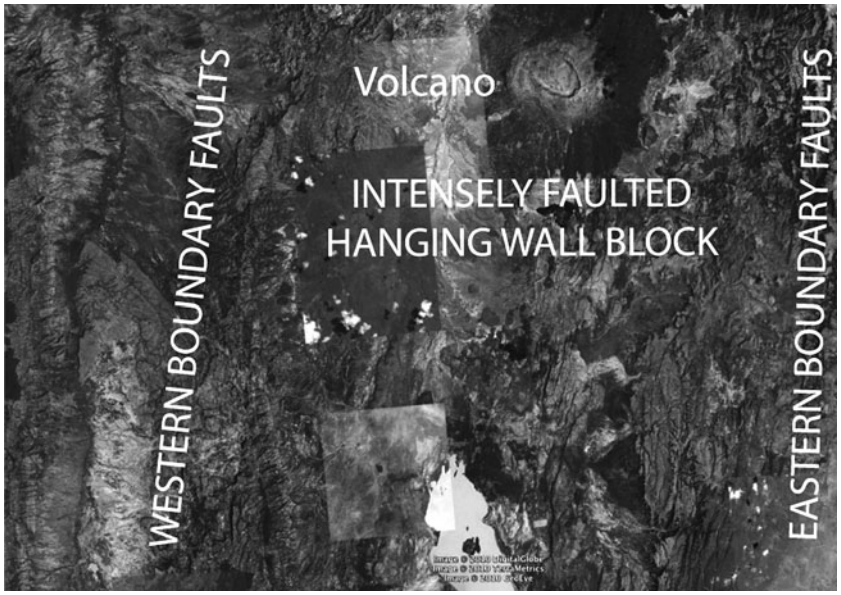
thickness (17 km) of sediment accumulation on continental crust (Dewey, 1982) Most rifts of average size, however, have sediment thicknesses ranging from 2 to 5 km. Individual rift lengths vary from about 1,000 to 100 km and widths from about 150 to 10 km. Rift size is a function not only of the amount of extension, but also the thickness of the lithosphere it disrupts. The thicker the lithosphere, the wider the rift. On Figure 5, the distribution of continental rifts on earth is shown. In areas of thick lithosphere (e.g., in the North American craton, eastern Europe, and Siberian craton), the rifts are very wide, whereas in areas of thin lithosphere such as the western USA or in the Aegean they are very narrow, in places less than 10 km wide.

The bounding faults of the rifts have long been thought to be listric (shovel-shaped), i.e., flattening with depth, although studies on earthquakes have not supported this. Wherever precise hypocenter location is possible and where the earthquake has caused a surface rupture (see *Earthquake, Focal Mechanism*; *Earthquakes and Crustal Deformation*; *Earthquake, Location Techniques*), it is seen that the associated fault is planar down to depths of 8–10 km. Eyidoğan and Jackson (1985) documented that the March 28, 1969, Demirci earthquake in western Turkey consists of two discrete sub-events: one along a northerly-dipping planar normal fault and another one on a horizontal surface north of, but at the same depth as, the hypocenter of the first one. The horizontal rupture is believed to be located in a region that deforms ductiley under long-term strain rates, but ruptures when that rate is suddenly increased as in an earthquake (Figure 2). This suggests a kinked cross-section for the master fault. Such "kinked" faulting events may be the explanation of why hanging wall blocks in some rifts are so much more disrupted than the rift shoulders. In areas of thicker lithosphere, however, hypocenters along the master faults of rifts may be as deep as 30 km as exemplified by the March 10, 1989, Malawi earthquake in east Africa (Figure 2; Jackson and Belkinsop, 1993) (see *Seismicity, Intraplate*).

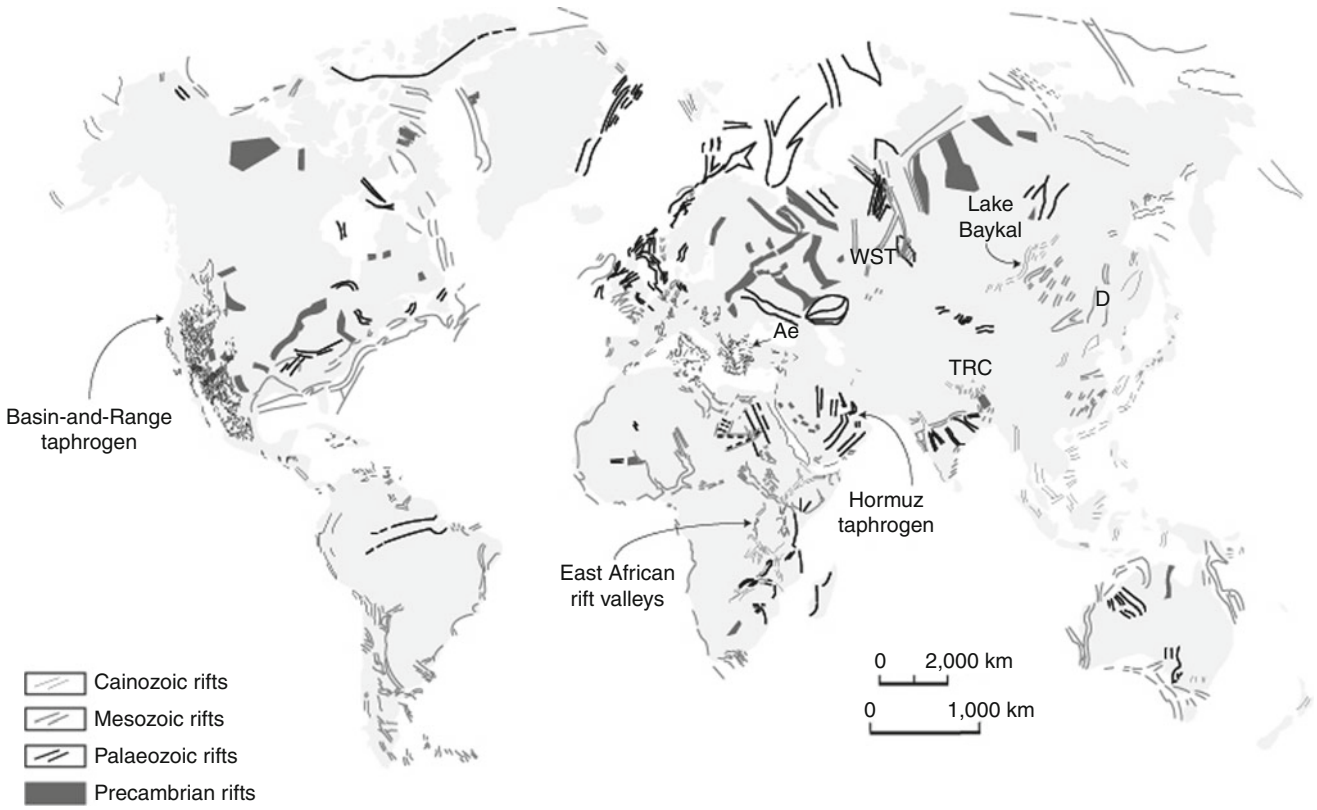
The straight fault hypothesis is difficult to generalize, especially where the master fault families bounding rifts are strongly curved in map view such as those that created the Lake Tanganyika rift system (Burgess et al., 1988). It is possible that individual segments of large faults rupture as straight fractures, but eventually the entire fault acquires a listric geometry. This seems supported by the fact that in Lake Tanganyika, where the rift floor has deeply subsided, faulting within the hanging wall block is less intense (Burgess et al., 1988). This is incompatible with a major kink of the master fault at depth. Wernicke and Burchfiel (1982) pointed out that where numerous, parallel-striking planar normal faults have rotated in unison around horizontal axes, one listric fault at the head of the series is a kinematic necessity to allow the others to rotate.

Asymmetric rifts are said to "face" in the direction of the dip of their master faults (Figure 3). Many major rifts are made up of half rifts that change their facing along the strike. In such cases, "transfer" or "accommodation zones" form along their boundaries that take the form of areas of complex oblique faulting and crustal block rotations (see especially the excellent discussions on transfer and accommodation zones in Faulds and Stewart, 1998; also Derer et al., 2005). Smaller and shallower transfer faults, essentially tear faults disrupting the hanging wall block with complicated displacement histories, form where the hanging wall is differentially stretched along the strike of the rift (cf. Şengör, 1987).

Rift shoulders are commonly uplifted. In symmetric rifts, both shoulders go up because of isostatic adjustment of the crustal wedges that are formed by the master faults dipping towards the rift (see *Isostasy*). Such uplift not only rotates the land surface on the shoulders away from the rift thus guiding much of the regional drainage in the same direction, but it rotates also the bounding normal faults (Buck, 1988). Each increment of faulting further rotates the shoulder and further rotates the previously formed fault segment. In Figure 3, in the case of a young asymmetric rift, the fault is shown to be straight. As the rift age and extension advance, the fault surface curves away from the rift as a result of incremental rotations and finally its oldest parts acquire a flat position, possibly with "horses" of the hanging wall block stuck on it as



Continental Rifts, Figure 4 A part of the eastern ("Gregory") rift in Kenya showing the intense faulting of the down-dropped hanging wall block between the master fault families. (Google Earth image.)



Continental Rifts, Figure 5 Distribution of continental rifts in the world. (Updated from Şengör and Natal'in, 2001.)

"extensional allochthons" (Figure 3). This seems to be the origin of many near-horizontal normal faults in areas of thin lithosphere and large amounts of extension such as the extensional metamorphic core complexes of the North American Cordillera (Armstrong, 1982), or those of the Tethysides (Verdel et al., 2007), although the existence of "primary" low-angle normal faults is also known both from detailed field studies (e.g., Miller and John, 1988) and from the seismicity (Eyidoğan and Jackson, 1985; Abers et al., 1997). The amount of extension in many rifts, especially those that are symmetric, do not go that far, however, and their bounding faults rarely dip less that 30°.

In places where intracontinental rifting along a narrow rift valley has advanced far, as, e.g., along the Wonji Fault Belt in Ethiopia, it has been noted that faulting is driven by dyke injection below the brittle upper carapace of about 10 km thickness and that segments with active deformation at the surface correlate with those of active magmatism (Kendall et al., 2005; Casey et al., 2006).

Cloos (1939) believed that the shoulder uplift in symmetric rifts resulted from the original doming that had led to keystone drop of the hanging wall block, but it was quickly realized that the observed angles of slope on crustal domes rarely exceed 1°, whereas to give rise to substantial rifts one needs about 10° dome slopes, which are excessive. Table 1 lists values of extension and shoulder uplift in some selected rifts. In none, except possibly in the Española Basin, is the shoulder uplift sufficient to cause the estimated amount of stretching.

Sedimentation in rifts

Because rifts form elongate basins they are ideal sediment receptacles. The type of sediment to be found in a rift varies according to the size and geometry of the rift, the climatic zone(s) in which it is located and its evolutionary history. Although most rifts are filled slowly by sedimentation in the sea, lakes, rivers and, in a subordinate degree, subaerially by wind, some are catastrophically invaded by the sea by breaching the barriers of sub-sea-level rift complexes (e.g., the Permian North Sea or the medial Cretaceous South Atlantic). In such cases, either vast evaporite deposits (as in the Aptian of the South Atlantic: Burke and Şengör, 1988) or thick shales (as in the late Permian of the North sea: Evans et al., 2003) are laid down in rift troughs.

Subaerial rifts are normally fed by rivers that locally create lakes within them because of endorheic conditions. Near the fault-bounded shoulders, stream valleys are commonly short (as the crest of the shoulder often is also the water-divide) but the streams in them are energetic because of the steep slopes. They thus carry clastic sediments of a variety of sizes that are horizontally sorted in large alluvial fans at the feet of the fault-escarpments. Along these escarpments, many alluvial fans form and coalesce and ongoing subsidence leads to the accumulation of large thicknesses of fanglomerates, sandstones, and shales near the faulted rift margin(s). Because of rotation towards the fault plane of the basin floors, the sedimentary thicknesses systematically decrease away from rift shoulders towards rift median lines. Steep slopes and catastrophic rain in both tropical and desert climates frequently lead to slope failures resulting in landslides complicating the sedimentary record. Permanent or ephemeral lakes may occupy the middle parts of rift floors. Ephemeral lakes commonly occur in dry climates in the form of playas and deposit evaporites and clays that may interfinger with marginal alluvial fan deposits. Permanent and deep lakes are important sediment repositories in rifts. They may accumulate very thick sedimentary sections in them, exhibiting complex facies dictated by the evolution of the geomorphology of the rift and its surrondings. Such lakes can be very deep (Lake Baykal in Sibearia: with 1,637 m, the world's deepest lake in addition to an 8 km rift sedimentary fill; Lake Tanganyika in

Continental Rifts, Table 1 The amount of stretching and shoulder uplift in some selected rifts (From Şengör, 2001)

Name of rift	Amount extended (km)	Uplift of shoulder area (km)
N Kenya	35–40	>1.4–1.7
Gregory	10	2
Malawi	7.5 (from transfer fault offset)	2.8 (strike-slip affected?)
Suez N	16.3	1.5
Central	17.7–32.6	1
S	29.3 (all from fault geometry)	~500 m
Upper Rhine	17 (from crustal configuration), 5–7 (from fault geometry)	2.2
Oslo	28–36 (from crustal configuration), 11–13 (from fault geometry)	1
Viking	100–130 (from crustal configuration), 30 (from fault geometry)	3
Central (North Sea)	100–105 (from crustal configuration), 15 from fault geometry)	1.5
Benue	100[a]	?
Baykal	15–20 (from crustal configuration), 10 (from fault geometry)	2–3
Española basin	5.5	1.5
Albuquerqe basin N	10 (from fault geometry)	~1
Albuquerqe basin S	16 (from fault geometry)	0.9

[a]Benkhelil et al. (1988) established at least 10 km of shortening during the following compression

East Africa: 1,457 m, but with half the sediment thickness of Lake Baykal). They are well-aerated in cold climates because the cold upper layers of the lake waters make convection and overturning possible (as in Lake Baykal), but they are stratifed with anoxic deeps in hot regions because hot surface waters cannot sink and thus inhibit overturning (as in Tanganyika). As the anoxic conditions in Lake Tanganyika exemplify, deep rift lakes in hot climates may generate good hydrocarbon source rocks in reducing environments. This is corroborated by ancient rift fills that formed under similar conditions such as the Permian oil-bearing sequences in the Junggar and Turfan rift basins in Central Asia (e.g., Graham et al., 1990; Miao et al., 2006) or the organic-rich (up to 8% organic C) Triassic-Jurassic lacustrine sediments of the Newark-type rifts of the eastern USA (Manspeizer and Olsen, 1981).

Rift valley floors are seldom level along the strike of the master bounding faults and this commonly induces longitudinal drainage in rifts. In fact, most of world's major rivers flow in two kinds of valleys: those that are located in rift troughs or in depressions inherited from them and those that are placed in fore- and hinterland basin troughs (cf. Audley-Charles et al., 1977). The Rhine in Europe, the Irrawady in Asia, flow along their almost entire lengths in active rift troughs. Along the western arm of the East African rift valleys, Lake Tanganyika receives (surface elevation at 773 m a.s.l.) the excess waters of Lake Kivu (at 1,460 m) by means of the south flowing river Ruzizi that came into being only 10,600 years ago and may have stopped flowing between 8 and 6,000 years ago (Felton et al., 2007), showing the sensitive dependence of rift sedimentation on the global climate. Farther north, just beyond Kivu, Lake Edward (at 912 m) is connected to Lake Albert (at 619 m) through the northerly flowing river Semliki. Lake Albert in turn feeds the Nile northwards via the river Albert Nile (White Nile or "Bahr al Jebel," i.e., "the river of the mountain").

In the sedimentary record, the rift facies is commonly characterized by thick (2–15 km) clastic sequences with rapid thickness and facies changes both laterally and vertically. If the rifts are underwater and far from clastic sources, they may house carbonate and even siliceous sedimentary sections as illustrated by the Alpine Triassic-early Jurassic facies. There, a continuous passage from late Triassic shallow water carbonates via a horizon of extensional-faulting-related neptunian dykes filled with progressively deepening facies and finally covered entirely by radiolarian cherts indicates a stretching basin floor as it deepened. That no shoulder uplift is indicated by any unconformities suggests that the Alpine rift was asymmetric with a major south-dipping detachment, and that the rocks on the hanging wall block never experinced normal-fault-related upheaval. If catastrophic marine invasion occurs, then sediments at once blanket the preexisting topography.

For good reviews of sedimentation in rifts, see Frostick et al. (1986), Lorenz (1988), Olsen and Schlische (1990), Leeder (1995), Şengör (1995), Beratan (1996), also the section entitled "Rift Basins" in Einsele (2000, pp. 543–560).

Magmatism in rifts

Rift magmatism is extremely variable because the tectonic environments in which rifts form are so diverse (see below). Rifts that form above mantle plumes (see *Mantle Plumes*) in plate interiors away from plate boundaries may begin their activity with restricted volumes of alkalic basalts indicating limited partial melting of mantle material at depth and may evolve towards abundant tholeiitic effusions as the temperature at depth increases and more and more mantle rock is melted, as, for example, illustrated by the evolution of the Afar plume in east Africa. Most geologists take alkalic to peralkalic magmatism as indication of extensional tectonics. The frequent association of nepheline syenites, kindred alkaline igneous rocks, and carbonatites (alkalic rocks and carbonatites [ARCs]) with intracontinental rifts has been taken by Burke et al. (2003, 2008) to argue that deformed rocks of the same compositions (deformed alkalic rocks and carbonatites [DARCs]) that have been shown to characterize suture zones in regions in Asia, Europe, Africa, and North America are nothing but ARC rocks that erupted into intracontinental rifts that became involved in a Wilson cycle of ocean opening and closing. This has been shown to form a powerful tool to map suture zones where other suture indicators such as ophiolites or subduction-related rocks may be missing. However, to think that rifts generate only alkalic magmatic rocks is clearly too simplistic. Those rifts that form along magmatic arc systems, for instance, inevitably have calc-alkalic magmas and those that form above recently thickened continental crust may generate vast quantities of calc-alkalic andesitic/rhyolitic melts. Rifts forming in back-arc regions, such as those in the Aegean, contain volcanic rocks showing evidence for a metasomatised mantle beneath them (e.g., the so-called Kulaites, i.e., hornblende basalts, in western Turkey: Richardson-Bunbury, 1996). Large amounts of silica-rich magmatic rocks may be generated if mantle plumes manage to melt large quantities of continental crust while ascending or if fractional crystallization is allowed to proceed undisturbed (e.g., the Cretaceous tin granites in Nigeria or the Cainozoic granites in Scotland; see Wiart and Oppenheimer, 2004 for an acitve example in Afar). Some rifts such as Tanganyika or Baykal are almost entirely devoid of magmatism, while others such as the late Palaeozoic Oslo Rift have floors made up entirely of magmatic rocks. Some Atlantic-type continental margins, which grew out of disrupted rifts, are called "volcanic," whereas others, "non-volcanic," because they descended from rifts with similar magmatic characteristics.

It is naive to think that lithospheric extension generates only certain kinds of magmatic rocks without considering the tectonic environments that gave rise to the extension and in which the extension proceeds. It is therefore

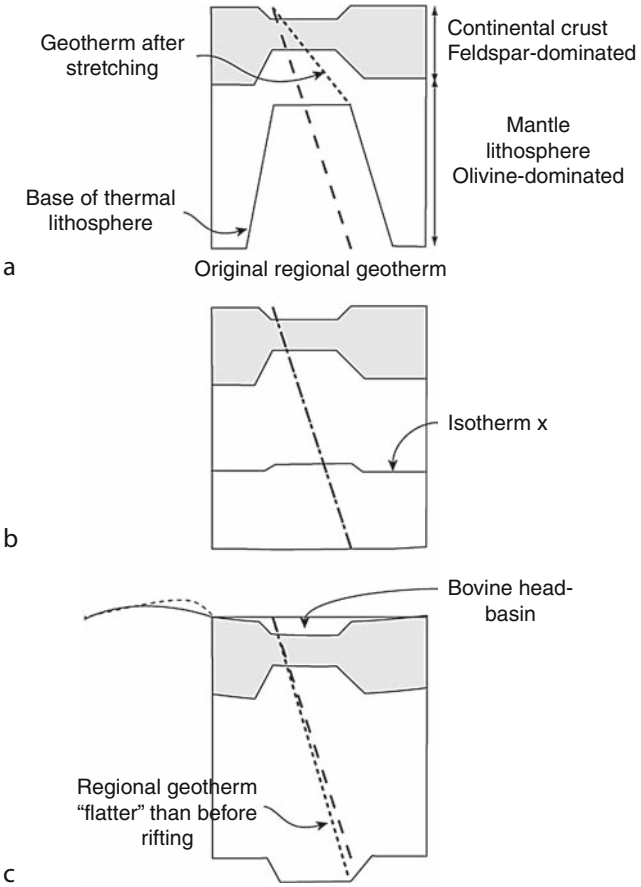
imperative to look at the diversity of the environments which generate rifts.

For good reviews of magmatism in rifts, see Kampunzu and Lubala (1991), Kinnaird (1998), Ulrych et al. (1999), Burke et al. (2003, 2008), Wilson et al. (2004), Yirgou et al. (2006); also the special section on magmatism and extension in *Journal of Geophysical Research*, vol. 100 (1995).

Metamorphism in rifts

Continental rifts are not commonly thought of as prime loci of metamorphic events, but four different kinds of metamorphic rocks do occur in them. In increasing importance: (1) contact metamorphic rocks around rift-related magmatic rocks, (2) hydrothermal metmorphism around vents, (3) burial metamorphism due to large sediment thicknesses, and (4) metamorphic rocks that form within the continental crust and are brought to surface by rifting.

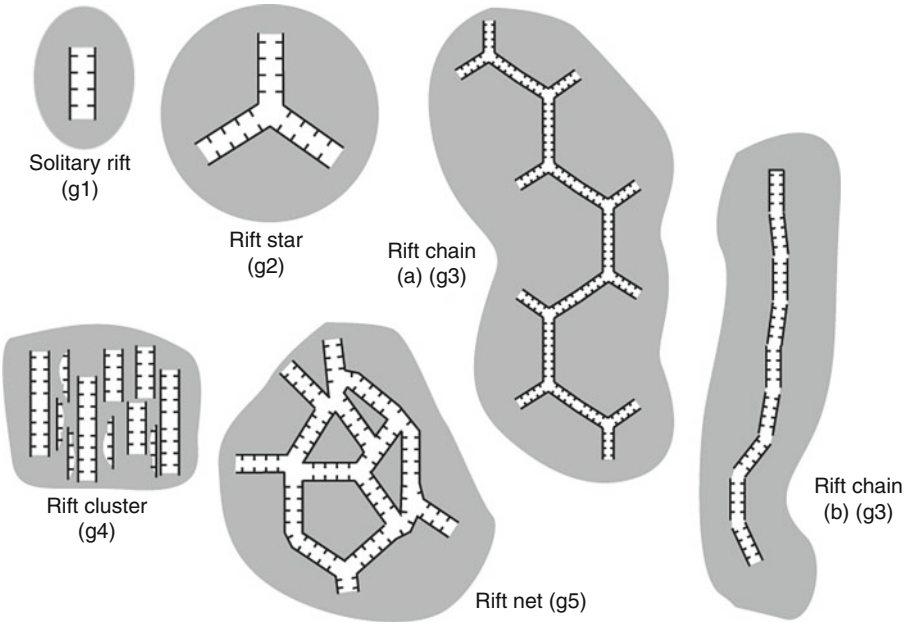
There is little to say about the first kind of metamorphic rocks in rifts as they are no different from any other contact metamorphic rock. Hydrothermal metamorphism is important in many rifts because of the high heat flow and the continuous opening of extensional fractures allowing hot waters to circulate (e.g., Crane et al., 1991). Burial metamorphism is not common in rifts unless they accumulate very large thicknesses of sediment quickly, sufficient to raise the temperature and pressure conditions to initiate metamorphism. If we assume a normal geothermal gradient (~3°C/100 m; see *Heat Flow, Continental*), an accumulation of about 4 km of sediment would be sufficient to initiate metamorphism by raising the temperature beyond 120°C. Jackson (1987) argued, however, that because earthquake hypocenters in rifts are seldom deeper than 10 km, at least beyond that depth rocks must be metamorphic. But since there are earthquakes down to 30 km in the east African rifts system, it is questionable how sound an argument that is. It is true, however, that rift sediments are seldom metamorphosed beyond greenshist grade because they are seldom stretched enough and seldom accumulate sufficient thicknesses of sediment to heat the rocks beyond 320°C. Smith (1976) pointed out, for example, that at the bottom of the about 16 km-thick Gulf Coast sediment pile, the temperature may be just about 320°C under 4 kb pressure. But this ignores both the radioactive heating of the pile (see *Radiogenic Heat Production of Rocks*) and raising the geotherm by stretching. However, even when the crust is stretched, every particle in the stretched crust must cool, unless the rift trough is rapidly sedimented (Figure 6; McKenzie, 1978; Jackson, 1987). Although it has been suggested that some high-grade metamorphic rocks did form in a rift environment (Wickham and Oxburgh, 1985), it was a result of mistaking an orogenic environment for a taphrogenic one, although from elsewhere allegedly rift-related metamorphism creating high-grade gneisses has been reported on the basis of condensed metamorphic zones and absence of associated plutonism



Continental Rifts, Figure 6 Thermal regime of rifts with a simple geotherm. (**a**) Immediately after the stretching, both the crust and the lithospheric mantle thin and isotherms become crowded under the rift increasing the heat flow in the rift. Although no particle in the rift basement will become any hotter. (**b**) After stretching ceases, isotherms begin to subside. (**c**) When the lithosphere "recovers," the rift will end up with more lithospheric mantle beneath it and will have a heavier substratum that will subside and pull its surroundings with it creating an "intracratonic basin" overlying the rift. This is called a "bovine head" cross-section because of its resemblance to a long horn ox. While the basin subsides, its margnis will be elastically uplifted (solid line above **c**), but may relax viscoelastically in time (dotted line above **c**). Here the margins are shown only on one side for graphic convenience, but naturally the events will take place all around the basin.

(e.g., St-Onge and King, 1987). However, I am unaware of any well-documented metamorphic terrane that formed as a consequence of rift burial.

By contrast, large normal faults of the kind shown for the advanced asymmetric rifts in Figure 3 may pull up rocks from the lower continental crust that may have been metamorphosed even up to granulite facies as shown by many extensional metamorphic core complexes. But in this case, the metamorphism is not *caused* by rifting, but it is *exposed* by rifting. In fact, as the rocks ascend



Continental Rifts, Figure 7 The geometric classification of taphrogens. (After Şengör, 1995.)

in the footwall, they get retrograded, which is directly caused by rifting.

For metamorphism in rifts, see Armstrong (1982), Jackson (1987), Lister and Davis (1989), Block and Royden (1990), Verdel et al. (2007).

Kinds of rifts

A classification of rifts must serve to answer the questions: how many different sorts of environments lead to lithospheric extension and what kinds of rifts form in these environments? In essence, one must specify where and how taphrogens commonly form. If taphrogeny stops before producing ocean, it causes subsidence and creates large basins (called "koilogens" by Spizaharsky and Borovikov, 1966, pp. 113ff.) overlying taphrogens (Figure 6; cf. McKenzie, 1978; Şengör, 1995). A comprehensive synthesis of various taphrogen types does not yet exist. Below is a restatement of Şengör's (Şengör, 1995; Şengör and Natal'in, 2001) classification of types of rifts that make up the main elements of taphrogens. That classification is hierarchical and goes from pure geometry to dynamics (Figure 8). Its hierarchical nature allows the environment and the path of formation of a given rift to be determined. For numerous examples of each of Şengör's categories, see Şengör and Natal'in (2001). In this section, references already given in Şengör (1995) and in Şengör and Natal'in (2001) are not repeated.

Şengör's classification has three different categories that do not completely overlap, namely *geometric*, *kinematic*, and *dynamic*. In the following, the three different categories are identified with their initials, i.e., *g*, *k*, and *d*, respectively.

Geometric classification of rifts

Rifts or groups of rifts forming individual taphrogenic units display five kinds of fundamental map patterns (see Figures 7 and 8). From simplest to more complex, these are:

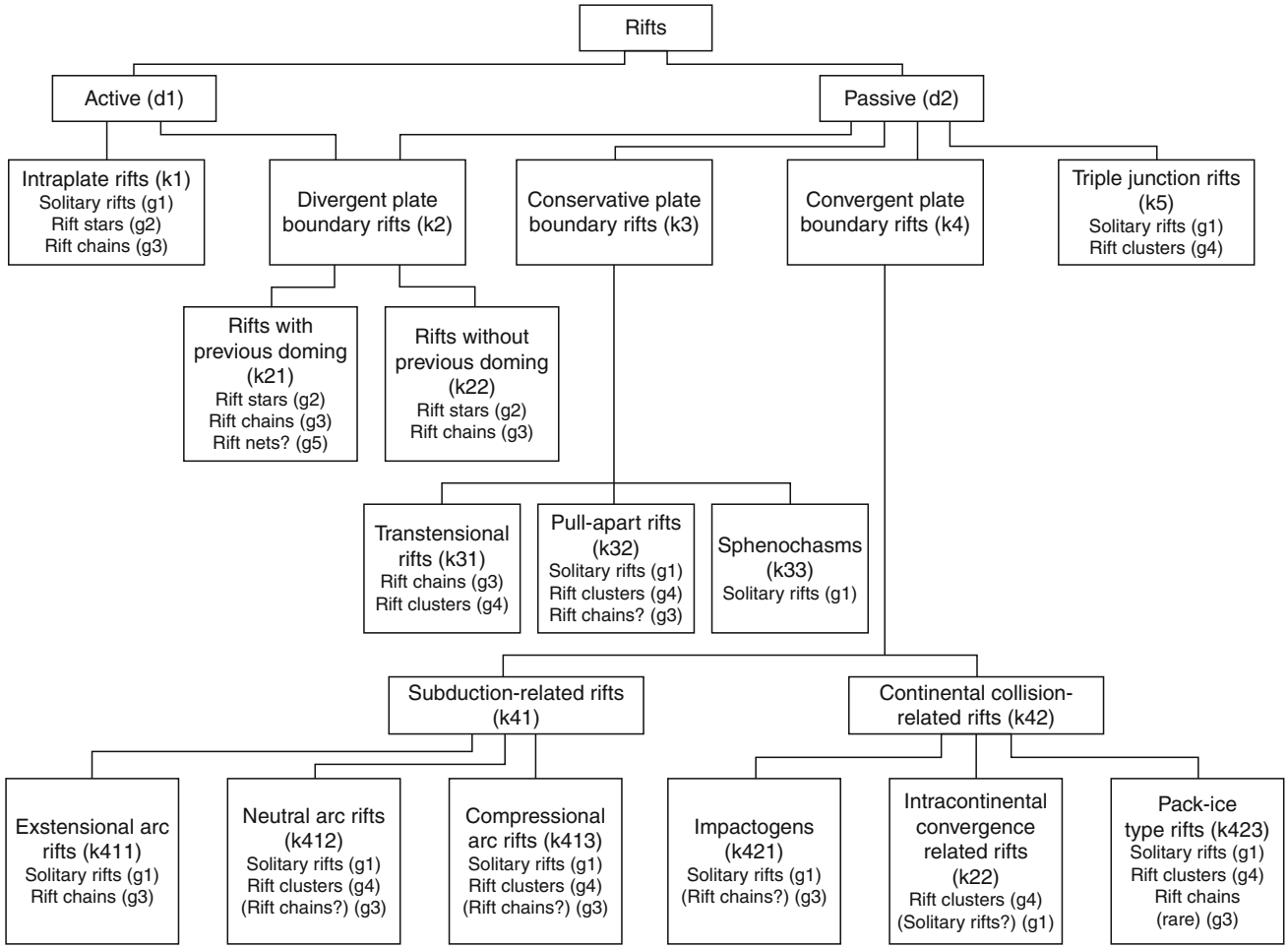
g1 – Solitary rifts: Solitary rifts form small, fairly insignificant, and very rare taphrogens and are extremely difficult to ascertain in the geological record because it is commonly hard to tell whether a given rift fragment is isolated or part of a larger taphrogen (for a comparable difficulty in shortening structures, imagine a big solitary fold or a large thrust fault!).

g2 – Rift stars: Rift stars form when more than two rifts radiate away from a common center, building a roundish taphrogen. Rift stars are very common features of the structural repertoire of our planet today.

g3 – Rift chains: When several rifts are aligned end-to-end along linear/arcuate belts of rifting, they form *rift chains*. The East African Rift System constitutes the best-known active rift chain in the world. Solitary rifts or rift stars or combinations of these may be connected to form different kinds of rift chains (e.g., Figure 7 rift chains a or b).

g4 – Rift clusters: When several subparallel rifts occur in roughly equant areas, they are said to form a *rift cluster*. The two best-known active rift clusters in the world are the Basin-and-Range extensional area in North America and the Aegean Sea and the surrounding regions (see Figure 5).

g5 – Rift nets: Rift nets constitute a rare pattern, which comes about when rifts form a roughly checkered pattern as in the Proterozoic basement of the East European platform or in the late Mesozoic in central North Africa (cf. Şengör and Natal'in, 2001). They resemble



Continental Rifts, Figure 8 The kinematic classification of rifts and taphrogens. (After Şengör, 1995.)

chocolate-bar boudinage, as seen in the Proterozoic basement of eastern Europe (Figure 5), and may have a similar origin, but more commonly rift nets form in complex and rapidly shifting stress environments in which dominant extension directions change fast. Many rift nets in fact may represent two superimposed rift clusters.

Kinematic classification of rifts

Because rifts are ubiquitous in all stages of the Wilson Cycle of ocean opening and closing, the kinematic characteristics of the plate boundaries have been taken as a basis for classifying them according to the environment of the overall displacement and strain in which they form (see Figure 8). There are three types of plate boundaries plus the plate interiors, to which four types of rift families correspond. In addition, incompatibilities arise around some unstable intracontinental triple junctions leading to complex rifting that should be treated separately from the other four classes, thus creating a fifth kinematic class, called "triple-junction rifts."

k1 – Intraplate rifts: Rifts surrounded entirely by undeformed lithosphere occupy this category. Such rifts are usually solitary, small, and rare (the difficulty in forming them is analogous to that forming a solitary fold or a nappe surrounded by entirely undeformed terrain), and are not easy to identify in the geological history. The Lake George and Lake Champlain rifts in the northeastern USA are active examples (see *Seismicity, Intraplate*).

k2 – Rifts associated with divergent plate boundaries: These rifts result from plate separation along nascent extensional boundaries. All the Cainozoic rifts in east Africa belong here. This category of rifts may be further subdivided into two classes as follows:

k21 – Rifts that form following an episode of doming: the divergent boundary along which rifts form is in this case preceded by an episode of lithospheric doming. The East African Rift Valleys are the best-known examples of such rifting (see *Isostasy, Thermal*).

k22 – Rifts that form with no pre-rift doming: in this case, rifts form without a prelude of uplift, as is the case in the Salton Trough and the Gulf of California.

A good fossil example is the rifting of the Alpine Neo-Tethys in the earlier Mesozoic.

k3 – Rifts that form in association with conservative plate boundaries: conservative, i.e., transform fault, boundaries are those along which neither extension nor shortening is required by the regional slip vectors. However, various reasons conspire to induce both extension and shortening to occur along considerable stretches of these boundaries. Rifts along conservative plate boundaries form in three different settings:

k31 – Transtensional conservative boundaries: if a conservative boundary is opening up all along its length because of a component of extension, it is called *transtensional*. Many active rifts have a transtensional component, and fossil examples of such rifts may be recognized largely through the structures they contain as shown by Olsen and Schlische (1990). Dewey (2002) gave an exhaustive analysis of the strain that develops in such rifts.

k32 – Pull-apart basins along conservative boundaries: major strike-slip faults commonly have bends along them that either facilitate ("releasing bends") or obstruct ("restraining bends") slip along them. Extensional basins form along the releasing bends, in which the magnitude of extension equals the magnitude of cumulative strike-slip offset along the strike-slip fault since the formation of the releasing bend. Such basins are called "pull-apart basins." Crowell's (1974) fault-wedge basins are nothing more than special cases of pull-apart basins.

k33 – Sphenochasms: not all basins created by secondary extension associated with strike-slip faults are pull-apart basins. Some represent tears caused by either an asperity or differential drag along the strike-slip fault in one of the fault walls, in which the amount of extension changes from a maximum along the fault to zero at the pole of opening of the tear-basin. S. W. Carey called such wedge-shaped rifts that open towards a major strike-slip fault *sphenochasms*.

k4 – Rifts that form in association with convergent plate boundaries: a large family of rifts forms in association with convergent plate boundaries. In this group, a first-order subdivision is between rifts associated with subduction zones and rifts associated with continental collision, although this may artificially split some genetic groups, such as those rifts that presumably form because of the tension generated by excessive crustal thickening. The usefulness of the present grouping is that it enables a rapid overview of the presently active rift environments and comparison with ancient ones.

k41 – Rifts associated with subduction zones: environments of rifting associated with subduction zones correspond to three different types of arc behavior, namely, extensional, neutral, and compressional arcs.

k411 – Rifts associated with extensional arcs: an extensional arc generally splits along the magmatic axis (if such an axis is already in existence) forming a small rift chain. Such a situation is today known from both the Okinawa rift and the Izu-Bonin arc system. Such rifts commonly do not get preserved intact, both because of the complications of the tectonic evolution of arcs involving common changes of behavior and because of later collisions with other arcs or continents.

In extensional arcs, rifts also develop orthogonal to arc trend owing to the extension of the arc as it bows out in front of an expanding marginal basin (as, for instance, in Crete).

k412 – Rifts associated with neutral arcs: neutral arcs have neither shortening nor extension across them. Therefore, the only rifts that may form in neutral arcs are those associated with arc-parallel strike-slip faults, which may be classified in the same way as the rifts that form along conservative plate boundaries. More complex rift basins may originate along such arc-parallel strike-slip faults if the sliver plate in the forearc area and its various pieces rotate about vertical axes.

Pull-apart basins in arcs are difficult to recognize, but the Sumatra Fault has several well-developed examples along it.

Sphenochasms along strike-slip faults in arcs are rarer still. Davis et al. (1978) have discussed two possible examples: the more recent of which may have created the "Columbia Embayment" by motion along the Straight Creek fault in the latest Cretaceous and the earliest Cainozoic.

k413 – Rifts associated with compressional arcs: in compressional arcs crust commonly thickens and lithosphere thins, both by heating and by eventual delamination. The arc becomes shortened across and elongated along its trend. This elongation commonly generates rifts at high angles to the trend of the arc. Rifts on the Altiplano in the Andes are examples of such high-angle rifts.

k42 – Rifts associated with zones of continental collision: three different environments of rifting form associated with the collision of continents: (1) *Lines of extension* that radiate from points at which collision commences, (2) *Regions of extension* abutting against sutures, and (3) *Nodes of extension* in areas of complex deformation in fore- and hinterlands shattered by collisions. Impactogens (k421), rifts forming in intracontinental convergence belts (k422), and pack-ice-type rifts (k423) correspond with these three environments, respectively.

k5 – Triple-junction rifts: Triple-junction rifts form at or near unstable intracontinental triple junctions, at which plate evolution dictates the generation of "holes" because continental lithosphere cannot be continuously subducted.

Dynamic (genetic) classification of rifts

Rifts are also classified according to the origin of forces (see *Geodynamics*) that lead to rifting (see Figure 6).

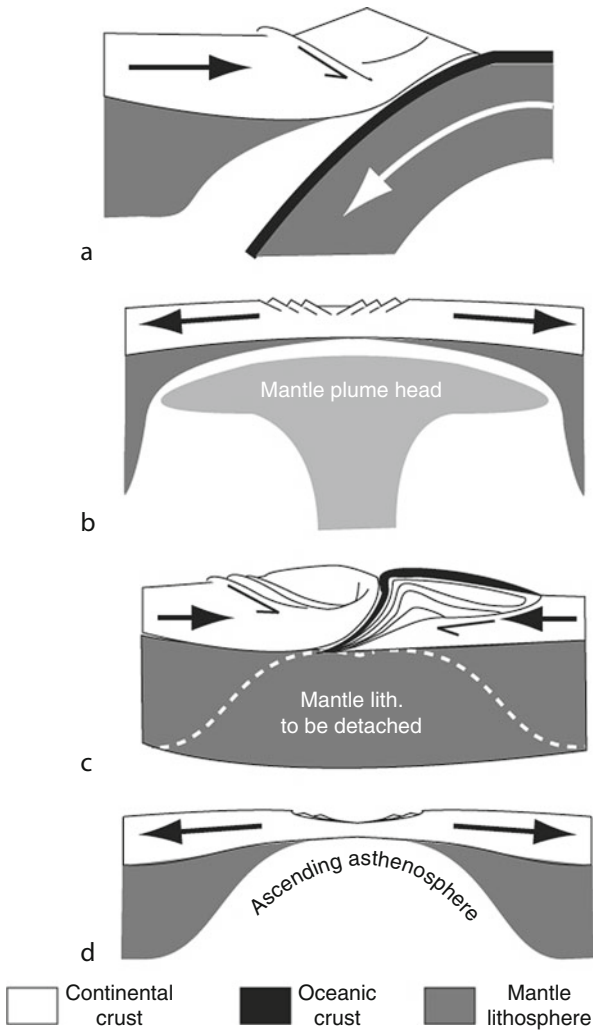
Şengör and Burke (1978) proposed that stresses which cause rifting may be imposed on the lithosphere directly by the mantle beneath it (cf. Şengör, 2001) or they may result from two-dimensional plate evolution. Accordingly, they termed these two modes of rifting "active" and "passive." Gans (1987) tried to replace these terms with "open-system' and "closed-system" rifting, respectively, which, however, did not find general acceptance.

d1 – Active rifting: "Active rifting" is rifting caused by mantle upwelling (see *Earth, Density Distribution*; *Mantle Convection*) associated with plumes in the mantle (see *Mantle Plumes*). Two views have been advanced to explain the origin of the extension in domes rising above hot-spot jets: one ascribes the rifting to basal shear stresses induced by a spreading plume head beneath a dome. The other holds the potential energy of the rising dome responsible for driving the rifting. All of these factors probably do contribute to maintaining the active rifting process at its habitually slow pace of considerably less than 1 cm/a (see *Geoid*).

Plume-caused taprogens may be also termed deeply rooted rifts and would be the extensional counterparts of subduction-related orogens (Figure 9a, b). These can be contrasted to collisional orogens that have lost their subductive "anchors" and the passive rifts with no deep mantle connections (Figure 9c, d).

d2 – Passive rifting: in the passive rifting mode, extension is caused by the two-dimensional motions of the lithospheric plates and not by an autonomous interference from the mantle (Figure 9d). In this mode of rifting, there is no pre-rifting doming associated with a hot-spot (Şengör and Burke, 1978). Kinematic mechanisms reviewed above under the headings k22, k31, k32, k33, k411, k412, k413, k421, k422, k423, and k5 all may form rifts in a "passive rifting mode."

There is only one kind of rift this classification does not consider: rifts that form by propagating from an already existing rift. Since propagation may take many forms, it might be sufficient to indicate such rifts with the notation d2 to indicate their passive mode of opening. The regional geologist may use Figure 8 as a "flow chart" to follow the evolutionary histories of the various kinds of rift basins he encounters.



Continental Rifts, Figure 9 Deeply rooted megastructures. (**a**) Subduction-controlled orogens; (**b**) mantle-plume controlled, i.e., active rifts and rootless megastructures; (**c**) collisional orogens; (**d**) plate motion-controlled, i.e., passive rifts.

Conclusions

Although they constitute one of the three great continental structure types and have serious relevance to human well-being, rifts remain poorly known. This is largely because there are still important technological difficulties in studying them directly because of the large sediment thicknesses they contain and because their diversity is little appreciated. They are so ubiquitous in all tectonic environments that they tend to be overlooked in favor of other, rarer, structures that may appear more prominent. It is commonly not recognized that all kinds of structures, such as magmatic arcs, collisional orogens, keirogens, intracratonic basins all have associated rifts and, consequently, people wishing to know about rifts tend not to concern themselves with them. A stereotypic rift that commonly bedevils textbooks is one that disrupts a craton and has alkalic or peralkalic vulcanism with a thick, mainly nonmarine sedimentary fill, although there are many more rifts in the world than those with such characteristics. What is now needed in rift studies is a comprehensive synthesis that takes into account this diversity.

Bibliography

Abers, G. A., Mutter, C. Z., and Fang, J., 1997. Shallow dips of normal faults during rapid extension: earthquakes in the Woodlark-d'Entrecasteaux rift system, Papua New Guinea. *Journal of Geophysical Research*, **102**, 15,301–15,307.

Armstrong, R. L., 1982. Cordilleran metamorphic core complexes – from Arizona to Southern Canada. *Annual Review of Earth and Planetary Science Letters*, **10**, 129–154.

Audley-Charles, M. G., Curray, J. R., and Evans, G., 1977. Location of major deltas. *Geology*, **5**, 341–4.

Beratan, K. K. (ed.), 1996. Reconstructing the history of basin and range extension using sedimentology and stratigraphy. Geological Society of America special paper 303, vi + 212 pp. + 1 folded plate in pocket.

Bernoulli, D., Manatschal, G., Desmurs, L., and Müntener, O., 2003. Where did Gustav Steinmann see the trinity? Back to the roots of an Alpine ophiolite concept. In Dilek, Y., and Newcomb, S. (eds.), *Ophiolite Concept and the Evolution of Geological Thought*. Geological Society of America special paper 373, pp. 93–110.

Block, L., and Royden, L. H., 1990. Core complex geometries and regional scale flow in the lower crust. *Tectonics*, **9**, 557–567.

Blundell, D. J., and Gibbs, A. D. (eds.), 1990. *Tectonic Evolution of the North Sea Rifts*. Oxford: Clarendon.

Boillot, G., Recq, M., Winterer, E. L., Meyer, A. W., Applegate, J., Baltuck, M., Bergen, J. A., Comas, M. C., Davies, T. A., Dunham, K., Evans, C. A., Girardeau, J., Goldberg, G., Haggerty, J., Jansa, L. F., Johnson, J. A., Kasahara, J., Loreau, J. P., Luna-Sierra, E., Moullade, M., Ogg, J., Sarti, M., Thurow, J., and Williamson, M., 1987. Tectonic denudation of the upper mantle along passive margins: a model based on drilling results (ODP leg 103, western Galicia margin, Spain). *Tectonophysics*, **132**, 335–342.

Buck, R., 1988. Flexural rotation of normal faults. *Tectonics*, **7**, 959–973.

Burgess, C. F., Rosendahl, B. R., Sander, S., Burgess, C. A., Lambiase, J., Derksen, S., and Meader, N., 1988. The structural and stratigraphic evolution of Lake Tanganyika: a case study of continental rifting. In Manspeizer, W. (ed.), *Triassic-Jurassic Rifting – Continental Breakup and the Origin of the Atlantic Ocean and Passive Margins B*. Amsterdam: Elsevier, pp. 859–881.

Burke, K., 1996. The African plate. *South African Journal of Geology*, **99**, 339–409.

Burke, K., and Dewey, J., 1973. Plume-generated triple junctions: key indicators in applying plate tectonics to old rocks. *Journal of Geology*, **81**, 406–433.

Burke, K., and Gunnell, Y., 2008. The African erosion surface: a continental-scale synthesis of geomorphology, tectonics, and environmental change over the past 180 million years. Geological Society of America memoir 201, iv + 66 pp.

Burke, K., and Şengör, A. M. C., 1988. Ten metre global sea-level change associated with South Atlantic Aptian salt deposition. *Marine Geology*, **83**, 309–312.

Burke, K., and Wilson, J. T., 1976. Hot spots on the earth's surface. *Scientific American*, **235**, 46–57.

Burke, K., Ashwal, L. D., and Webb, S., 2003. New way to map old sutures using deformed alkaline rocks and carbonatites. *Geology*, **31**, 391–394.

Burke, K., Khan, S. D., and Mart, R. W., 2008. Grenville Province and Monteregian carbonatite and nepheline syenite distribution related to rifting, collision, and plume passage. *Geology*, **36**, 983–986.

Casey, M., Ebinger, C., Keir, D., Gloaguen, R., and Mohamed, F., 2006. Strain accommodation in transitional rifts: extension by magma intrusion and faulting in Ethiopian rift magmatic segments. In Yirgou, G., Ebinger, C. J., and Maguire, P. K. H. (eds.), *The Afar Volcanic Province Within the East African Rift System*. Geological Society special publication 259, pp. 143–163.

Cloos, H., 1939. Hebung-Spaltung-Vulkanismus – Elemente einer Geometrischen Analyse Irdischer Großformen. *Geologische Rundschau*, **30**(Zwischenheft 4A), 405–527.

Corti, G., Cuffaro, M., Doglioni, C., Innocenti, F., and Manetti, P., 2006. Coexisting geodynamic processes in the Sicily Channel. In Dilek, Y., and Pavlides, S. (eds.), *Postcollisional Tectonics and Magmatism in the Mediterranean Region and Asia*. Geological Society of America special paper 409, pp. 83–96.

Coward, M. P., Dewey, J. F., and Hancock, P. L. (eds.), 1987. Continental extensional tectonics. Geological Society special publication 321 (Albert M. Quennell volume), 637 pp.

Crane, K., Hecker, B., and Golubev, V., 1991. Heat flow and hydrothermal vents in Lake Baikal, U.S.S.R. *EOS Transactions of the American Geophysical Union*, **72**, 585–588.

Crowell, J. W., 1974. Sedimentation along the San Andreas Fault, California. In Dickinson, W. R. (ed.), *Plate Tectonics and Sedimentation*. Society of Economic Paleontologists and Mineralogists special publication 19, pp. 292–303.

Davis, G. A., Monger, J. W. H., and Burchfiel, B. C., 1978. Mesozoic construction of the Cordilleran "collage," central British Columbia to central California. In Howell, D. G., and McDougall, K. A. (eds.), *Mesozoic Paleogeography of the Western United States*. Los Angeles: Society of Economic Paleontologists and Mineralogists. Pacific Coast Paleogeography Symposium 2, Pacific Section, pp. 1–32.

Derer, C. E., Schumacher, M. E., and Schäfer, A., 2005. The Upper Rhine Graben: basin geometry and early syn-rift tectonosedimentary evolution. *International Journal of Earth Sciences (Geologische Rundschau)*, **94**, 640–656.

Dewey, J. F., 1982. Plate tectonics and the evolution of the British Isles. *Journal of the Geological Society of London*, **139**, 371–412.

Dewey, J. F., 2002. Transtension in arcs and orogens. *International Geology Review (George A. Thompson Symposium volume)*, **44**, 402–439.

Dickinson, W. R., 2002. The basin and range province as a composite extensional domain. *International Geology Review (George A. Thompson Symposium volume)*, **44**, 1–38.

Durand, L., Jolivet, F., Horváth, F., and Séranne, M. (eds.), 1999. *The Mediterranean basins: Tertiary extension within the Alpine Orogen*. Geological Society special publication 156, 570 pp.

Einsele, G., 2000. *Sedimentary Basins – Evolution, Facies and Sediment Budget*. Berlin: Springer, Vol. 2 (completely revised).

Evans, D., Graham, C., Armour, A., and Bathurst, P., 2003. *The Millenium Atlas: Petroleum Geology of the Central and Northern North Sea*. London: The Geological Society of London.

Eyidogan, H., and Jackson, J., 1985. Seismological study of normal faulting in the Demirci, Alasehir and Gediz earthquakes of 1969–70 in western Turkey: Implications for the nature and geometry of deformation in the continental crust. *Geophysical Journal of the Royal Astronomical Society*, **81**, 569–607.

Faulds, J. E., and Stewart, J. H. (eds.), 1998. *Accommodation zones and transfer zones: the regional segmentation of the Basin and Range Province*. Geological Society of America special paper 323, vii + 257 pp. + 1 folded plate in pocket.

Felton, A. A., Russell, J. M., Cohen, A. S., Baker, M. E., Chesley, J. T., Lezzar, K. E., McGlue, M. M., Pigati, J. S., Quade, J., Curt Stager, J., and Tiercelin, J. J., 2007. Paleolimnological evidence for the onset and termination of glacial aridity from Lake Tanganyika Tropical East Africa. *Palaeogeography Palaeoclimatology Palaeoecology*, **252**, 405–423, doi:10.1016/j.palaeo.2007.04.003.

Frostick, L. E., 1997. The East African rift basins. In Selley, R. C. (ed.), *African Basins*. Amsterdam: Elsevier, pp. 187–209.

Frostick, L. E., Renaut, R. W., Reid, I., and Tiercelin, J. J. (eds.), 1986. Sedimentation in the African rifts. Geological Society special publication 25, 382 pp.

Gans, P. B., 1987. An open-system, two-layer crustal stretching model for the eastern Great Basin. *Tectonics*, **6**, 1–12.

Graham, S. A., Brassell, S., Carroll, A. R., Xiao, X., Demaison, G., McKnight, C. L., Liang, Y., Chu, J., and Hendrix, M. S., 1990. Characteristics of selected petroleum source rocks Xianjiang Uygur autonomous region, Northwest China. *American Association of Petroleum Geologists Bulletin*, **74**, 493–512.

Gregory, J. W., 1894. Contributions to the physical geography of British East Africa. *The Geographical Journal*, **4**, 289–315.

Harding, T. P., 1983. Graben hydrocarbon plays and structural styles. *Geologie en Mijnbouw*, **62**, 3–23.
Hüttner, R., 1991. Bau und Entwicklung des Oberrheingrabens - Ein Überblick mit historischer Rückschau. *Geologisches Jahrbuch*, **E48**, 17–42.
Ingersoll, R. V., 2001. Structural and stratigraphic evolution of the Rio Grande Rift, northern New Mexico and southern Colorado. *International Geology Review*, **43**, 867–891.
Jackson, J. A., 1987. Active continental deformation and metamorphism. *Philosophical Transactions of the Royal Society of London*, **A321**, 47–66.
Jackson, J., and Belkinsop, T., 1993. The Malawi earthquake of 10 March 1989: deep faulting within the east African rift system. *Tectonics*, **12**(5), 1131–1139.
Jaffé, F. C., 1971. Geothermal energy: a review. *Bulletin der Vereinigung Schweizerischer Petroleum-Geologen und - Ingenieure*, **38**, 17–40.
Kampunzu, A. B., and Lubala, R. T. (eds.), 1991. *Magmatism in Extensional Structural Setting – The Phanerozoic African Plate*. Berlin: Springer.
Kendall, J.-M., Stuart, G. W., Ebinger, C. J., Bastow, I. D., and Keir, D., 2005. Magma-assisted rifting in Ethiopia. *Nature*, **433**, 146–148.
Kinnaird, J. A. (ed.), 1998. Aspects of tensional magmatism. *Journal of African Earth Sciences*, **26**, 1–150.
Kravchinsky, V. A., Krainov, M. A., Evans, M. E., Peck, J. A., King, J. W., Kuzmin, M. I., Sakai, H., Kawai, T., and Williams, D. F., 2003. Magnetic record of Lake Baikal sediments: chronological and paleoclimatic implication for the last 6.7 Myr. *Palaeogeography Palaeoclimatology Palaeoecology*, **195**, 281–293.
Kuz'min, M. I., Karabanov, E. B., Kawai, T., Williams, D., Bychinsky, V. A., Kerber, E. V., Kravchinsky, V. A., Bezrukova, E. V., Prokopenko, A. A., Geletii, V. F., Kalmychkov, G. V., Goreglyad, A. V., Antipin, V. S., Khomutova, M. Y., Soshina, N. M., Ivanov, E. V., Khrusevich, G. K., Tkachenko, L. L., Solotchina, E. P., Ioshida, N., and Gvozdkov, A. N., 2001. Deep drilling on Lake Baikal: main results. *Russian Geology and Geophysics*, **42**, 3–28.
Lambiase, J. J. (ed.), 1995. *Hydrocarbon habitat in rift basins*. Geological Society special publication 80, 381 pp.
Landon, S. M. (ed.), 1994. *Interior rift basins*. American Association of Petroleum Geologists memoir 59, 276 pp.
Le Pichon, X., and Sibuet, J. C., 1981. Passive margins: a model of formation. *Journal of Geophysical Research*, **86**, 3708–3720.
Leeder, M. R., 1995. Continental rifts and proto-oceanic rift troughs. In Busby, C. J., and Ingersoll, R. V. (eds.), *Tectonics of Sedimentary Basins*. Oxford: Blackwell, pp. 119–148.
Lister, G. S., and Davis, G. A., 1989. The origin of metamorphic core complexes and detachment faults formed during Tertiary continental extension in the northern Colorado River region. U.S.A. *Journal of Structural Geology*, **11**, 65–94.
Littke, R., Bayer, U., Gajewski, D., and Nelskamp, S., 2008. *Dynamics of Complex Intracontinental Basins – The Central European Basin System*. Berlin: Springer.
Lorenz, J. C., 1988. *Triassic-Jurassic Rift-Basin Sedimentology – History and Methods*. New York: Van Nostrand Reinhold.
MacKenzie, A. S., and McKenzie, D., 1983. Isomerization and aromatization of hydrocarbons in sedimentary basins formed by extension. *Geological Magazine*, **120**, 417–470.
Manspeizer, W. (ed.), 1988. *Triassic-Jurassic Rifting – Continental Breakup and the Origin of the Atlantic Ocean and Passive Margins*. Amsterdam: Elsevier, Vol. 2.
Manspeizer, W., and Olsen, P. E., 1981. Rift basins of the passive margin: tectonics, organic-rich lacustrine sediments, basin analysis. In Hobbs, G. W. (ed.), *Field Guide to the Geology of the Palaeozoic, Mesozoic, and Tertiary Rocks of New Jersey and the Central Hudson Valley*. New York: Petroleum Exploration Society of New York, pp. 25–103.
McKenzie, D., 1978. Some remarks on the development of sedimentary basins. *Earth and Planetary Sscience Letters*, **40**, 25–32.
McQuarrie, N., and Wernicke, B. P., 2005. An animated tectonic reconstruction of southwestern North America since 36 MA. *Geosphere*, **1**, 147–172, doi:10.1130/GES00016.1.
Miao, J. Y., Kou, H. S., Zhou, L. F., and Han, Z. Y., 2006. Sedimentary environments of organic matter from Middle Permian source rocks in northern Xinjiang China. *Chinese Journal of Geochemistry*, **25**, 258–265.
Miller, J. M. G., and John, B. E., 1988. Detached strata in a Tertiary low-angle normal fault terrane, southeastern California: a sedimentary record of unroofing, breaching, and continued slip. *Geology*, **16**, 645–648.
Morley, C. K., 1999. *Geoscience of Rift Systems – Evolution of East Africa*. AAPG studies in geology 44. Tulsa: The American Association of Petroleum Geologists, 242 pp. + 5 Appendices + 16 pp. Index.
Olsen, P. E., and Kent, D. V., 1999. Long-period Milankovich cycles from late Triassic and early Jurassic of eastern North America and their implications for the calibration of the early Mesozoic time-scale and the long-term behaviour of the planets. *Philosophical Transactions of the Royal Society of London*, **A357**, 1761–1786.
Olsen, P. E., and Schlische, R. W., 1990. Transtensional arm of the early Mesozoic Fundy rift basin: penecontemporaneous faulting and sedimentation. *Geology*, **18**, 695–698.
St Onge, M. R., and King, J. E., 1987. Evolution of regional metamorphism during back-arc stretching and subsequent crustal shortening in the 1.9. Ga Wopmay Orogen, Canada. *Philosophical Transactions of the Royal Society of London*, **A321**, 199–218.
Quennell, A. M. (ed.), 1985. *Continental Rifts*. New York: Van Nostrand Reinhold. Benchmark Papers in Gelogy Series.
Quennell, A. M. (ed.), 1982. *Rift Valleys Afro-Arabian*. Benchmark papers in geology 60. Stroudsburg: Hutchison Ross.
Richardson-Bunbury, J. M., 1996. The Kula volcanic field, western Turkey: the development of a Holocene alkali basalt province and the adjacent normal-faulting graben. *Geological Magazine*, **133**, 275–283.
Ring, U., and Wernicke, B. (eds.), 2009. *Extending a continent: architecture, rheology and heat budget*. Geological Society special publication 321, 272 pp.
Rotstein, Y., Edel, J.-B., Gabriel, G., Boulanger, D., Schaming, M., and Munschy, M., 2006. Insight into the structure of the Upper Rhine Graben and its basement from a new compilation of Bouguer Gravity. *Tectonophysics*, **425**, 55–70.
Sawkins, F. J., 1990. *Metal Deposits in Relation to Plate Tectonics*. Berlin: Springer.
Schlüter, T., 1997. *Geology of East Africa*, with contributions by Craig Hampton. Beiträge zur Regionalen Geologe der Erde 27. Berlin: Gebrüder Borntraeger.
Schumacher, M. E., 2002. Upper Rhine Graben: role of preexisting structures during rift evolution. *Tectonics*, **21**, 1006–1022, doi:10.1029/2001TC900022.
Şengör, A. M. C., 1995. Sedimentation and tectonics of fossil rifts. In Busby, C. J., and Ingersoll, R. V. (eds.), *Tectonics of Sedimentary Basins*. Oxford: Blackwell, pp. 53–117.
Şengör, A. M. C., 1987. Cross-faults and differential stretching of hangingwalls in regions of low-angle normal faulting: examples form western Turkey. In Coward, M. P., Dewey, J. F., and Hancock, P. L. (eds.), *Continental Extensional Tectonics*. Geological Society special publication 321 (Albert M. Quennell volume), pp. 575–589.
Şengör, A. M. C., 2001. Elevation as indicator of mantle plume activity. In Ernst, R., and Buchan, K. (eds.), Geological Society of America special paper 352, pp. 183–225.

Şengör, A. M. C., and Natal'in, B. A., 2001. Rifts of the world. In Ernst, R., and Buchan, K. (eds.), Geological Society of America special paper 352, pp. 389–482.
Şengör, A. M. C., and Burke, K., 1978. Relative timing of rifting and volcanism on Earth and its tectonic implications. *Geophysical Research Letters*, **5**, 419–421.
Smith, A. G., 1976. Orogeny: a review. *Tectonophysics*, **33**, 215–285.
Snow, J. K., and Wernicke, B., 2000. Cenozoic tectonism in the Central Basin and Range: magnitude, rate and distribution of upper crustal strain. *American Journal of Science*, **300**, 659–719.
Spizaharsky, T. N., and Borovikov, L. I., 1966. Tectonic map of the Soviet Union on a scale of 1: 2 500 00. In *Scientific Communications Read to the Commission for the Geological Map of the World*. 22nd International Geological Congress, Delhi, pp. 111–120
Surkov, V. S., Smirnov, L. V., and Zhero, O. G., 1994. Early Mesozoic rifting and evolution of the West Siberian Basin. In Roure, F., Elluz, N., Shein, V. S., and Skvortsov, I. (eds.), *Geodynamic Evolution of Sedimentary Basins*. International Symposium, Moscow, Paris, Éditions Technip, pp. 135–143.
ten Brink, U. S., and Taylor, M. H., 2002. Crustal structure of central Lake Baikal: insights into intracontinental rifting. *Journal of Geophysical Research*, **107**, ETG-2-1–ETG-2-15, doi:10.1029/2001JB000300.
Ulmishek, G. F., Bogino, V. A., Keller, M. B., and Poznyakevich, Z. L., 1994. Structure, stratigraphy, and petroleum geology of the Pripyat and Dnieper-Donets Basins, Byelarus and Ukraine. In Landon, S. M. (ed.), *Interior Rift Basins*. American Association of Petroleum Geologists memoir 59, pp. 125–156.
Ulrych, J., Cajz, V., and Adamovič, J. (eds.), 1999. Magmatism and rift basin evolution. *GeoLines*, **9**, 1–135.
Verdel, C., Wernicke, B. P., Ramezani, J., Hassanzadeh, Renne, P. R., and Spell, T. L., 2007. Geology and thermochronology of tertiary Cordilleran-style metamorphic core complexes in the Saghand region of central Iran. *Geological Society of America Bulletin*, **119**, 961–977.
Wernicke, B., 2009. The detachment era (1977–1982) and its role in revolutionizing continental tectonics. In Ring, U., and Wernicke, B. (eds.), Geological Society special publication 321, pp. 1–8.
Wernicke, B. (ed.), 1990. *Basin and range extensional tectonics near the latitude of Las Vegas, Nevada*. Geological Society of America memoir 176, xii + 511 pp. + numerous separate plates.
Wernicke, B., and Burchfiel, B. C., 1982. Modes of extensional tectonics. *Journal of Structural Geology*, **4**, 105–111.
Wiart, P., and Oppenheimer, C., 2004. Large magnitude silicic volcanism in north Afar: the Nabro volcanic range and Ma'alalta volcano. *Bulletin of Volcanology*, **67**, 99–115.
Wickham, S., and Oxburgh, E. R., 1985. Continental rifts as a setting for regional metamorphism. *Nature*, **318**, 330–333.
Wilson, M., Neumann, E.-R., Davies, G. R., Timmermann, M. J., Heremans, M., and Larsen, B. T. (eds.), 2004. *Permo-Carboniferous magmatism and rifting in Europe*. Geological Society special publication 223, 498 pp.
Woodcock, N. H., 1986. The role of strike-slip fault systems at plate boundaries. *Philosophical Transactions of the Royal Society of London*, **A317**, 13–29.
Yirgou, G., Ebinger, C. J., and Maguire, P. K. H. (eds.), 2006. *The Afar volcanic province within the East African rift system*. Geological Society special publication 259, 331 pp.
Ziegler, P. A. (ed.), 1992. *Geodynamics of Rifting*. Amsterdam: Elsevier.
Ziegler, P. A., 1994a. Hydrocarbon habitat in rift basins. In Roure, F., Elluz, N., Shein, V. S., and Skvortsov, I. (eds.), *Geodynamic Evolution of Sedimentary Basins*. International Symposium, Moscow, Paris, Éditions Technip, pp. 85–94.
Ziegler, P. A., 1994b. Geodynamic processes governing development of rift basins. In Roure, F., Elluz, N., Shein, V. S., and Skvortsov, I. (eds.), *Geodynamic Evolution of Sedimentary Basins*. International Symposium, Moscow, Paris, Éditions Technip, pp. 19–67.

Cross-references

Earth, Density Distribution
Earthquake, Focal Mechanism
Earthquake, Location Techniques
Earthquakes and Crustal Deformation
Geodynamics
Geoid
Geothermal Heat Pumps
Isostasy
Isostasy, Thermal
Lithosphere, Continental
Lithosphere, Mechanical Properties
Mantle Convection
Mantle Plumes
Plate Motions in Time: Inferences on Driving and Resisting Forces
Plates and Paleoreconstructions
Radiogenic Heat Production of Rocks
Seafloor Spreading
Sedimentary Basins
Seismicity, Intraplate

CORE DYNAMO

Ulrich R. Christensen
Max-Planck Institut für Sonnensystemforschung, Katlenburg-Lindau, Germany

Synonyms

Geodynamo

Definition

Dynamo. Process for generating electrical current and magnetic field by electromagnetic induction in a moving conducting medium.
Geodynamo. Dynamo process in the fluid outer core that generates Earth's main magnetic field.

Introduction

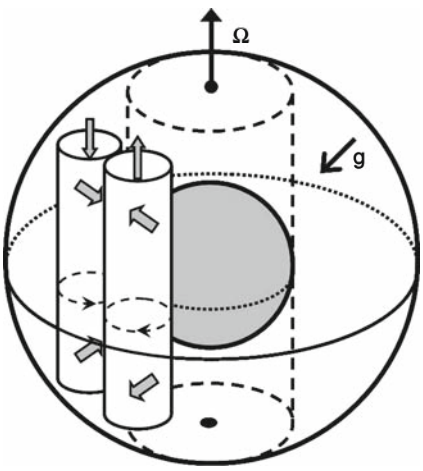
It has been firmly established that the geomagnetic field must be generated by a magnetohydrodynamic dynamo operating in the Earth's outer core. Thermal and compositional buoyancy forces drive a convective circulation of the liquid iron alloy. Assuming that a magnetic field is present, electrical currents are induced in the moving fluid. When the magnetic field associated with these currents has the strength and geometry that is suitable for the induction process and no external source for the field is required, this is called a self-sustained dynamo. The first approximation of the Earth's core is a sphere with uniform electrical conductivity, in contrast to technical dynamos, where the currents are guided by a highly inhomogeneous distribution of electrical conductors. In the former case, we speak of a homogeneous dynamo. From a theoretical point of view, homogeneous dynamos are more difficult to understand.

Simple motions, such as differential rotation (as in an electrical generator), are unable to drive a homogeneous dynamo, and flow patterns of a certain complexity are required. Until late into the twentieth century, dynamo theory has been concerned mainly with the conceptual understanding of how, in principle, a magnetic field can be generated in such an environment. Starting with models by Glatzmaier and Roberts (1995), Kageyama and Sato (1995) and Kuang and Bloxham (1997), realistic self-consistent numerical simulations of the geodynamo became available and have been successful in reproducing many of the observed properties of the geomagnetic field. The more fundamental aspects of the geodynamo are discussed elsewhere (see *Geomagnetic Field, Theory*). Here, the progress in understanding the geodynamo based on numerical modeling and comparing its results with specific properties of the geomagnetic field is addressed.

Dynamo model concept and equations

Model setup

In contrast to earlier kinematic dynamo models, where the flow is prescribed, modern geodynamo models aim at a fully self-consistent treatment of fluid flow and magnetic field generation in the Earth's outer core. There are some basic requirements for a realistic model: (1) The mechanism for driving flow by thermal or compositional buoyancy must be part of the model. (2) Because an axisymmetric field cannot be generated by a dynamo (Cowling's theorem), the model must be fully three-dimensional. (3) The model must be rotating because Coriolis forces are important to generate a flow pattern that is conducive for the dynamo process. The basic setup of geodynamo models is that of a rotating spherical shell of outer radius r_o and inner radius r_i, filled with a fluid of uniform conductivity, which represents Earth's outer core (Figure 1). For this system, the coupled equations for convection-driven flow and electromagnetic induction are solved. A detailed account on fundamental aspects of convection in rotating spheres is given in Jones (2007), and modeling aspects and the commonly employed numerical schemes are discussed in Christensen and Wicht (2007).



Core Dynamo, Figure 1 Columnar convection in a rotating spherical shell. The inner core tangent cylinder is shown by broken lines. Under Earth's core conditions, the columns would be thinner and more numerous. (From Christensen, 2010b, Copyright: Cambridge University Press.)

Dynamo equations

The relevant magnetohydrodynamic equations are usually written in non-dimensional form. A possible scheme for scaling the equations is to use the shell thickness $D = r_o - r_i$ as length scale, the viscous diffusion time D^2/ν as timescale (ν is kinematic viscosity), $(\rho\Omega/\sigma)^{1/2}$ for the scale of the magnetic field $\mathbf{B}$ and the imposed temperature contrast ΔT between inner and outer boundary for temperature T (ρ is density, σ electrical conductivity and Ω rotation rate). The Navier–Stokes equation for the velocity $\mathbf{u}$, augmented by rotational and electromagnetic forces, is

$$E\left(\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u}\cdot\nabla\mathbf{u}\right) + 2\hat{\mathbf{z}}\times\mathbf{u} + \nabla\Pi = E\nabla^2\mathbf{u} + \frac{Ra\,E}{Pr}\frac{\mathbf{r}}{r_o}T + \frac{1}{Pm}(\nabla\times\mathbf{B})\times\mathbf{B}. \quad (1)$$

The terms on the left-hand side describe in order the inertial force, the Coriolis force (with $\hat{\mathbf{z}}$ the unit vector parallel to the rotation axis) and the gradient of the non-hydrostatic pressure Π. The terms on the right-hand side stand for viscous friction, thermal buoyancy forces and the Lorentz force.

The magnetic induction equation, obtained from Maxwell's equations and Ohm's law for a moving incompressible conductor, is

$$\frac{\partial \mathbf{B}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{B} = (\mathbf{B}\cdot\nabla)\mathbf{u} + \frac{1}{Pm}\nabla^2\mathbf{B}, \quad (2)$$

where the second term on the LHS describes magnetic field advection and the terms on the RHS magnetic field generation and diffusion, respectively. Magnetic diffusion is a consequence of the ohmic resistance that damps the electrical currents associated with the magnetic field.

The advection–diffusion equation for temperature is

$$\frac{\partial T}{\partial t} + \mathbf{u}\cdot\nabla T = \frac{1}{Pr}\nabla^2 T + \varepsilon, \quad (3)$$

with a heat source term ε on the RHS. For compositional convection, an equivalent equation holds where the concentration of light components replaces temperature. The set of equations is completed by the condition of incompressibility, which seems to be justified for Earth's core where density differences are moderate, and the condition that $\mathbf{B}$ is solenoidal:

$$\nabla\cdot\mathbf{u} = 0, \quad \nabla\cdot\mathbf{B} = 0. \quad (4)$$

The equations are complemented by boundary conditions for the velocity, usually $\mathbf{u} = 0$, fixed temperature or fixed heat flux at r_o and r_i, and a continuity condition for the magnetic field at r_o that links the field in the dynamo with an external potential field that decays with radius. Simple models assume the inner core to be insulating, but in many cases it is taken as an electrical conductor with the same conductivity as in the outer core. It has been suggested that the finite conductivity of the inner core plays an important role for preventing frequent dipole reversals of the geodynamo, because a dipole reversal can only be completed successfully if the new polarity persists for the timescale of diffusive dipole decay in the inner core, $r_i^2/(\pi^2\lambda)$, which is several thousand years. However, this is contested by dynamo models that show little differences between simulations with a conducting or an insulating inner core.

Parameters

The four non-dimensional control parameters in these equations are defined in Table 1. In dynamo models, the values of most of these parameters fall short of their extreme values in the Earth's core. For practical reasons, it is not possible to resolve the very small spatial structures that occur at realistic parameter values. The *Rayleigh number Ra* describes the ratio between buoyancy driving convection and retarding effects. In the core, it is much larger than the critical value for the onset of convection, Ra_c, while models are more moderately supercritical (Table 1). The discrepancy is by ten orders of magnitude for the *Ekman number E*, the ratio of viscous forces to Coriolis forces. The very small core value of E indicates that viscous forces are negligible, except in very thin Ekman layers at the boundaries to the solid mantle and inner core. The *magnetic Prandtl number Pm*, the ratio of viscosity to magnetic diffusivity, is very small for liquid metals, but must be set to a value of order one to obtain a self-sustained dynamo in present models. Only the (hydrodynamic) *Prandtl number Pr*, the ratio of viscosity to thermal diffusivity, is of order one in the core and in dynamo models. In terms of physical parameters, the viscosity and the thermal diffusivity are far too large in the models and, in most cases, the rotation rate is too small.

Several dimensionless diagnostic numbers can be formed with the characteristic flow velocity U and magnetic field strength B in the dynamo (Table 1). The velocity in Earth's core can be estimated from the geomagnetic secular variation under the assumption that the magnetic flux at the core surface is approximately frozen into the moving fluid (Alfvén's theorem), which is the case when the diffusion term in Equation 2 can be neglected. In this case, the variation of the radial magnetic field component B_r at the core-mantle boundary can be written

$$\partial B_r/\partial t + \nabla_h \cdot (\mathbf{u} B_r) = 0. \tag{5}$$

where the suffix h indicates the horizontal part of the divergence operator. Additional assumptions are needed to infer the velocity at the top of the core (below the Ekman layer) from the observed B_r and its time derivative, but the flow pattern obtained from these inversions are broadly similar and the characteristic velocity is approximately 0.5 mm/s (Holme, 2007). The characteristic field strength inside Earth's core is probably in the range 1–4 mT, as will be discussed further below.

The most important diagnostic number is the *magnetic Reynolds number Rm*, which describes the ratio of magnetic advection and induction to magnetic diffusion. In order for induction effects in a dynamo to overcome the diffusive dissipation of magnetic field, it is expected that *Rm* must be larger than one. Practically, it turns out that self-sustained magnetic field generation occurs in geodynamo models when *Rm* exceeds approximately 40. Using $U \approx 0.5$ mm/s, the magnetic Reynolds number is of order 1,000, safely above the critical limit. Also, $Rm \gg 1$ is a condition for Equation 5 to apply. Still, the value of *Rm* in the Earth's core is moderate and can be handled in direct numerical simulations. This contrasts with the much larger values of *Rm* in other cosmic dynamos, which requires that magnetic induction and diffusion

Core Dynamo, Table 1 Dynamo parameters. α thermal expansivity, g_o gravity at core surface, ΔT superadiabatic temperature contrast across core, κ thermal diffusivity, ν kinematic viscosity, Ω rotation rate, D outer core thickness, $\lambda = 1/(\mu_o\sigma)$ magnetic diffusivity, U characteristic flow velocity, B characteristic magnetic field strength, μ_o magnetic permeability, ρ density

	Control parameters			
	Rayleigh number	Ekman number	Magn. Prandtl number	Prandtl number
Definition	$Ra = \alpha g_o \Delta T D^3/(\kappa\nu)$	$E = \nu/(\Omega D^2)$	$Pm = \nu/\lambda$	$Pr = \nu/\kappa$
Core	$10^4 Ra_c$	10^{-15}–10^{-14}	10^{-6}–10^{-5}	0.1–1
Models	(1–100) Ra_c	10^{-6}–10^{-2}	0.1–10	1
	Diagnostic numbers			
	Magn. Reynolds no.	Reynolds no.	Rossby no.	Elsasser no.
Definition	$Rm = UD/\lambda$	$Re = UD/\nu$	$Ro = U/(\Omega D)$	$\Lambda = B^2/(2\mu_o\lambda\rho\Omega)$
Core	10^3	10^9	10^{-7}	0.1–10
Models	40–2,000	<2,000	10^{-2}–10^{-4}	1–100

effects at small unresolved scales are parameterized by applying concepts of the mean-field dynamo theory. The ability to solve for magnetic induction and diffusion in Earth's core directly without parameterizations is probably the main reason for the success of geodynamo models.

The (hydrodynamic) *Reynolds number Re* is much smaller in the models than it is in the core where it is of the order 10^9, indicating a highly turbulent regime. Consequently, dynamo models miss the small eddies that are expected to exist in the core. Viscosity kicks in only at length scales of some meters to suppress even smaller structures in the flow. However, the small eddies are unimportant for the magnetic induction process, because the magnetic diffusivity, which is $Pm^{-1} = 10^6$ times larger than the kinematic viscosity, homogenizes the magnetic field at scales of less than several tens of kilometers.

The *Elsasser number* Λ describes the ratio of Lorentz forces to Coriolis forces. In the Earth's core, both forces are assumed to be of leading order. A balance in Equation 1 that is dominated by these two forces (plus the pressure gradient force) is called a magnetostrophic balance. The magnetostrophic state is often associated with an Elsasser number of order one. Many present dynamo models match this value of Λ.

The *Rossby number Ro* describes the ratio of inertial forces to the Coriolis force and is very small in the core. In models, it is less than one, although larger than the core value. Instead of forming the Rossby number with the global length scale D, a more appropriate measure for the ratio of the two forces is obtained by using the characteristic length scale ℓ of the flow (local Rossby number Ro_ℓ). The value of Ro_ℓ is difficult to estimate for the core, but model values and the Earth value may be in better agreement than they are for *Ro*.

Energetics of the geodynamo

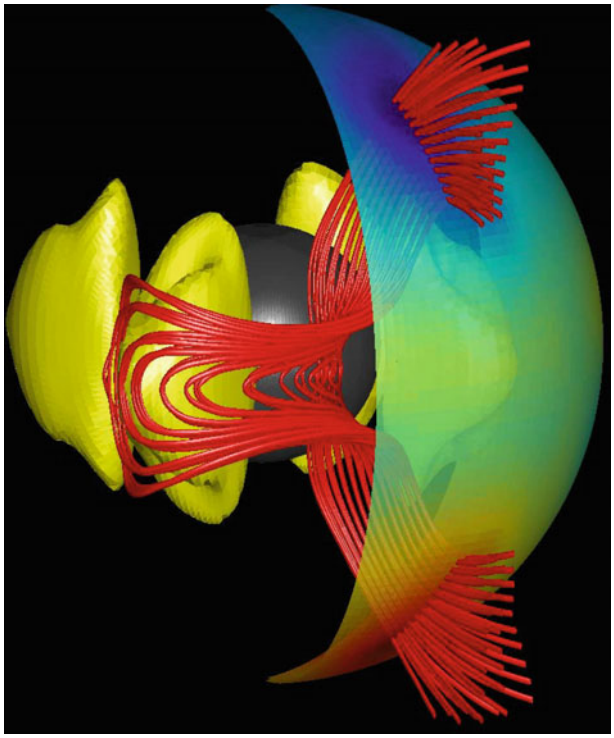
Although the decay of radioactive isotopes in the core may contribute to some degree to the core heat flux, most of the buoyancy that drives convection in the outer core is associated with secular cooling. The sluggish convection in the mantle controls the heat flux through the core-mantle boundary in terms of its magnitude and its spatial distribution, which is expected to be rather heterogeneous. Recent estimates put the heat flow at the core-mantle boundary into the range of 5–15 TW (Lay et al., 2008). For thermal convection to occur, the temperature gradient must reach or exceed the adiabatic gradient. Because iron is a comparatively good conductor, 3–8 TW can be transported by thermal conduction along an adiabatic gradient. This fraction of the heat flux is lost for driving the dynamo. For values of the heat flux near the low end of estimates and of the potential conductive flux near the upper end, a stable thermal stratification at the top of the core could exist. A consequence of core cooling is the slow growth of the solid inner core by freezing iron on its outer boundary. The associated latent heat release balances roughly one half of the heat loss from the core and renders the deep fluid core thermally unstable even if the top layer is not. The relative depletion of light alloying elements in the solid core is balanced by an enrichment in the residual fluid above the growing inner core, which drives compositional convection. Estimates for the work done by thermal and compositional buoyancy forces are of the order $W \approx 1$ TW (Nimmo, 2007). Compositional buoyancy contributes roughly two thirds to it. This energy flux is available to generate magnetic field and balance the associated ohmic dissipation. Viscous dissipation is usually assumed to play a minor role in the core.

Flow pattern and dynamo mechanism

The pattern of flow is strongly controlled by the Coriolis force. It is different for the regions inside and outside the *inner core tangent cylinder*. This is an imaginary cylinder aligned with the rotation axis and touching the inner core at the equator, which is shown by broken lines in Figure 1.

Flow outside tangent cylinder

Outside the tangent cylinder, the flow is organized in convection columns aligned with the vector of rotation, with the flow encircling the column axis (Figure 1). Superimposed is a circulation along the column axis that is directed towards the equatorial plane in columns with a cyclonic sense of rotation and away from the equator in anticyclonic columns. This results in helical flow, in which particles move on corkscrew-like trajectories. The helicity, $H = \mathbf{u} \cdot (\nabla \times \mathbf{u})$, is coherently negative in the northern hemisphere and positive in the southern hemisphere. Such helical flow is the basis for the α-effect, which can convert an axisymmetric toroidal field into an axisymmetric poloidal field, or vice versa, through intermediate steps involving non-axisymmetric small-scale fields (Moffatt, 1978). Because purely axisymmetric fields cannot be maintained by a dynamo, this mechanism plays an important role in the understanding of the dynamo process. The α-effect has been introduced in the context of mean-field dynamo theory, where it describes the induction effects of small unresolved flow scales. In geodynamo models and probably to some extent in the Earth's core, it is due to fairly large-scale convection columns (Busse, 1975). The models suggest that the axisymmetric toroidal field in the Earth's core (which is invisible from the outside) consists of a torus of westward-directed field north of the equatorial plane and eastward field south of the equatorial plane. The helical columnar flow converts this to a dipolar poloidal field whose field lines can leave the core (Figure 2). In some dynamo models, the toroidal magnetic field is regenerated by the same 'macroscopic α-effect from the poloidal magnetic field (α^2 dynamo, Olson et al., 1999). In other geodynamo models, the shearing of poloidal field lines by differential rotation, the Ω-effect, has been found to play an important role for toroidal field generation ($\alpha\Omega$-dynamo), but essentially all models confirm that the helical columnar convection produces the dipolar poloidal field component.



Core Dynamo, Figure 2 Bundle of magnetic field lines in a simple geodynamo model ($E = 10^{-3}$, Ra/Ra_c=1.8, Pm = 5, Pr = 1). The yellow blobs are anticyclonic vortices. Strong inward magnetic flux is shown in blue on the outer core boundary and outward flux in red. The NS-running part of the field lines represents poloidal magnetic field, and the EW-directed segments form the axisymmetric toroidal magnetic field, which consists of two bundles of opposite directions in the two hemispheres. The same field structure repeats periodically in longitude. (Courtesy of Julien Aubert.)

Although at low Ekman number and high Rayleigh number the simple columns give way to a more complex and turbulent pattern, the vortices are still elongated in the direction of the rotation axis and retain a coherent helicity in each hemisphere.

Strong concentrations of magnetic flux in the present geomagnetic field centered below North America and Sibera, which have counterparts at the same longitudes in the southern hemisphere, are interpreted as evidence for the presence of convection columns in the fluid core (Gubbins and Bloxham, 1987) (Figure 3a). The convergence at the core-mantle boundary associated with the equatorward flow in the axis of cyclonic columns (Figure 1) concentrates the poloidal magnetic field lines into these flux lobes, assuming that the frozen-flux assumption is approximately valid. In simple models, such as that in Figure 2, the fluxes lobes are associated with a single columnar vortex. Dynamo models at lower Ekman number and strongly supercritical Rayleigh number show similar high-latitude flux lobes in their magnetic fields when they are low-pass filtered to the resolution of the geomagnetic core field, roughly up to spherical harmonic degree 13 (e.g., Figure 3b). In this case, and probably in the Earth's core, the large-scale magnetic flux concentrations are due to the superposition of strong but complex and smaller-scale magnetic field structures, each associated with a separate convection column.

Flow inside tangent cylinder

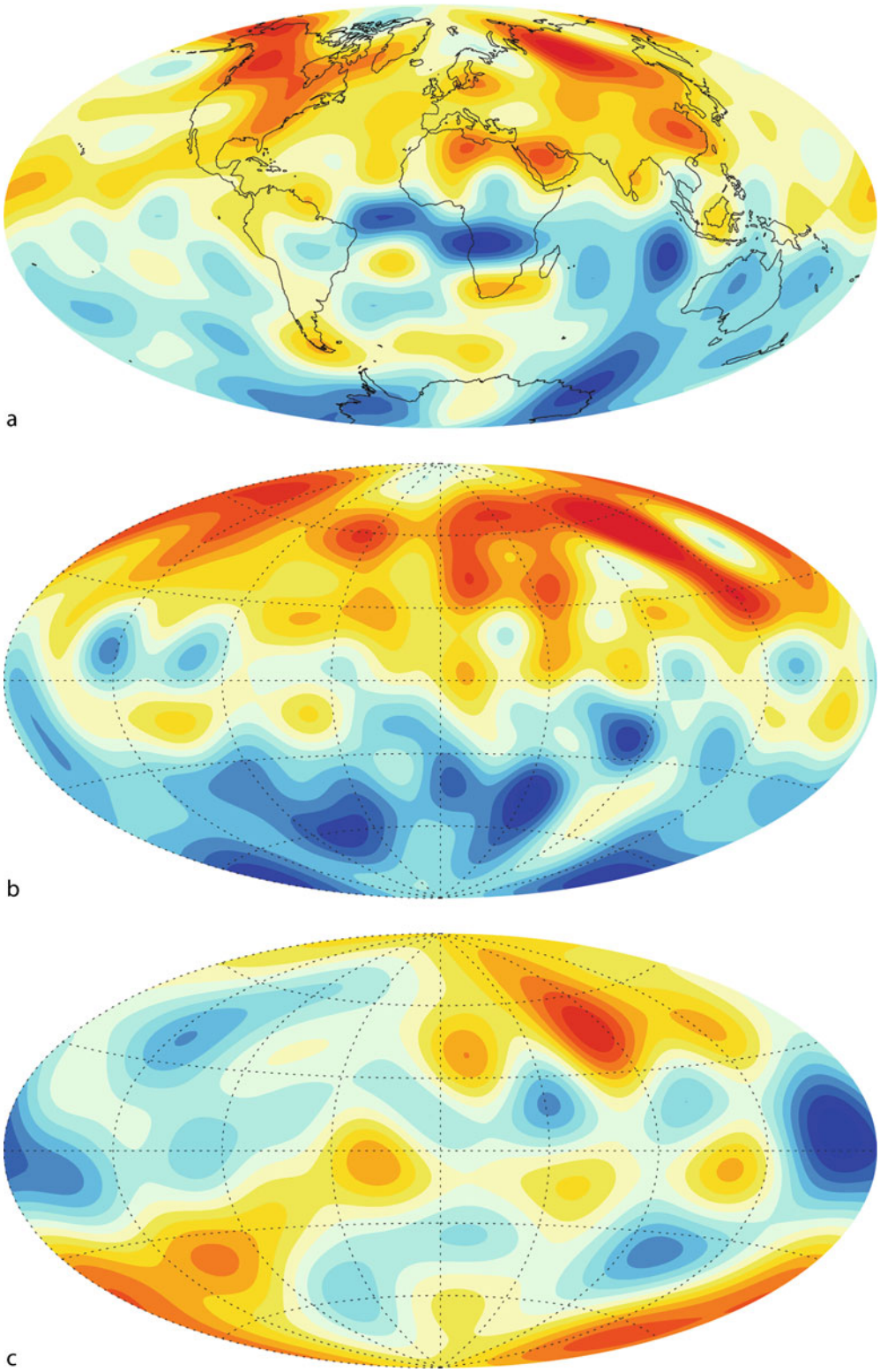
In the polar regions inside the tangent cylinder, many dynamo models show an upwelling plume at or near the rotation axis. Because of the action of Coriolis forces, the plume is associated with an eastward vortex motion near the inner core and by a westward vortex motion near the core-mantle boundary, which is called a thermal wind circulation. A westward vortex motion in the polar cap region at the top of the core has also been inferred from the observed geomagnetic secular variation (Hulot et al., 2002). A polar plume is associated with flow that diverges away from the rotation axis below the core surface, which disperses poloidal magnetic field lines. This is seen in many geodynamo models (e.g., Figure 3b) and can explain the weak magnetic flux in the north polar region of the core (Figure 3a). It has been suggested that the solid inner core couples electromagnetically to the eastward vortex motion of the overlying fluid, which leads to a superrotation of the inner core with respect to the mantle and crust. While perfect coupling to the outer core flow would lead to superrotation rates in excess of a degree per year (Glatzmaier and Roberts, 1996), gravitational coupling between the inner core and the mantle can slow down the rate. Seismological evidence indicates a superrotation at a fraction of a degree per year.

Field properties

Field strength inside the core

How strong is the magnetic field inside Earth's core? Observations allow to constrain the strength of the poloidal field at the core-mantle boundary, which is presently 0.26 mT for the dipole field alone and 0.39 mT in spherical harmonic degrees from 1 to 13. It is not known how much the small unresolved length scale add to the field at the top of the core, but possibly they double this value. Inside the Earth's core, the poloidal field is expected to be stronger than at the outer boundary, and the toroidal component also contributes. Suggestions that the toroidal field, if generated by an intense Ω-effect, could be much stronger than the poloidal field are not supported by geodynamo simulations. In most of them, the magnetic energy is equipartitioned between the two field components. The dynamo models suggest that the characteristic field strength inside the dynamo is around 5–15 times the dipole field strength on the outer boundary, which puts it into the range of 1–4 mT.

The field strength inside Earth's core can be probed directly by the observation of torsional waves and oscillations. To the extent that viscous and inertial forces are negligible, the azimuthal Lorentz force integrated over the



Core Dynamo, Figure 3 Radial magnetic field at the core-mantle boundary, *red* (*full lines*) inward and *blue* (*broken lines*) outward. (**a**) Geomagnetic field in 2010. (**b**) Dynamo model with dominant dipole contribution ($E = 10^{-5}$, $Pm = 0.8$, $Pr = 1$, $Ra/Ra_c = 114$). (**c**) Multipolar dynamo model ($E = 10^{-5}$, $Pm = 0.5$, $Pr = 1$, $Ra/Ra_c = 161$). The model magnetic fields are low-pass filtered to the resolution of the geomagnetic field.

surface of cylinders that are aligned with the rotation axis (geostrophic cylinders) must vanish in the so-called Taylor state. Perturbations of the Taylor state will results in torsional oscillations or torsional waves, consisting of an oscillatory differential rotation between the geostrophic cylinders. Magnetic field lines that permeate adjacent cylinders act like elastic strings and provide a restoring force when the cylinders are displaced relative to each other. The frequency of torsional eigenmodes and the wavespeed of traveling waves is directly related to the *rms* strength of the field component B_s in the cylindrically radial direction. The identification of torsional modes and waves in the secular variation data set is somewhat tentative, but is supported by a correlation between the inferred large-scale motion in the core and changes in the rotation rate of the solid part of the Earth (length-of-day changes) caused by angular momentum exchange between core and mantle (e.g., Holme, 2007). Torsional waves have led to estimates for the characteristic B_s-value ranging from 0.2 to 2 mT. Values near the upper end of this range for B_s agree with the dynamo model estimates for the total field assuming that the core field is statistically isotropic.

Scaling laws

An important question is what controls the magnetic field strength of the geodynamo and of planetary dynamos in general. It has been suggested that the magnitude of the internal core field is determined by a magnetostrophic force balance, meaning that the magnetic field inside Earth's core is driven towards a value that makes the Elsasser number (Table 1) to be of order one, i.e.,

$$B \propto (2\rho\Omega/\sigma)^{1/2}. \tag{6}$$

This results in $B \approx 2$ mT in good agreement with the above estimates. A different line of argument is based on the power available to balance ohmic dissipation. With some assumptions on how the length scale of the magnetic field in the core depends on the vigor of convection, this leads to

$$B \propto \mu_o^{1/2}\, \rho^{1/6}\, P^{1/3}, \tag{7}$$

where $P = W/(4\pi r_o^2)$ stands for the power per unit area of the core surface. With the current estimates for W Equation 7 results in a similar value of the field strength as the Elsasser number rule. However, the two scaling rules are radically different; in Equation 6, the field strength is independent of the energy flux, and in Equation 7, it does not depend on rotation rate or electrical conductivity. Sets of dynamo simulations that cover a substantial range in these properties, or their counterparts in terms of the non-dimensional control parameters, support the power-based scaling law (Equation 7). Furthermore, they show no significant dependence of the field strength on the viscosity. This may indicate that although current dynamo simulations are far too viscous (too high Ekman number), the viscosity is sufficiently low to not play a dominant role. However, testing this by further lowering the viscosity is desirable. The validity of the power-based scaling rule is also supported by the finding that it gives the right order of magnitude of the field strength for various convection-driven dynamos in rapidly rotating cosmic objects including Earth, Jupiter, and low-mass stars. A detailed discussion of dynamo scaling laws is given in Christensen (2010a).

Magnetic field geometry

While the rotation rate seems to have little influence on the field strength inside dynamo models, it probably controls the magnetic field morphology. In some geodynamo models, the magnetic field on the outer boundary resembles the Earth's field at the core-mantle boundary (Figure 3a and b) in terms of a dominant axial dipole and also of structures associated with multipole contributions. In other dynamo models, the dipole contribution is weak. In some cases, the field is dominated by the axial quadrupole or is restricted to one hemisphere, but in most cases it is irregular with contributions from a broad spectrum of multipole components without a pronounced dominance of a particular mode (Figure 3c). The selection of the field geometry was found to depend on the (local) Rossby number. At low values (rapid rotation and moderate flow velocities), dynamos are dipolar, whereas for high values (slower rotation and fast flow), they tend to be multipolar. In addition, boundary conditions and the mode of driving convection, i.e., either by secular cooling of the fluid core or by the latent heat and light element flux from the growing solid inner core, can play a role. The driving mode has changed over geological time. Before the Earth's inner core nucleated, which is estimated to have occurred about 0.5–1.5 Gyr ago (Nimmo, 2007), the dynamo has been driven by secular cooling alone, whereas thereafter the buoyancy source is mainly located at the inner core boundary. However, dynamo simulations for the different scenarios (Aubert et al., 2009) suggest that neither the geometry of the magnetic field nor its strength has changed very strongly. This is in accord with the lack of paleomagnetic evidence for drastic changes in the geomagnetic field since the Archean.

Secular variation

Timescales

The characteristic timescale of secular variation can be defined by the square root of the power of the magnetic field over the power in the time derivative of the field. This can be done separately for each harmonic degree n. The timescales decrease with increasing n. Aside from this scale dependence, the secular variation time has been found in dynamo models to vary with the inverse of the magnetic Reynolds number Rm, as would be expected if most of the secular variation is caused by the frozen-flux advection of magnetic field. Using the observed secular variation timescales to estimate the magnetic Reynolds number (or flow velocity) in the core gives essentially the same result, $Rm \approx 1,000$, as the more complex inversion of the secular variation for the core flow.

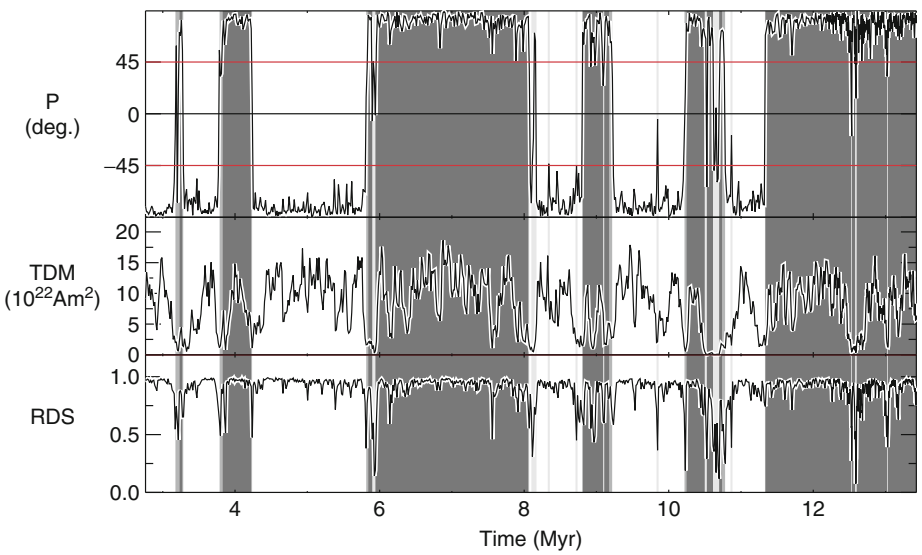
Advection versus diffusion of field

The westward drift of magnetic field structures is a long-standing issue. Some dynamo models show a persistent and global westward drift, whereas others do not. Over the past 400 years, magnetic field structures have systematically migrated westward in the Atlantic hemisphere, but not in the Pacific hemisphere, where the amplitude of secular variation is generally weaker. Dynamo models support the notion that such hemispherical dichotomy can be caused by coupling between the mantle and the core, in particular by the heterogeneity of the core heat flow caused by lower mantle thermal structure.

While much of the geomagnetic secular variation can be attributed to frozen-flux advection in a large-scale circulation, magnetic diffusion must play some role, in particular diffusion in the radial direction. It has been suggested that intense flux spots of both polarities, which are found in the geomagnetic field at low latitudes under Africa and the Atlantic (Figure 3a), have been formed by the expulsion of toroidal magnetic flux bundles through the core surface Bloxham (1986), somewhat analogous to the formation of sunspots. Because the toroidal field is confined to the core, the emergence of such spots necessarily involves diffusion, whereas their migration can be caused by frozen-flux advection. Some geodynamo simulations support such origin of equatorial flux spots. In these simulations, the columnar convection acts on a pair of toroidal flux bundles with opposite polarities and produces (anti) symmetric pairs of spots north and south of the equator. However, the geomagnetic low-latitude field pattern conforms at best marginally with this picture. An alternative explanation to the concept of field expulsion by convective columns and migration due to a large-scale westward flow is that, propagating hydromagnetic waves in the core create a periodic pattern of upwelling and downwelling that produces the belt of wave-train-like magnetic field structures in the low-latitude Atlantic hemisphere.

Magnetic field reversals

Several dynamo simulations show stochastic dipole reversals. Often the details resemble remarkably well what is known from the paleomagnetic record about geomagnetic reversals (e.g., Wicht et al., 2009). They are rather short events that interrupt long periods of stable polarity; both complete reversals and aborted reversals (excursions) occur; during a reversal the magnetic field is weak and not dipole dominated (Figure 4). Most model studies of reversal behavior have been performed with simple dynamo models at rather high values of the Ekman number, because long simulation times are needed to capture a fair number of these events. Reversing dipolar dynamos are typically found in a parameter regime between that of stable dipolar dynamos and of multipolar dynamos. Driving convection more strongly for otherwise identical conditions leads to very frequent and erratic dipole reversals that have little significance because the dipole is hardly ever dominating the magnetic field. Lowering the Prandtl number has the same effect, suggesting that strengthening the nonlinear inertial forces relative to the Coriolis force favors non-dipolar dynamo. Reducing the convective driving or increasing the Prandtl number results in strong dipolar fields that rarely or never reverse. In addition, boundary conditions



Core Dynamo, Figure 4 Time series of dipole tilt, true dipole moment (TRM) and relative dipole strength (RDS) from a numerical dynamo simulation. (Wicht, 2005, Copyright Blackwell.)

affect the reversal frequency. In particular, it has been shown that the pattern of heat flux at the core-mantle boundary, which is controlled by the thermal state of the lowermost mantle and varies on the long timescale of mantle convection, has a strong influence on the occurrence of reversals. This supports the assumption that the drastic secular change in the reversal rate over the past 200 Myr, which is observed in the paleomagnetic record, is caused by changes in the amplitude and pattern of core heat flow associated with a reorganization of mantle convection, such as formation and fading of superplumes. Although model studies have led to a number of suggestions concerning the cause for individual reversals, for example by the formation of strong new plumes in the core, so far no generally accepted concept for the reversal mechanism has emerged.

Summary

Helical convection columns outside the inner core tangent cylinder region play an essential role for our present understanding of the geodynamo. Theory and numerical simulations suggest that they are a fundamental element of the circulation in the outer core, and that they play a key role for the generation the Earth's dipole-dominated magnetic field. Certain details of the magnetic field structure at the core-mantle boundary are explained by such columns. Somewhat less certain, but supported by several lines of evidence, is the notion of rising plumes in the tangent cylinder near the rotation axis that are associated with weak magnetic field and westward vortex flow in the polar regions at the core-mantle boundary. However, much of our current understanding is based on numerical dynamo models that are far too viscous compared to the core. Although there is some evidence suggesting that this does not affect basic properties of the magnetic field, lowering the viscosity in dynamo models is important to test the robustness of our present results, even if models that use actual core values of the viscosity are out of reach. Modeling supports the notion that the magnetic field strength is controlled by the available power that drives the dynamo, not necessarily by a balance of Lorentz and Coriolis forces, as has been often assumed. Several dynamo models show dipole reversals that resemble, in many respects, geomagnetic reversals, but a full understanding of what causes reversals or controls the frequency of geomagnetic reversals is missing.

Bibliography

Aubert, J., Labrosse, S., and Poitou, C., 2009. Modelling the paleo-evolution of the geodynamo. *Geophysical Journal International*, **179**, 1414–1429.

Bloxham, J., 1986. The expulsion of magnetic flux from the Earth's core. *Philosophical Transactions of the Royal Society of London*, **87**, 669–678.

Busse, F. H., 1975. A model of the geodynamo. *Geophysical Journal of the Royal Astronomical Society*, **42**, 437–459.

Christensen, U. R., 2010a. Dynamo scaling laws: applications to the planets. *Space Science Reviews*, **152**, 565–590.

Christensen, U. R., 2010b. Planetary fields and dynamos. In Schrijver, C. J., and Siscoe, G. L. (eds.), *Heliophysics: Evolving Solar Activity and the Climates of Earth and Space*. Cambridge: Cambridge University Press.

Christensen, U. R., and Tilgner, A., 2004. Power requirement of the geodynamo from Ohmic losses in numerical and laboratory dynamos. *Nature*, **429**, 169–171.

Christensen, U. R., and Wicht, J., 2007. Numerical dynamo simulations. In Schubert, G. (ed.), *Treatise on Geophysics, Core Dynamics*. Amsterdam: Elsevier, Vol. 8, pp. 245–282.

Glatzmaier, G. A., and Roberts, P. H., 1995. A three-dimensional self-consistent computer simulation of a geomagnetic field reversal. *Nature*, **337**, 203–209.

Glatzmaier, G. A., and Roberts, P. H., 1996. Rotation and magnetism of Earth's inner core. *Science*, **274**, 1887–1891.

Gubbins, D., and Bloxham, J., 1987. Morphology of the geomagnetic field and implications for the geodynamo. *Nature*, **325**, 509–511.

Holme, R., 2007. Large-scale flow in the core. In Schubert, G. (ed.), *Treatise on Geophysics, Core Dynamics*. Amsterdam: Elsevier, Vol. 8, pp. 107–130.

Hulot, G., Eymin, C., Langlais, B., Mandea, M., and Olson, N., 2002. Small-scale structure of the geodynamo inferred From Oersted and Magsat Satellite data. *Nature*, **416**, 620–623.

Jones, C. A., 2007. Thermal and compositional convection in the outer core. In Schubert, G. (ed.), *Treatise on Geophysics, Core Dynamics*. Amsterdam: Elsevier, Vol. 8, pp. 131–185.

Kageyama, A., and Sato, T., 1995. Computer simulation of a magnetohydrodynamic dynamo II. *Physics of Plasmas*, **2**, 1421–1431.

Kuang, W., and Bloxham, J., 1997. An earth-like numerical dynamo model. *Nature*, **389**, 371–374.

Lay, T., Hernlund, J., and Buffett, B. A., 2008. Core mantle boundary heat flow. *Nature Geoscience*, **1**, 25–32.

Moffatt, H. K., 1978. *Magnetic Field Generation in Electrically Conducting Fluids*. Cambridge: Cambridge University Press.

Nimmo, F., 2007. Energetics of the core. In Schubert, G. (ed.), *Treatise on Geophysics, Core Dynamics*. Amsterdam: Elsevier, Vol. 8, pp. 31–65.

Olson, P., Christensen, U. R., and Glatzmaier, G. A., 1999. Numerical modeling of the geodynamo: mechanism of field generation and equilibration. *Journal of Geophysical Research*, **104**, 10383–10404.

Wicht, J., 2005. Palaeomagnetic interpretation of dynamo simulations. *Geophysical Journal International*, **162**, 371–380.

Wicht, J., Stellmach, S., and Harder, H., 2009. Numerical models of the geodynamo: from fundamental cartesian models to 3D simulations of field reversals. In Glassmeier, K. H., Soffel, H., and Negendank, J. F. W. (eds.), *Geomagnetic Field Variations*. Berlin: Springer, pp. 107–158.

Cross-references

Core-Mantle Coupling
Differential Rotation of the Earth's Inner Core
Earth's Structure, Core
Energy Budget of the Earth
Geomagnetic Excursions
Geomagnetic Field, Global Pattern
Geomagnetic Field, Polarity Reversals
Geomagnetic Field, Secular Variation
Geomagnetic Field, Theory
Paleomagnetic Field Intensity
Radioactivity in Earth's Core

CORE-MANTLE COUPLING

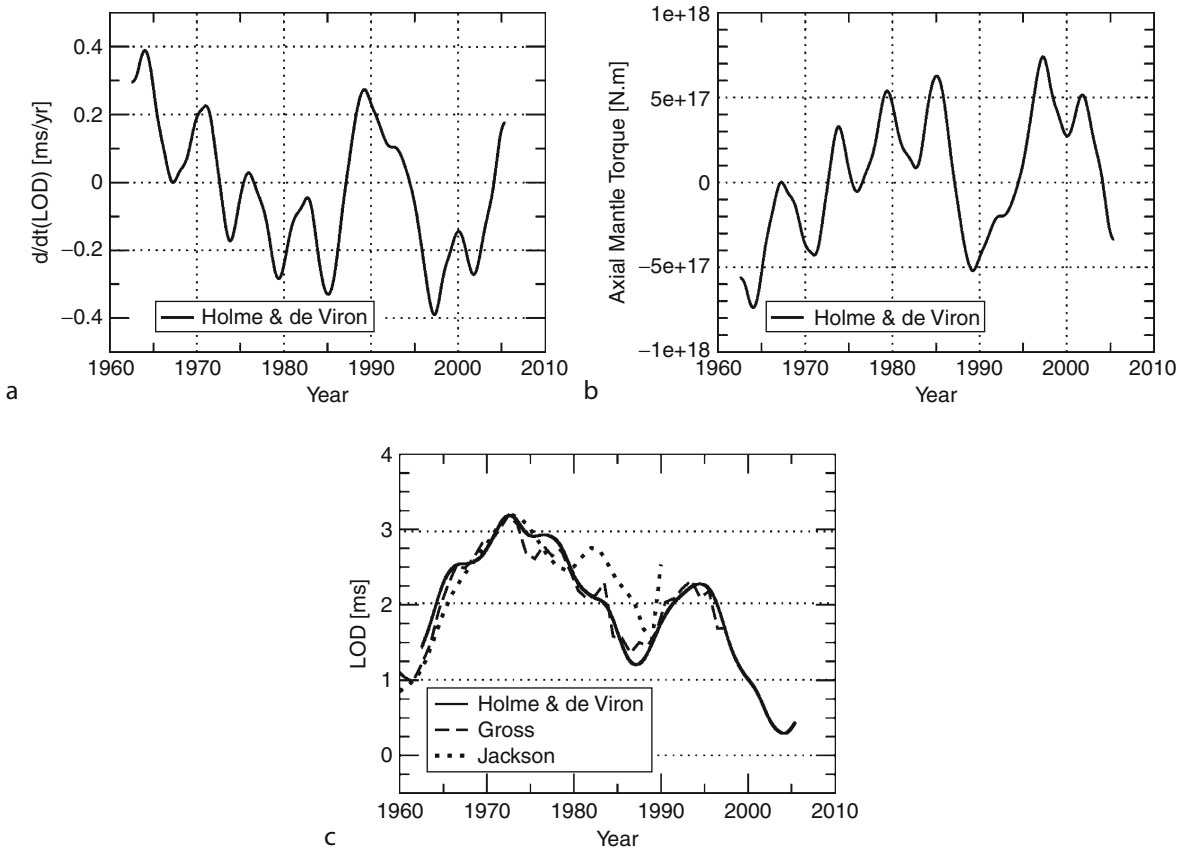
Paul H. Roberts[1], Jonathan M. Aurnou[2]
[1]Institute of Geophysics and Planetary Physics, University of California, Los Angeles, CA, USA
[2]Department of Earth and Space Sciences, University of California, Los Angeles, CA, USA

Definition, scope, and aims

The Earth is not a perfect timekeeper, and the spectrum of the variations in the mantle's angular velocity $\widehat{\mathbf{\Omega}}$ spans a wide range of frequencies. Of particular interest here are the comparatively large amplitude decadal and semi-decadal variations in which changes in length of day, P, of up to 2 ms occur. These would not be explained even if the global circulations of the atmosphere and oceans could be reversed. This is confirmed by a more detailed argument given in our recent review (Roberts and Aurnou, 2012), which will be referred to here as "RA12."

The origin of these length of day (LOD) variations must be sought in the Earth's core, and Figure 1 suggests that the task is not an easy one. Figure 1b shows dP/dt, derived by differentiating smoothed LOD data from the last half century, with atmospheric, oceanic, and tidal signals removed; semi-decadal time variations are clearly seen with a period τ_{LOD} of about 6 years (e.g., Abarca del Rio et al., 2000). Figure 1c shows the implied $\widehat{\Gamma}_z$ as a function of time t, where $\widehat{\Gamma}_z$ is the component parallel to the polar axis O_z of the torque $\widehat{\mathbf{\Gamma}}$ exerted by the core on the mantle. This shows that $\widehat{\Gamma}_z$ of nearly 10^{18} Nm is generated. We shall call this the "target torque" and seek its origin.

In addition, to the semi-decadal oscillations, larger variations seem to exist having longer periods, one of which is estimated to have a roughly 60 year period (e.g., Roberts et al., 2007). This time scale is reminiscent of the geomagnetic secular variation, and it is natural to seek



Core-Mantle Coupling, Figure 1 (**a**) LOD time series data from Holme and de Viron (2005) and Gross (2001) compared with LOD model from the "smooth" core flow inversion of Jackson (1997). The time series include the variation due to lunar tidal drag. (**b**) Temporal derivative, dP/dt, of the smoothed time series data of Holme and de Viron (2005), where P is the length of day. (**c**) The implied axial torque, $\widehat{\Gamma}_z = -(2\pi\widehat{C}/P^2)\mathrm{d}P/\mathrm{d}t$, on the mantle, where $\widehat{C}$ is the axial moment of inertia of the mantle. The LOD in panel (**a**) is arbitrary to within an additive constant, chosen here to obtain agreement at 1972.5 with Gross (2001).

a connection between them (e.g., Braginsky, 1970; Gillet et al., 2010). Figure 1c supports this quest. It shows the LOD data (Gross, 2001; Holme and de Viron, 2005) plotted against the estimated LOD variations inferred from the core flow models of Jackson (1997) that are based on inversion of geomagnetic secular variation data. The qualitative agreement implies that the variations in LOD are due to core-mantle angular momentum exchange associated with magnetohydrodynamic (MHD) processes in the core.

This review will focus on variations in LOD, i.e., changes in $\Omega_z(\approx \Omega)$. Precession and nutation of the Earth's axis, which describe variations in Ω_x and Ω_y, are also phenomena that cannot be satisfactorily explained without invoking core-mantle torques, but they are beyond the scope of this review. It will become clear that torques on the Solid Inner Core (SIC) are also of interest. Variables in the SIC will be distinguished by a tilde ($\widetilde{\ }$) and those in the mantle by a hat ($\widehat{\ }$). Except when making general statements, unadorned letters will refer to variables in the Fluid Outer Core (FOC).

The four coupling processes

The torque on the mantle about the geocenter O is

$$\widehat{\mathbf{\Gamma}} = \oint_{\widehat{\mathrm{S}}} \mathbf{r} \times \widehat{\mathbf{T}}\, \mathrm{dA}, \quad \text{where} \quad \widehat{T}_i = -\widehat{S}_{ij} n_j, \tag{1a,b}$$

sometimes called the "surface traction," is the stress associated with the normal and $\widehat{S}_{ij}$ is the total stress tensor; $\mathbf{r} = r\mathbf{1}_r$ is the radius vector from O, $r = |\mathbf{r}|$. The minus sign in Equation 1b arises because our unit normal, $\mathbf{n}$, to $\widehat{\mathrm{S}}$, the Core Mantle Boundary (CMB), points into the mantle. The axial torque, i.e., the component of $\widehat{\mathbf{\Gamma}}$ along the rotation axis Oz, is

$$\widehat{\Gamma}_z = -\oint_{\widehat{\mathrm{S}}} s\widehat{S}_{\phi n}\, \mathrm{dA}, \tag{1c}$$

where s is distance from the z-axis. For the torque on the SIC, we have similarly

$$\widetilde{\mathbf{\Gamma}} = \oint_{\widetilde{\mathrm{S}}} \mathbf{r} \times \widetilde{\mathbf{T}}\, \mathrm{dA}, \qquad \widetilde{T}_i = \widetilde{S}_{ij} n_j,$$
$$\widetilde{\Gamma}_z = \oint_{\widetilde{\mathrm{S}}} s\widetilde{S}_{\phi n}\, \mathrm{dA}; \tag{1d,e,f}$$

there is no minus sign in Equations 1e, f because our unit normal, $\mathbf{n}$, to the Inner Core Boundary (ICB) points out of the SIC.

Equation 1a tacitly assumes that the core alone exerts a torque on the mantle; sources of torque from outside the Earth are ignored, and therefore

$$\mathbf{\Gamma} + \widehat{\mathbf{\Gamma}} + \widetilde{\mathbf{\Gamma}} = \mathbf{0},$$
$$\mathbf{M} + \widehat{\mathbf{M}} + \widetilde{\mathbf{M}} = \text{constant}, \tag{1g,h}$$

where $\mathbf{\Gamma}$ is the torque on the FOC; $\mathbf{M}$, $\widehat{\mathbf{M}}$ and $\widetilde{\mathbf{M}}$ are the corresponding angular momenta. Equation 1g implies that, if one of the three torques changes, so do one or two of the others, in the opposite sense. The system is, in this respect, self-regulating. We shall treat the mantle and SIC as rigid bodies below.

Stress is exerted on the CMB in four ways: through viscosity, topography, gravity, and magnetic field, and there are correspondingly four parts to each of S_{ij}, $\mathbf{T}$ and $\mathbf{\Gamma}$, e.g., $\widehat{\mathbf{\Gamma}} = \widehat{\mathbf{\Gamma}}^V + \widehat{\mathbf{\Gamma}}^T + \widehat{\mathbf{\Gamma}}^G + \widehat{\mathbf{\Gamma}}^M$. The topographic and gravitational torques depend on the non-sphericity of $\widehat{\mathrm{S}}$ and $\widetilde{\mathrm{S}}$. For the others, negligible error is made by replacing the CMB and ICB by spheres, $\widehat{\mathrm{S}}_\bullet$ and $\widetilde{\mathrm{S}}_\bullet$, of radii r_o and r_i. The torques will be estimated below.

The viscous torque

Assuming uniform fluid density, ρ, and kinematic viscosity, ν, the part of the viscous stress tensor, $\widehat{S}^V_{ij}$, responsible for $\widehat{T}^V_i$ is $\rho\nu\nabla_j V_i$, i.e., $\widehat{\mathbf{T}}^V = -\rho\nu(\mathbf{n}\cdot\nabla)\mathbf{V} \approx -\rho\nu\partial_r\mathbf{V}$ so that, by Equations 1a, c,

$$\widehat{\mathbf{\Gamma}}^V = -\rho\nu \oint_{\widehat{\mathrm{S}}} \mathbf{r} \times \partial_r \mathbf{V}\, \mathrm{dA},$$
$$\widehat{\Gamma}^V_z = -\rho\nu \oint_{\widehat{\mathrm{S}}} s\partial_r V_\phi\, \mathrm{dA}. \tag{2a,b}$$

These torques involve only the radial gradient of the fluid velocity, $\mathbf{V}$. This tends to drag the mantle in the direction of the subsurface flow.

The viscosity of core fluid is hard to estimate. First principles calculations (de Wijs et al., 1998; Dobson et al., 2000; Vočadlo et al., 2000) suggest that ν at the CMB is within a factor of 3 of 10^{-6} m^2 s^{-1}. Suppose the viscous stress transfers momentum between the core and mantle over a length scale $d_\nu = E^{1/2} r_o \sim 0.2$ m, which is the thickness of the laminar Ekman boundary layer. Here

$$E = \nu/\Omega\mathcal{L}^2 \tag{2c}$$

is the Ekman number ($\approx 3 \times 10^{-15}$), which quantifies the ratio of the viscous and Coriolis forces; $\mathcal{L}$ is a characteristic scale of motion outside the boundary layer, for which we take $\mathcal{L} = r_o$. Assuming velocities of order $\mathcal{V} \approx 10^{-4}$ m s^{-1} (e.g., Jackson, 1997), the surface traction, $\widehat{T}^V \approx \rho\nu\mathcal{V}/d_\nu$, is about 10^{-5} N m^{-2}, implying that $\widehat{\Gamma}^V \approx 5 \times 10^{14}$ Nm.

In a highly turbulent medium such as the FOC, small scale motions transport macroscale quantities, such as angular momentum, far more effectively than molecular diffusion. When these scales are too small to be resolved by numerical computations, they are termed "sub-grid scales." Their effect must then be included in some other way. Appeal is often made to an analogy with molecular diffusion. Molecular transport depends on the molecular mean-free-path, ℓ, and the rms molecular speed, u. Simple

dynamical arguments show that the molecular diffusivity, ν, for momentum is of order $u\ell$. The analogy pictures small scale eddies replacing molecules as the transporters of macroscale momentum. The correlation length, $l_{\rm cor}$, of the turbulence replaces ℓ and the rms turbulent velocity, υ, replaces u. The transport of macroscale momentum is then governed by a kinematic "turbulent viscosity" $\nu_T \approx \upsilon l_{\rm cor}$, which greatly exceeds ν. We use $\upsilon = 10^{-4}$ m s^{-1} for the velocity scale. The length scale is harder to estimate. Here we take $l_{\rm cor} \sim E^{1/3} r_o \approx 100$ m, which is based upon the characteristic length scale for finite amplitude rapidly rotating convection (e.g., Stellmach and Hansen, 2004; Sprague et al., 2006). This gives $\nu_T \approx 10^{-2}$ m^2 s^{-1}. Though it is strictly inconsistent to use υ and $l_{\rm cor}$ when estimating the shear on the resolved scales, an upper bound on the turbulent traction follows from doing so: $\widehat{T}^V \approx \rho \nu_T \upsilon / l_{\rm cor} \approx \rho(\upsilon l_{\rm cor})\upsilon / l_{\rm cor} \approx 10^{-4}$ N m^{-2}, implying $\widehat{\Gamma}^V_z \lessapprox 5 \times 10^{15}$ Nm.

This may overestimate $\widehat{\Gamma}^V_z$. A stable layer may exist at the top of the FOC; see, e.g., Loper (2007), Buffett (2010). Braginsky (1999) pointed out that the light material released during the freezing of the SIC may preferentially congregate near the ICB, and that this may answer unresolved questions about the geomagnetic secular variation (Braginsky, 1984). Turbulent motions in a buoyantly stable layer tend to be damped preferentially in the direction of stratification, and this reduces the macroscale momentum transport across the layer (e.g., Gargett, 1984; Davidson, 2004). Waves in such a layer may increase what ν alone can do in transporting macroscale momentum (e.g., Rogers and Glatzmaier, 2006). But it is doubtful if they can transport it as effectively as ν_T for fully convective turbulence far from boundaries.

Though these and similar arguments lack rigor, the above estimates of $\widehat{\Gamma}^V$ are less than the target torque, but not vastly so (cf. Kuang and Bloxham, 1997; Brito et al., 2004; Deleplace and Cardin, 2006; Buffett and Christensen, 2007).

The topographic torque

The likelihood that there are inverted mountains and valleys on the CMB, and that these might create topographic torques large enough to explain the observed changes in LOD, was first suggested by Hide (1969). These irregularities are often collectively called "bumps," and their study was jokingly christened "geophrenology" by the late Keith Runcorn. Hide's idea generated much interest and literature, e.g., Anufriev and Braginsky (1975; 1977a; b), Jault and Le Mouël (1989), Kuang and Bloxham (1993, 1997), Buffett (1998) and Kuang and Chao (2001). Seismic studies infer the bump height, $\mathcal{H}$, is of order 1 km (e.g., Tanaka, 2010). See *Earth's Structure, Lower Mantle*.

Topographic coupling between the FOC and its boundaries depends on deviations from sphericity in the shapes of the CMB and ICB. The fluid pressure, p, creates surface tractions, $\widehat{\mathbf{T}}^T = p\mathbf{n}$ and $\widetilde{\mathbf{T}}^T = -p\mathbf{n}$, that are not purely radial. The resulting topographic torques on the CMB and ICB are

$$\widehat{\mathbf{\Gamma}}^T = \oint_{\widehat{\rm S}} p\mathbf{r} \times \mathbf{n}\, {\rm dA}, \qquad \widetilde{\mathbf{\Gamma}}^T = -\oint_{\widetilde{\rm S}} p\mathbf{r} \times \mathbf{n}\, {\rm dA}. \tag{3a,b}$$

Equal but opposite torques act on the FOC, so that (cf. Equation 1g)

$$\mathbf{\Gamma}^T = -\widehat{\mathbf{\Gamma}}^T - \widetilde{\mathbf{\Gamma}}^T = -\int_{\rm V} \mathbf{r} \times \boldsymbol{\nabla} p\, {\rm dV}. \tag{3c}$$

We define the CMB and ICB by

$$r = r_o + \widehat{h}(\theta, \phi), \quad r = r_i + \widetilde{h}(\theta, \phi), \tag{4a,b}$$

where $(r,\ \theta,\ \phi)$ are spherical coordinates. Equations 3a, b can be simplified if it is assumed that $\left|\widehat{h}\right|/r_o \ll 1$, $\left|\widetilde{h}\right|/r_i \ll 1$, $\left|\boldsymbol{\nabla}\widehat{h}\right| \ll 1$ and $\left|\boldsymbol{\nabla}\widetilde{h}\right| \ll 1$; see RA12. The projection of Equations 3a, b onto the spheres $\widehat{\rm S}_\bullet$ and $\widetilde{\rm S}_\bullet$ are then

$$\widehat{\mathbf{\Gamma}}^T = \oint_{\widehat{\rm S}_\bullet} \widehat{h}\mathbf{r} \times \boldsymbol{\nabla} p\, {\rm dA}_\bullet, \qquad \widetilde{\mathbf{\Gamma}}^T = -\oint_{\widetilde{\rm S}_\bullet} \widetilde{h}\mathbf{r} \times \boldsymbol{\nabla} p\, {\rm dA}_\bullet. \tag{4c,d}$$

The torques $\widehat{\mathbf{\Gamma}}^T$ and $\widetilde{\mathbf{\Gamma}}^T$ obviously depend on variations in p in the FOC, and we consider next the causes and magnitudes of these. This necessitates a considerable digression. The first step is to develop a reference state.

Convection mixes the FOC so well that, except in thin boundary layers at the CMB and ICB, it is chemically and thermodynamically homogeneous. It is therefore isentropic, i.e., its specific entropy, S, is uniform. The core is an unknown mixture of all elements, but the basics can be understood by assuming that it is a binary alloy of Fe and a lighter element, X, whose mass fraction is X. It is usually supposed that X is mostly Si or S, but it is unnecessary to be specific here. Except in boundary layers,

$$S = S_a = \text{Constant}, \qquad X = X_a = \text{Constant, in the FOC}, \tag{5a,b}$$

where the suffix $_a$ stands for "adiabatic".

Although fast enough to mix the core thoroughly, core flows are slow compared with the speed of sound, u_s ($\approx 10^4$ m s^{-1}). The primary dynamical balance is therefore hydrostatic and, allowing for centrifugal forces, it is governed by

$$\boldsymbol{\nabla} p_a = \rho_a(\mathbf{g}_a - \mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{r})) = \rho_a\left(\mathbf{g}_a + \frac{1}{2}\nabla(\mathbf{\Omega} \times \mathbf{r})^2\right), \tag{5c}$$

where $\mathbf{\Omega} = \Omega \mathbf{1}_z$ is the angular velocity of the Earth and $\mathbf{g}$ is the gravitational acceleration. Newtonian gravitation theory requires that

$$\boldsymbol{\nabla} \times \mathbf{g} = \mathbf{0}, \quad \boldsymbol{\nabla} \cdot \mathbf{g} = -4\pi G\rho, \tag{5d,e}$$

where G is the constant of gravitation. It follows that

$$\boldsymbol{\nabla} p_a = \rho_a \mathbf{g}_e = -\rho_a \boldsymbol{\nabla} \Phi_e, \quad \text{where} \quad \mathbf{g}_e = -\boldsymbol{\nabla}\Phi_e \tag{5f,g}$$

is the effective gravitational field and $\Phi_e = \Phi_a - \frac{1}{2}(\mathbf{\Omega} \times \mathbf{r})^2$ is the "effective gravitational potential." Equation 5d shows that p_a, ρ_a, ... are constant on equipotential surfaces:

$$p_a = p(\Phi_e, S_a, X_a),$$
$$\rho_a = \rho(\Phi_e, S_a, X_a) = -\partial p(\Phi_e, S_a, X_a)/\partial \Phi_e, \ldots \tag{5h,i}$$

The effect of the centrifugal forces is quantified by $\varepsilon_\Omega = \Omega^2 r/g$. This is small, varying from 1.7×10^{-3} at the CMB to 1.5×10^{-3} at the ICB. The deviation of the equipotential surfaces from spheres is therefore small, as are deviations caused by the gravitational field produced by density anomalies in the mantle and SIC, quantified below. It is therefore helpful to refer to departures from a "reference state" that is spherically symmetric: $\Phi = \Phi^s(r)$, $p = p^s(r)$, $\rho = p^s(r)$, $T = T^s(r)$, etc., where T is temperature. There are large departures from adiabaticity in boundary layers and smaller departures throughout the FOC that are of two main types:

(a) *Deviations from Φ^s due to rotation and density anomalies in the mantle and SIC.* These will be denoted by a prime, e.g., $p'_a = p_a - p^s$. Rotation produces the larger deviation but, as will be seen below, this is less relevant to LOD variations than the anomaly created by the mantle, which in the FOC is dominantly

$$\Phi'_a = A'(r/r_o)^2 \sin^2\theta \cos\ 2\phi, \quad (A' > 0) \tag{6}$$

Wahr and deVries (1989), Forte et al. (1994). The value of the constant A' is about 1,300 $m^2\ s^{-2}$, according to Defraigne et al. (1996) and about 2,300 $m^2\ s^{-2}$ according to Forte et al. (1994); our compromise will be A' =1,800 $m^2\ s^{-2}$. By Equation 5i, $p'_a \approx -\rho^s_a \Phi'_a$. Therefore $\varepsilon_a = p'_a/p^s_a$ is 10^{-4} on the CMB equator and 10^{-5} at the ICB equator.

(b) *Deviations created by convection.* We use the notation

$$\rho = \rho_a + \rho_c, \tag{7a}$$

and similarly for other variables. The Earth radiates energy into space at a rate estimated to exceed 43 TW (e.g., §4.1.5 of Schubert et al., 2001). We take an extreme position by supposing the entire 43 TW emerges from the core. The outward heat flux in the FOC is the sum of the convective heat flux q_c and the adiabatic heat flux $q_a = -KdT^s/dr$, where K is thermal conductivity. Though the latter may be as much as 5 TW, we ignore it. Then q_c = 0.28 W m^{-2} at the CMB. We take

$$q_c = \rho c_p \overline{T_c V_r}, \tag{7b}$$

where c_p (≈800 J kg^{-1} K^{-1}) is the specific heat at constant pressure, p. The overline denotes a horizontal average over the flow, which is undoubtedly highly turbulent. For $V_r = 10^{-4}$ m s^{-1}, it follows that $T_c = 3.5 \times 10^{-4}$ K. As $T_a \approx 4,000$ K (e.g., Kawai and Tsuchiya, 2009), $\varepsilon_c = T_c/T_a$ is less than 10^{-7}, which is 3–4 orders of magnitude smaller than ε_a. Even this is an overestimate; it has never been suggested that Q_c is larger than 15 TW. The smallness of ε_Ω, ε_a, and ε_c means that all these effects can be treated as perturbations of the reference state.

As $p'_a(r,\ \theta,\ \phi) \gg p_c(r,\ \theta,\ \phi)$, it is sensible to consider first the effect of p'_a by introducing the "adiabatic topographic torques" on the mantle, ICB, and FOC:

$$\widehat{\mathbf{\Gamma}}^T_a = \oint_{\widehat{S}} p_a\, \mathbf{r} \times \mathbf{n}\, dA,$$
$$\widetilde{\mathbf{\Gamma}}^T_a = -\oint_{\widetilde{S}} p_a\, \mathbf{r} \times \mathbf{n}\, dA, \tag{8a,b,c}$$
$$\mathbf{\Gamma}^T_a = -\int_V \mathbf{r} \times \boldsymbol{\nabla} p_a\, dV.$$

If the CMB (ICB) were an equipotential or if it were spherical, $\widehat{\mathbf{\Gamma}}^T_a \left(\widetilde{\mathbf{\Gamma}}^T_a\right)$ would vanish according to Equation 8a, b, but generally these torques are nonzero. They can be evaluated only if $\widehat{p}$ ($\widetilde{p}$) is known on the CMB (ICB). It will be shown in the next section that $\widehat{\mathbf{\Gamma}}^T_a$ and $\widetilde{\mathbf{\Gamma}}^T_a$ are intimately related to what we shall call "adiabatic gravitational torques." We therefore postpone further discussion and estimation of adiabatic topographic torques.

Consider the torque $\widehat{\mathbf{\Gamma}}^T_c$ associated with the convective motions in the FOC:

$$\widehat{\mathbf{\Gamma}}^T_c = \oint_{\widehat{S}_\bullet} \widehat{h}\mathbf{r} \times \boldsymbol{\nabla} p_c\, dA_\bullet,$$
$$\widehat{\Gamma}^T_{c,z} = \oint_{\widehat{S}_\bullet} \widehat{h} \partial_\phi p_c\, dA_\bullet. \tag{9a,b}$$

Reasons will be given in the section on the "Magnetic Torque" why core flow may be well described by the magnetostrophic approximation Equation 20a, and why, deep in the core, Coriolis and Lorentz forces are comparable, implying a magnetic field strength $\mathcal{B}$ there of about 2 mT, or about four times greater than the typical field strength $\mathcal{B}_o$ on the CMB. The Lorentz force is therefore 16 times less on the CMB than in the bulk of the core. Also, as g_ϕ, is small, $\partial_\phi p_c \approx -2\Omega\rho_0 r_o V_\theta \cos\theta \sin\theta$ should be a good approximation to the ϕ-component of Equation 20a on the CMB. Therefore (Hide et al., 1993)

$$\widehat{\Gamma}^T_{c,z} = -2\Omega\rho_0 r_o \oint_{\widehat{S}_\bullet} \widehat{h}(\theta,\phi) V_\theta(r_o,\theta,\phi)\cos\theta\sin\theta\,\mathrm{dA}_\bullet. \tag{9c}$$

In principle, $\widehat{\Gamma}^T_{c,z}$ can be estimated by extracting $\widehat{h}$ from seismological analysis, and by using the V_θ (r_o, θ, ϕ) inferred from the core surface motion. In practice, this is difficult and has generated controversy. Equation 9c suggests that

$$\widehat{\Gamma}^T_{c,z} = \mathrm{O}\left(2\Omega\rho\mathcal{V}\mathcal{H}r_o^3\right), \tag{9d}$$

which is 10^{18} Nm for a bump height of $\mathcal{H}$ = 100 m. Such a bump height is well within the bounds set by recent seismic investigations (e.g., Tanaka, 2010). Equation 9b indicates, however, that Equation 9d may be a serious overestimate because p_c is a single-valued function and $\partial_\phi p_c$ is as often positive as negative in the integrand of Equation 9b. Though $\pm 2\Omega\rho\mathcal{V}r_o$ is a reasonable estimate of $\partial_\phi p_c$ at most points on the CMB, considerable cancellation is likely when evaluating the integral in Equation 9b. There is even an outside chance that the cancellation might be complete; see Anufriev and Braginsky (1978b).

Reliable estimation of the convective topographic torque must probably await careful experiments and allied theory. No argument has so far convincingly demonstrated that topography can create torques of the target magnitude of 10^{18} Nm but, equally, none have shown that it cannot.

The gravitational torque

The gravitational torque on a body V of density $\rho(\mathbf{x})$ in a gravitational field $\mathbf{g}(\mathbf{x})$ is

$$\mathbf{\Gamma}^G = \int_V \rho\,\mathbf{r}\times\mathbf{g}\,\mathrm{dV}. \tag{10a}$$

This volume integral can be usefully transformed into a surface integral by drawing on the analogy between the theories governing Newtonian gravitation and electrostatics, the only difference between these theories being one of sign: like charges repel but all bodies attract gravitationally. It can be shown that

$$\rho g_i = \nabla_j S^G_{ij}, \quad \text{where} \quad S^G_{ij} = -\frac{1}{4\pi G}\left(g_i g_j - \frac{1}{2}g^2\delta_{ij}\right) \tag{10b,c}$$

is the gravitational stress tensor, the gravitational pressure $-g^2/8\pi G$ being also the gravitational energy density. Equations 10b, c enable Equation 10a to be written as a surface integral:

$$\mathbf{\Gamma}^G = -\frac{1}{4\pi G}\oint_S \mathbf{r}\times\left[(\mathbf{n}\cdot\mathbf{g})\mathbf{g} - \frac{1}{2}g^2\mathbf{n}\right]\mathrm{dA}, \tag{10d}$$

where $\mathbf{n}$ points out of V. See RA12 and Appendix B of Braginsky and Roberts (1995) for derivations of these results. When they are applied below, $\mathbf{n}$ on the CMB and SIC will, as previously defined, be oriented approximately parallel to $\mathbf{r}$.

By Equation 10a, the gravitational torque on the FOC due to the mantle and SIC is

$$\mathbf{\Gamma}^G = \int_V (\rho_a+\rho_c)\mathbf{r}\times(\mathbf{g}_a+\mathbf{g}_c)\mathrm{dV} = \mathbf{\Gamma}^G_a + \mathbf{\Gamma}^G_c, \tag{11a}$$

where $\mathbf{\Gamma}^G_a$ and $\mathbf{\Gamma}^G_c$ are the adiabatic and convective parts of $\mathbf{\Gamma}^G$:

$$\mathbf{\Gamma}^G_a = \int_V \rho_a\mathbf{r}\times\mathbf{g}_a\,\mathrm{dV}, \qquad \mathbf{\Gamma}^G_c = \int_V \mathbf{r}\times(\rho_c\mathbf{g}_a + \rho_a\mathbf{g}_c + \rho_c\mathbf{g}_c)\mathrm{dV}. \tag{11b,c}$$

It was pointed out earlier that $\rho_c = \mathrm{O}(10^{-4}\rho'_a)$, $g_c = \mathrm{O}(10^{-4}g'_a)$, etc. Apparently therefore $\Gamma^G_c = \mathrm{O}(10^{-4}\Gamma^G_a)$, making it sensible to focus first on the adiabatic torque.

Consistent with hydrostatic balance in the FOC (see Equation 5f), Equations 8c and 11b, give

$$\mathbf{\Gamma}^{G+T}_a = \mathbf{\Gamma}^G_a + \mathbf{\Gamma}^T_a = \int_V \mathbf{r}\times(\rho_a g_a - \boldsymbol{\nabla}p_a)\mathrm{dV} = \mathbf{0}. \tag{12a}$$

By Equations 8a, b and 10d and the continuity of $\mathbf{g}$ and p, we also have

$$\widehat{\mathbf{\Gamma}}^{G+T}_a = \oint_{\widehat{S}} \mathbf{r}\times\left\{p_a\mathbf{n} + (4\pi G)^{-1}\left[(\mathbf{n}\cdot\mathbf{g}_a)g_a - \frac{1}{2}g_a^2\mathbf{n}\right]\right\}\mathrm{dA}, \tag{12b}$$

$$\widetilde{\mathbf{\Gamma}}^{G+T}_a = -\oint_{\widetilde{S}} \mathbf{r}\times\left\{p_a\mathbf{n} + (4\pi G)^{-1}\left[(\mathbf{n}\cdot\mathbf{g}_a)g_a - \frac{1}{2}g_a^2\mathbf{n}\right]\right\}\mathrm{dA}, \tag{12c}$$

from which, in agreement with Equation 12a,

$$\widehat{\mathbf{\Gamma}}^{G+T}_a + \widetilde{\mathbf{\Gamma}}^{G+T}_a = \mathbf{0}. \tag{12d}$$

If the torques $\widehat{\mathbf{\Gamma}}^{G+T}_a$ and $\widetilde{\mathbf{\Gamma}}^{G+T}_a$ are nonzero, $\widehat{\mathbf{\Omega}}$ and $\widetilde{\mathbf{\Omega}}$ evolve. The consequent change in the relative orientation of the mantle and SIC modifies $\mathbf{g}_a$ and p_a in Equation 12b, c so that the system evolves toward a configuration of minimum energy E. In this configuration

$$\widehat{\mathbf{\Gamma}}^{G+T}_a = \mathbf{0}, \qquad \widetilde{\mathbf{\Gamma}}^{G+T}_a = \mathbf{0}. \tag{13a,b}$$

If this minimum energy, torque-free state is perturbed, the restoring GT-torques (as we shall call them) set up a "gravitational oscillation" of the SIC relative to the mantle.

To give a simple example, suppose that the gravitational anomaly defined by Equation 6 is created by sources entirely within the mantle, the CMB having no

bumps, so that $\widehat{\Gamma}^{G+T}_{a,z} = \widehat{\Gamma}^{G}_{a,z}$. Similarly, suppose the SIC is spherical but has internal sources that above the ICB produce the gravitational anomaly

$$\Phi''_a = A''(r_i/r)^3 \sin^2\theta \cos 2(\phi - \varphi), \quad (A'' > 0), \tag{14a}$$

where φ is the angular displacement of the system from the stable state, $\varphi = 0$, in which Equations 13a, b hold. Equations 12b, c show that

$$\widehat{\Gamma}^{G}_{a,z} = -\widetilde{\Gamma}^{G}_{a,z} = \Gamma^{G}_{0} \sin 2\varphi, \quad \text{where} \quad \Gamma^{G}_{0} = 8A'A''r_i^3/3Gr_o^2 (> 0). \tag{14b,c}$$

These torques vanish for the stable minimum energy states $\varphi = 0, \pi$ (and also for the unstable $\varphi = \pm\frac{1}{2}\pi$). Small departures from a stable state satisfy

$$\widehat{C} d_t^2 \widehat{\varphi} = 2\Gamma^{G}_{0}(\widetilde{\varphi} - \widehat{\varphi}), \quad \widetilde{C} d_t^2 \widetilde{\varphi} = 2\Gamma^{G}_{0}(\widetilde{\varphi} - \widehat{\varphi}), \tag{14d,e}$$

where $\widehat{C}$ (= 7.12×10^{37} kg m^2) and $\widetilde{C}$(= 5.86×10^{34} kg m^2) are the polar moments of inertia of mantle and SIC, respectively. The frequency, ω^G, of the oscillation is

$$\omega^G = [2(\widehat{C} + \widetilde{C})\Gamma^{G}_{0}/\widehat{C}\widetilde{C}]^{1/2} \approx [2\Gamma^{G}_{0}/\widetilde{C}]^{1/2}. \tag{14f}$$

Although an anomaly of the form of Equation 14a could be produced by density variations within the SIC, it is more plausibly created by SIC topography for the following reasons. It is generally believed that the SIC is the result of freezing of the FOC, an ongoing process even today (Jacobs, 1953). An alloy generally changes its composition when it changes phase. The rather large density jump at the ICB, $\widetilde{\Delta} = \widetilde{\rho}(r_i) - \rho(r_i) \approx 600$ kg m^{-3}, is hard to attribute to contraction on freezing but can be readily explained as a small reduction in X on freezing. Phase equilibrium at the ICB implies $T^s_a(\Phi_a, S_a, X_a) = T_m(p_a, X_a)$, where T_m is the melting temperature. This implies that the ICB is an equipotential surface. Since $\Phi = \Phi^s_a(r) + \Phi'_a(r,\theta,\phi)$ where $|\Phi'_a/\Phi^s_a| \ll 1$, Taylor expansion, using Equation 4b, shows that $\Phi^s_a(r_i) + \widetilde{g}\widetilde{h}(\theta,\phi) + \Phi'_a(r_i,\theta,\phi)$ is approximately constant, where $\widetilde{g} = -g^s_{a,r}(r_i) = \partial_r \Phi^s_a(r_i) (> 0)$ is gravity at the ICB. It follows that

$$\widetilde{h} = -\Phi'_a(r_i,\theta,\phi)/\widetilde{g} = \widetilde{\varepsilon} \sin^2\theta \cos 2\phi, \quad \widetilde{\varepsilon} = -A'(r_i/r_o)^2/\widetilde{g}. \tag{14g,h}$$

This shows how the gravitational anomaly in the mantle imposes its $n = m = 2$ preference on the SIC. It makes the otherwise ad hoc assumption of Equation 14a seem perfectly reasonable. The condition that $\widetilde{h}$ creates Equation 14a, for $\varphi = 0$ and $r - r_i \ll \widetilde{h}$ is

$$A''/A' = 4\pi G r_i^3 \widetilde{\Delta}/5r_o^2\widetilde{g} \approx 0.0034. \tag{14i}$$

The maximum bump height on the ICB is $|\widetilde{\varepsilon}| \approx 50$ m. Nonzero φ corresponds to a rotated SIC. Such a rotation is to be expected; core turbulence continually subjects the SIC to (topographic) torques that continuously change its orientation. Two relaxation processes act to restore the ICB to its equipotential: (i) flow within the SIC, (ii) new freezing/melting on the ICB. If either were instantaneous, there would be no torque between the mantle and SIC, but both appear to act on much longer time scales than core turbulence, so that SIC topography is almost "frozen" to the SIC as it turns.

Concerning (i), the viscosity of the SIC is plausibly much less than the viscosity of the mantle. According to Schubert et al. (2001) $\widehat{\nu} \approx 5 \times 10^{19}$m^2 s^{-1} in the deep mantle but, according to Mound and Buffett (2006), $\widetilde{\nu} \gtrsim 10^{13}$m^2 s^{-1}. Whereas mantle anomalies are essentially "frozen in," slow motions within the SIC created by stresses exerted by the FOC on the ICB can gradually restore equilibrium (Yoshida et al., 1996; Buffett, 1997). Concerning (ii), the thermodynamic disequilibrium created by the misalignment of the ICB from its equipotential surface is slowly removed by new freezing of the FOC or new melting of the SIC. This processes has not been fully explored (but see Fearn et al., 1981; Morse, 1986). The possible significance of melting/freezing processes on SIC structure has been recently investigated by Alboussière et al. (2010) and Monnereau et al. (2010).

Interest in gravitational torques and oscillations was sparked by Buffett (1996). We follow him but by a different method, making use of Equations 14c, i:

$$\Gamma^{G}_{0} = 32\pi r_i^6 \widetilde{\Delta} A'^2/15 r_o^4 \widetilde{g} \approx 6.7 \times 10^{19}\,\text{Nm} \tag{15a}$$

By Equation 14f, the frequency of the oscillation is $\omega^G \approx 4.8 \times 10^{-8}$ s^{-1}, with a period of

$$\tau^G = 2\pi/\omega^G \approx 4.1 \text{ years}. \tag{15b}$$

According to Buffett et al. (2009), gravitational oscillations are mainly responsible for the LOD variations shown in Figure 1. Mound and Buffett (2006) obtain $\Gamma^{G}_{0} \approx 1.5 \times 10^{20}$ Nm.

Equations 14d, e imply, for some t_0,

$$\widehat{\Omega} = \widehat{\Omega}_0 + d_t\widehat{\phi} = \widehat{\Omega}_0 + (2\Gamma^{G}_{0}/\widehat{C}\omega^G)(\widetilde{\phi} - \widehat{\phi})_{\max} \sin\omega^G(t - t_0). \tag{15c}$$

The amplitude of the gravitational oscillation is therefore related to that of the variation, ΔP, in LOD by

$$\left(\widetilde{\phi} - \widehat{\phi}\right)_{\max} = \pi\widehat{C}\omega^G \Delta P/\Gamma^{G}_{0} P_0^2 \approx 1.2^\circ, \tag{15d}$$

for $\Delta P = 1$ ms. This gives a maximum angular velocity difference $\omega^G\left(\widetilde{\phi} - \widehat{\phi}\right)_{\max} \approx 2^\circ$ year^{-1}. Furthermore the peak-to-peak variation in the radial gravitational acceleration at the Earth's surface, $r = r_E$, is

$$\Delta g_r''(r_E) = \frac{12A''}{r_i}\left(\frac{r_i}{r_E}\right)^4 (\widehat{\phi} - \widetilde{\phi})^2_{\max} \simeq 4\,\text{nGal}. \quad (15e)$$

This value is too small by roughly a factor of 5 to be detectable by the GRACE satellite system (e.g., Wahr et al., 2006).

One limitation of this analysis is the neglect of electromagnetic stresses at the boundaries when they are in motion relative to the core fluid. It has been implicitly assumed that the SIC is completely decoupled from the fluid within the tangent cylinder (TC), the imaginary cylinder that touches the ICB at its equator. This is particularly significant because the fluid dynamics inside and outside the TC are quite dissimilar. See Hide and James (1983), Heimpel and Aurnou (2007), and the next section. Because the SIC is as good an electrical conductor as the FOC (or better), it may be tightly coupled magnetically to $\mathcal{C}^N$ and $\mathcal{C}^S$, as suggested by Braginsky (1970); see the next section. To examine the effect of this coupling, we make the extreme assumption that all the fluid in the TC is completely locked to the SIC. Because the mantle is a poor electrical conductor, this fluid is not well coupled to the mantle, so that the entire fluid column within the TC can co-rotate about O_z almost freely with the SIC. This suggests that, instead of Equations 1g, h, a more useful division of the total angular momentum of the Earth might be based on

$$\widehat{\mathbf{\Gamma}} + \mathbf{\Gamma}^{\text{TC}} + \mathbf{\Gamma}^{\text{XTC}} = \mathbf{0}, \qquad \widehat{\mathbf{M}} + \mathbf{M}^{\text{TC}} + \mathbf{M}^{\text{XTC}} = \text{constant}, \quad (16a,b)$$

where $^{\text{TC}}$ refers to the TC and SIC locked together, and $^{\text{XTC}}$ refers to the part of the FOC exterior to the TC. The moment of inertia of the fluid within the TC is 2.12×10^{35} kg m^2 which, when added to $\widetilde{C}$, gives $C^{\text{TC}} = 2.71 \times 10^{35}$ kg m^2. Using this instead of $\widetilde{C}$ in Equation 14f, τ^{G} is lengthened from 4.1 years to

$$\tau^G_{\text{TC}} = 2\pi/\omega^G_{\text{TC}} \approx 8.9 \text{ years}, \quad \text{where } \omega^G_{\text{TC}} = [2(\widehat{C} + C^{\text{TC}})\Gamma^G_0/\widehat{C}C^{\text{TC}}]^{1/2}. \quad (16c,d)$$

See also Mound and Buffett (2006).

Even though ρ_c/ρ_a, p_c/p_a, etc., are of order 10^{-4}, this does not mean that $\Gamma_c^{G+T}/\Gamma_a^{G+T}$ is as small as that. In fact, Equation 13a shows that $\mathbf{\Gamma}_a^{G+T} = \mathbf{0}$ in the minimum energy state. The adiabatic GT-torques dominate the convective torques only if φ is sufficiently large. Stated another way, a convective torque can be nullified by a small departure from the minimum energy state. Earlier, the torque on the SIC created by core turbulence was held responsible for causing φ to deviate from zero. This torque is essentially (magneto-)convective, and is nullified by the GT-torque for a tiny change in φ. Another way of estimating how tiny this φ is equates the magnitudes of the GT-torque, taken as $1.3 \times 10^{20}\,\varphi$ Nm (see Equations 14b, c and 15d), and the convective torque, taken to have its target value of 10^{18} Nm. This gives $\varphi \approx 0.5°$. Within this angle, the mantle and SIC are gravitationally locked together, over short time scales compared with those of the relaxation processes in the SIC described above. See Buffett and Glatzmaier (2000).

The magnetic torque

This section assumes that readers are familiar with pre-Maxwell EM theory and the fundamentals of MHD, including the frozen flux theorem and Alfvén waves. Davidson (2001) contains the necessary background.

It may be useful to remind readers that the magnetic torque about O on a body V carrying a current of density $\mathbf{J}$ is the integrated moment of the Lorentz force, $\mathbf{J} \times \mathbf{B}$:

$$\mathbf{\Gamma}^M = \int_V \mathbf{r} \times (\mathbf{J} \times \mathbf{B})\,\text{dV} = \int_V r[B_r\mathbf{J} - J_r\mathbf{B}]\,\text{dV}. \quad (17a)$$

The Lorentz force can be expressed as a divergence:

$$(\mathbf{J} \times \mathbf{B})_i = \nabla_j S^M_{ij}, \quad \text{where } S^M_{ij} = \mu_0^{-1}\left(B_iB_j - \frac{1}{2}B^2\delta_{ij}\right) \quad (17b,c)$$

is the magnetic stress tensor. An alternative form of Equation 17a is therefore

$$\mathbf{\Gamma}^M = \mu_0^{-1}\oint_S \mathbf{r} \times \left[(\mathbf{n}\cdot\mathbf{B})\mathbf{B} - \frac{1}{2}B^2\mathbf{n}\right]\text{dA}, \quad (17d)$$

where the unit normal $\mathbf{n}$ to S points out of V. Therefore

$$\widehat{\Gamma}^M_z = -\mu_0^{-1}\oint_{\widehat{S}} s\widehat{B}_r\widehat{B}_\phi\,\text{dA}, \quad \widetilde{\Gamma}^M_z = \mu_0^{-1}\oint_{\widehat{S}} s\widetilde{B}_r\widetilde{B}_\phi\,\text{dA}. \quad (17e,f)$$

These results can be used as they stand to assess the magnetic coupling between the inferred core surface flow and the mantle. See Stix and Roberts (1984), Love and Bloxham (1994), Holme (1998). Here, however, we are more interested in forging a link between the observed changes in LOD and torsional waves. To explain what the latter are, it is necessary to consider some dynamical issues.

Most studies of core MHD are based on the Boussinesq approximation; see, e.g., Braginsky and Roberts (2007). This assumes constant density, $\rho_0 (\approx 10^4 \text{kg m}^{-3})$, and expresses conservation of mass and momentum as

$$\nabla\cdot\mathbf{V} = 0, \quad \partial_t\mathbf{V} + \mathbf{V}\cdot\nabla\mathbf{V} + 2\Omega\times\mathbf{V} = -\nabla(p_c/\rho_0) + \mathcal{C}\mathbf{g}_e + \mathbf{J}\times\mathbf{B}/\rho_0 + \nu_T\nabla^2\mathbf{V}. \quad (18a,b)$$

The accelerations in Equation 18b are from inertia ($\partial_t\mathbf{V}$ and $\mathbf{V}\cdot\nabla\mathbf{V}$), rotation ($2\mathbf{\Omega}\times\mathbf{V}$), pressure ($p_c$), buoyancy ($C\mathbf{g}_e$), magnetic field ($\mathbf{J}\times\mathbf{B}/\rho_0$), and viscosity ($\nu_T\nabla^2\mathbf{V}$). Thermal and compositional buoyancy, combined in the

codensity C (Braginsky and Roberts, 1995), maintains the flow and the magnetic field; see *Core Dynamo*.

The Coriolis force is generally more significant than the inertial and viscous forces. This is indicated by the smallness of the Ekman and Rossby numbers:

$$Ro = \mathcal{V}/\Omega\mathcal{L}. \tag{19}$$

See Equation 2c for the definition of E. From $\mathcal{V} = 10^{-4}$ m s^{-1}, $\mathcal{L} = r_o$ and $\nu_T = 10^{-2}$ m^2 s^{-1} follow $Ro \approx 10^{-6}$ and $E \approx 10^{-11}$. This suggests that the inertial and viscous terms can be safely omitted from Equation 18b, except on small length scales.

If the inertial and viscous forces are ejected from Equation 18b, it becomes

$$2\rho_0\mathbf{\Omega} \times \mathbf{V}^{\mathrm{N}} = -\nabla p_c + \rho_0 \mathcal{C}\mathbf{g}_{\mathrm{e}} + \mathbf{J} \times \mathbf{B}, \tag{20a}$$

where $^{\mathrm{N}}$ stands for "non-geostrophic," and "geostrophic" is defined below. Equations 18a and 20a define the "magnetostrophic approximation," often used to describe the quasi-steady macroscales of core MHD. As the viscous term has been ejected, the only boundary condition that $\mathbf{V}^{\mathrm{N}}$ must obey is

$$\mathbf{n} \cdot \mathbf{V}^{\mathrm{N}} = 0, \quad \text{on the CMB and ICB.} \tag{20b}$$

The full boundary conditions of continuous $\mathbf{B}$ and $\mathbf{n} \times \mathbf{E}$ still apply. Because $\partial_t \mathbf{V}$ has been ejected from Equation 20a, there are no Alfvén waves. Instead, the system evolves on the much longer ageostrophic time scale,

$$\tau^{\mathrm{N}} = 2\Omega\mathcal{L}^2/V_{\mathrm{A}}^2 = \tau_\eta/\Lambda, \quad \text{where } \tau_\eta = \mathcal{L}^2/\eta \approx 10^5 \text{ years} \tag{20c,d}$$

is the free decay time for magnetic fields of scale $\mathcal{L}$, and Λ is the "Elsasser number":

$$\Lambda = V_{\mathrm{A}}^2/2\Omega\eta, \quad \text{where} \quad V_{\mathrm{A}} = \mathcal{B}/\sqrt{(\mu_0\rho_0)} \tag{20e,f}$$

is the Alfvén velocity. Elsasser (1946) suggested that $\mathcal{B}$ is determined by a balance of Lorentz and Coriolis forces. Taking $\mathcal{J} \approx \sigma\mathcal{V}\mathcal{B}$, this implies $\Lambda = 1$, $\mathcal{B} \approx 2$ mT, and $V_{\mathrm{A}} \approx 2$ cm s^{-1}. It also gives $\tau^{\mathrm{N}} \approx \tau_\eta \approx 2 \times 10^5$ years for $\mathcal{L} = r_o$.

In cylindrical coordinates (s, ϕ, z), the ϕ-component of Equation 20a is

$$2\Omega\rho_0 V_s^{\mathrm{N}} = -\partial_\phi p_c + (\mathbf{J} \times \mathbf{B})_\phi. \tag{21a}$$

Integrate this over the surface, $\mathcal{C}(s)$, of the circular cylinder of radius s ($> r_i$) coaxial with Oz. The left-hand-side vanishes by mass conservation, as can be verified by integrating Equation 18a over the interior, $\mathcal{I}(s)$, of $\mathcal{C}(s)$ and applying Equation 20b to $\mathcal{N}(s)$ and $\mathcal{S}(s)$, the spherical caps of $\mathcal{C}(s)$ on the CMB that complete the boundary of $\mathcal{I}(s)$. It follows that

$$\int_{\mathcal{C}(s)} (\mathbf{J} \times \mathbf{B})_\phi \, \mathrm{dA} = 0. \tag{21b}$$

If $s < r_i$, there are two cylinders, $\mathcal{C}^N(s)$ and $\mathcal{C}^S(s)$, of radius s to the north and south of the SIC for which

$$\int_{\mathcal{C}^N(s)} (\mathbf{J} \times \mathbf{B})_\phi \, \mathrm{dA} = 0, \quad \int_{\mathcal{C}^S(s)} (\mathbf{J} \times \mathbf{B})_\phi \, \mathrm{dA} = 0. \tag{21c,d}$$

Equations 20b–d are examples of "Taylor's constraint" (Taylor, 1963). The cylinders $\mathcal{C}(s)$ are termed "Taylor cylinders." Of these, $\mathcal{C}(r_i)$ is the tangent cylinder (TC).

It is obviously possible to assign a $\mathbf{J}$ which creates a $\mathbf{B}$ that contradicts Equations 21b–d, at least initially. This shows that Equation 20a is an oversimplification. That approximation rested on discarding the inertial force in comparison with the Coriolis force. Consider however the class of "geostrophic flows":

$$\mathbf{v} = \upsilon(s,\ t)\, \mathbf{1}_\phi. \tag{22a}$$

The corresponding Coriolis acceleration is

$$2\rho_0\mathbf{\Omega} \times \mathbf{v} = -\nabla\mathcal{X}, \quad \text{where} \quad \mathcal{X} = 2\Omega\rho_0 \int \upsilon(s,t)\, \mathrm{d}s \tag{22b,c}$$

can be absorbed into p_c. Coriolis forces are therefore totally ineffective when they act on geostrophic flows. Other forces previously abandoned in comparison with Coriolis forces become influential, especially the inertial forces, which must be restored when analyzing the geostrophic part of core flow. This recovers the Alfvén wave, or something very like it, called the "torsional wave." They share a common time scale:

$$\tau_{\mathrm{A}} = r_o/V_{\mathrm{A}} \approx 6 \text{ years}. \tag{22d}$$

That this is also the time scale $\mathrm{t_{LOD}}$ of the semi-decadal variations of in Figures 1a, b may not be a coincidence, as argued by Gillet et al. (2010).

The geostrophic part $\mathbf{v}$ of $\mathbf{V}$ can be extracted from $\mathbf{V}$ by taking the "geostrophic average," $\langle V_\phi \rangle$, of V_ϕ: for $s > r_i$, this average is defined by

$$\upsilon = \langle V_\phi \rangle \equiv \frac{1}{\widehat{A}(s)} \int_{\mathcal{C}(s)} V_\phi \, \mathrm{dA}, \quad \text{so that} \quad \mathbf{V}^{\mathrm{N}} = \mathbf{V} - \upsilon\mathbf{1}_\phi, \tag{22e}$$

where $\widehat{A}(s) = 4\pi s z_1$ is the area of $\mathcal{C}(s)$, and $z_1(s) = \sqrt{(r_o^2 - s^2)}$ is the semi-length of its sides. The axial angular momentum of the FOC is carried by υ. Therefore, insofar as the rotation of the SIC is locked to that of the fluid in the TC, the angular momentum, $M_z + \widetilde{M}_z$, of the entire core can be derived from the zonal part of the inferred core surface flow. The LOD record provides $\widetilde{M}_z$. Therefore Equation 1h can be tested: $\widehat{M}_z = -(M_z + \widetilde{M}_z)$, Results have been gratifying; see Jault et al. (1988), Jackson (1997) and Figure 1c above. The previous section indicates however that the mantle and SIC are not locked together but take part in

a gravitational oscillation having a period (4–9 years) similar to the torsional wave period $\tau_A \approx 6$ years. The implied convolvement of gravitational and magnetic coupling complicates the task of extracting information about either; see Buffett et al. (2009).

In a torsional wave, the geostrophic cylinders are in relative angular motion about their common (polar) axis; see Figure 2a. The response **b** of **B** to the motion **v** can, as for an Alfvén wave, be visualized using the frozen flux theorem, the field lines behaving like elastic strings threading the cylinders together and opposing their relative motion; see Figure 2b. The resulting torque on a cylinder supplies the restoring force for a wave, the mass of the cylinders providing the inertia. Whenever **J** and **B** contradict Equations 21b–d, a torsional wave is launched that travels in the $\pm s$-indirections.

The canonical torsional wave equation is

$$\frac{\partial^2 \zeta}{\partial t^2} = \frac{1}{s^2 \widehat{A}(s)} \frac{\partial}{\partial s}\left[s^2 \widehat{A}(s) V_A^2(s) \frac{\partial \zeta}{\partial s}\right], \tag{23}$$

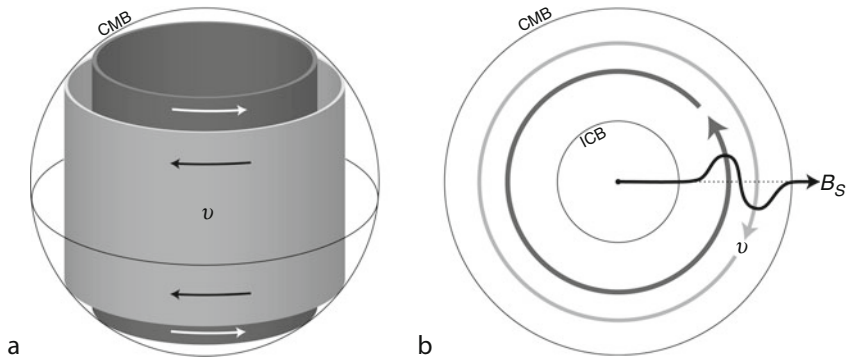
where $\zeta(s, t) = \upsilon/s$ is the angular velocity of $\mathcal{C}(s)$ and $V_A(s) = \mathcal{B}_s(s)/(\mu_0 \rho_0)^{1/2}$ is the Alfvén velocity based on the mean $\left(B_s^N\right)^2$ over $\mathcal{C}(s)$: $\mathcal{B}_s^2(s) = \left\langle \left(B_s^N\right)^2 \right\rangle$. Equation 23 is called canonical because it displays the essential nature of torsional waves clearly. It is however not only incomplete but also ignores magnetic coupling to the mantle and SIC. Equation 23 presupposes that the field, $\mathbf{B}^N$, on which the waves ride is axisymmetric. In this case, Equation 23 has a severe singularity at $s = 0$ which excludes physically acceptable solutions. This difficulty can be evaded by supposing that the TC rotates as a solid body, as suggested by Braginsky (1970), and by applying Equation 23 only in XTC. For general, non-axisymmetric $\mathbf{B}^N$, the regular singularity at $s = 0$ is harmless, but unfortunately Equation 23 is then incomplete. The terms missing from Equation 23 represent the transmission of torque from one Taylor cylinder to another by the potential field outside the core. As the terms are troublesome if retained, they are usually abandoned, with an unsubstantiated claim that they are too small to be worth keeping. Whether they are retained or abandoned, the wave equation admits normal mode solutions, i.e., solutions in which ζ is proportional to $\exp(-\imath\omega t)$, where every ω is real, as shown in RA12. When magnetic coupling to the mantle is included, ω acquires a negative imaginary part, representing the ohmic losses in the mantle. The inclusion of magnetic coupling is highly relevant to our focus here, but it clearly adds another layer of complexity.

In view of these theoretical obstacles, the reader may wonder whether computer simulation would not provide a simpler approach. It is however difficult to extract evidence of torsional waves from geodynamo simulations. This is because it is not yet possible to attain geophysically realistic magnetic Prandtl numbers, Pm, in numerical integrations. The importance of viscous damping of torsional waves can be assessed by the ratio of the torsional wave time scale, $\tau_A = r_o/V_A$, to the spin-up time scale, $\tau_{SU} = \sqrt{\left(r_o^2/2\Omega\nu\right)}$, which is their viscous decay time; see Roberts and Soward (1972). This ratio,

$$\tau_A/\tau_{SU} = (2\Omega\nu)^{1/2}/V_A = \sqrt{(Pm/\Lambda)}, \tag{24}$$

is small for the Earth ($\sim 10^{-3}$) but inconveniently large in simulations ($\gtrsim 0.1$). See Wicht and Christensen (2010) for recent simulations of torsional waves.

Space limitations do not permit the mathematical theory of torsional waves to be presented here; see Braginsky (1970), Roberts and Soward (1972), Jault (2003), Dumberry (2007) and RA12. The principal aim in what follows is to outline the underlying physics, that has been employed in theoretical studies, and that will have to be incorporated in numerical models in the future, when computer technology has advanced far enough to permit core MHD to be modeled more faithfully. Our discussion here sidesteps interesting and as yet incompletely answered questions, the first of which is whether torsional waves are involved in any essential way with the semi-



Core-Mantle Coupling, Figure 2 Schematics showing (**a**) geostrophic flows in the core, V^g, and (**b**) plan view of an initially cylindrical magnetic field (*dashed line*) distorted by *v*. The restoring Lorentz torques on the distorted magnetic field, B_s (*solid line*), lead to the cylindrical propagation of torsional waves (Adapted from Dumberry, 2007).

decadal LOD periodicity; might not the gravitational oscillation described in the last section be mainly responsible? The answer is unclear but core field strengths of 2 mT suggest that torsional waves very effectively link together geostrophic motions throughout the core, and that torsional waves therefore necessarily accompany gravitational oscillations. This obviously does not imply that torsional waves couple well to the mantle; if the mantle were an electrical insulator, there would be no magnetic coupling of core and mantle at all. The waves would then be detectable only through the inferred core surface motion. Strong magnetic coupling of the waves to the mantle raises difficult questions about both the excitation and dissipation of the waves. As shown below, the waves are mainly and heavily damped by ohmic dissipation in the mantle. What maintains the waves against this energy loss? No definitive answer has been given to this very significant question. Core turbulence may be directly responsible, or indirectly through the Reynolds magneto-stresses it exerts on the SIC or on the TC. Buffett et al. (2009) find evidence for wave generation at the TC. Core turbulence would presumably excite many damped normal modes. This might help to explain some of the recently discovered short time scales of core surface motions (Olsen and Mandea, 2008; Finlay et al., 2010). It has not been established unequivocally that the 6 year period is the fundamental mode. Though often questioned, evidence of a 60 periodicity exists, This longer period signal might be representative of the fundamental mode but would require current estimates of $\mathcal{B}_s(s)$ to be revised downward.

These matters are beyond the scope of this review. We shall merely sketch, mostly in a qualitative way, the approximate theory that is currently the most complete. We also aim to expose its strengths and weaknesses.

Underlying the entire theory is the idea that core flow can be neatly separated into geostrophic motions of short time scale, τ_A, and non-geostrophic motions of long time scale τ^{N}obeying Equation 20a. In other words, the theory focusses on the large length scales of core flow, in the belief that this includes the fundamental torsional wave mode of greatest interest. The torsional wave therefore rides on a flow satisfying Taylor's constraints Equations 21a–c. Because the Lehnert number, $\lambda = \tau_A/\tau^{\mathrm{N}} = V_A/2\Omega r_o = \omega_A/2\Omega \approx 3 \times 10^{-5}$, is small, the time dependence of $\mathbf{B}^{\mathrm{N}}$ can be ignored in torsional wave theory.

It was pointed out earlier that torsional waves are geostrophic motions in which the inertial force is crucial. The first step in deriving the torsional wave equation is therefore to restore the time-dependent inertial force to Equation 21a, obtaining

$$\rho_0 s\partial_t V_\phi + 2\Omega\rho_0 s V_s = -\partial_\phi p_c + s(\mathbf{J}\times\mathbf{B})_\phi, \qquad (25a)$$

from which $\partial_t\zeta$ is extracted, by taking the geostrophic average; see Equation 22e.

The evaluation of $\langle \mathbf{J}\times\mathbf{B}\rangle_\phi$ is simplified because E is small and the Lundquist number, $Lu = \tau_\eta/\tau_{\mathrm{A}} = V_{\mathrm{A}} r_o/\eta \approx 3\times 10^4$, of the waves is large. In a first approximation, $E = Lu^{-1} = 0$. Then viscous and ohmic diffusion and the associated boundary conditions are discarded. In particular, the electric field created in the FOC by the waves simplifies to

$$\mathbf{e} = -\mathbf{v}\times\mathbf{B}^{\mathrm{N}} - \mathbf{V}^{\mathrm{N}}\times\mathbf{b} - \mathbf{v}\times\mathbf{b}, \qquad (25b)$$

where $\mathbf{b}$ is the magnetic field of the waves. The term $-\mathbf{V}^{\mathrm{N}}\times\mathbf{B}^{\mathrm{N}}$ does not appear because it already acts on the Taylor state. The last term in Equation 25b is also ignored because the wave amplitude is assumed to be small. The ratio of $-\mathbf{V}^{\mathrm{N}}\times\mathbf{b}$ to $-\mathbf{v}\times\mathbf{B}^{\mathrm{N}}$ is of order $A^{\mathrm{N}}[b/\upsilon\sqrt{(\mu_0\rho_0)}]^{-1}$, where $A = V/V_{\mathrm{A}}$ is the Alfvén number, which for the non-geostrophic flow is about 0.1 (see above). As in an Alfvén wave, $\upsilon = \mathrm{O}[b/\sqrt{(\mu_0\rho_0)}]$, so that $|\mathbf{V}^{\mathrm{N}}\times\mathbf{b}| \approx 0.1|\mathbf{v}\times\mathbf{B}^{\mathrm{N}}|$. This, combined with the fact that the inclusion of $-\mathbf{V}^{\mathrm{N}}\times\mathbf{b}$ in $\mathbf{e}$ adds severe complications, encourages the neglect of $-\mathbf{V}^{\mathrm{N}}\times\mathbf{b}$ in Equation 25b, leaving

$$\mathbf{e} = -\mathbf{v}\times\mathbf{B}^{\mathrm{N}}. \qquad (25c)$$

The mantle and core are linked across a boundary layer of Ekman-Hartmann type; e.g., see Dormy et al. (2007). Because $Pm \ll 1$, this has a double structure. Viscosity acts only in an Ekman layer of thickness $d_\nu = \sqrt{(\nu/2\Omega)} \approx 0.1$ m; magnetic diffusion acts in a layer whose thickness is comparable with the EM skin depth, $d_\eta = \left(\frac{1}{2}|\omega|\mu_0\sigma\right)^{-1/2} \approx 10\,\mathrm{km}$. We therefore call this a "magnetic diffusion layer" (MDL), even though Coriolis and Lorentz forces affect its structure too. In the MDL, Equation 25c is replaced by

$$\mathbf{e} = -\mathbf{v}\times\mathbf{B}^{\mathrm{N}} - \eta\boldsymbol{\nabla}\times\mathbf{b} \approx -\mathbf{v}\times\mathbf{B}^{\mathrm{N}} - \eta\mathbf{1}_r\times\partial_r\mathbf{b}. \qquad (26)$$

As $d_\eta/d_\nu \gg 1$, the Ekman layer is only a tiny part of the MDL. Ekman-Hartmann theory simplifies by discarding the Ekman part, setting $\nu = 0$ and abandoning the no-slip boundary conditions. The structure of the MDL still depends on rotation, and on the Elsasser number, Λ, defined in Equation 20e.

The boundary layers play a vital role in linking ζ in the main bulk of the FOC to $\widehat{\mathbf{b}}$ on the CMB and $\widetilde{\mathbf{b}}$ on the ICB. They are therefore essential in determining $\widehat{\mathbf{\Gamma}}^M$ and $\widetilde{\mathbf{\Gamma}}^M$. At the CMB, $\mathcal{B} \approx 0.5$ mT and $\Lambda \approx 0.07$. The prevailing magnetic field, $\mathbf{B}^{\mathrm{N}}$, therefore has very little effect on the boundary layer, which is controlled almost entirely by Coriolis forces and magnetic diffusion. At the ICB, where $\mathcal{B}$ may be even an order of magnitude greater than at the CMB, $\Lambda > 1$, and Lorentz forces are too significant to ignore in the boundary layer. Further details are given in RA12.

The torque exerted by the waves on the mantle is proportional to the electrical conductivity, $\widehat{\sigma}(v, \theta, \phi)$, of the mantle, which we assume is nonzero only in the layer

$r_o < r < r_1 = r_o + d$ at the base of the mantle. The conductance of this layer is

$$\widehat{\Sigma}(\theta,\phi) = \int_{r_o}^{r_1} \widehat{\sigma}\ (r,\theta,\phi)\, dr. \tag{27a}$$

It is commonly assumed that $10^7 S < \widehat{\Sigma} < 10^9$S; we take $\widehat{\Sigma} = 10^8$S below.

Electric currents, $\widehat{\mathbf{j}}$, flow in the mantle, $\widehat{\mathrm{V}}$, either by leaking from the core or by electromagnetic induction, through the time dependence of the EM field on the CMB. We shall be interested in the penetration of the fields **b** and **e** of the waves, at frequencies ω of order 5×10^{-8} s^{-1}. The resulting magnetic and electric fields, $\widehat{\mathbf{b}}$ and $\widehat{\mathbf{e}}$, in the mantle depend on ω and on L, the horizontal length scale imposed by **b** and **e** on the CMB. Associated with ω is the skin depth of the mantle:

$$\widehat{d}_\eta = \left(\frac{1}{2}|\omega|\mu_0\widehat{\sigma}_o\right)^{-1/2}. \tag{27b}$$

Starting with Glatzmaier and Roberts (1995), theoreticians have usually simplified EM theory in the mantle by adopting the "thin-layer approximation" (TLA). This originated from the modeling of laboratory MHD experiments (see, e.g., Müller and Bühler, 2001). It is easily applied and therefore popular, although the conditions for its validity are seldom mentioned or questioned. The TLA demands that $d \ll \widehat{d}_\eta\ (\ll L)$. The horizontal part, $\widehat{\mathbf{e}}_\mathrm{H}$, of the electric field $\widehat{\mathbf{e}}$ is then independent of r, so that

$$\widehat{\mathbf{j}}_\mathrm{H} = \widehat{\sigma}\,\widehat{\mathbf{e}}_\mathrm{H}, \quad \text{and} \quad \widehat{\mathcal{J}} = \widehat{\Sigma}\,\widehat{\mathbf{e}}_\mathrm{H}, \tag{27c,d}$$

where $\widehat{\mathcal{J}}$ is the total horizontal current carried by the layer.

It may be helpful to visualize the TLA as a mathematical limit, $d \to 0$, $\widehat{\sigma} \to \infty$ with $\widehat{\Sigma}$ held fixed. Then $\widehat{\mathcal{J}}$ is a surface current responsible for a discontinuity in the magnetic field. If $\check{\mathbf{b}}(= -\boldsymbol{\nabla}\check{\mathrm{w}})$ is the potential field above the conducting layer,

$$\begin{aligned}\check{\mathbf{b}}(r_o,\theta,\phi) - \widehat{\mathbf{b}}(r_o,\theta,\phi) &= \mu_0\widehat{\mathcal{J}}(\theta,\phi)\times\mathbf{1}_r \\ &= \mu_0\widehat{\Sigma}(\theta,\phi)\widehat{\mathbf{e}}_\mathrm{H}(r_o,\theta,\phi)\times\mathbf{1}_r.\end{aligned} \tag{27e}$$

The potential field does not affect the torque, $\boldsymbol{\Gamma}^M$, on the mantle, although it does contribute to the torque that each Taylor cylinder exerts on the others.

We contrast two proposed conductivity distributions. Buffett et al. (2002) inferred from their studies of nutational resonances that d is only 200 m and that $\widehat{\sigma} = 3 \times 10^5$ S m^{-1}, which is comparable to the core conductivity σ, and gives $\widehat{d}_\eta = 10$ km. The TLA should therefore be excellent in most applications. In contrast, the laboratory experiments of Ohta et al. (2008) suggest $\widehat{\sigma} = 100$ S m^{-1} and $d = 3 \times 10^5$ m; see also Yoshino (2010). This gives $\widehat{d}_\eta = 2 \times 10^6$ m, so it is doubtful if the TLA can be validly applied. The similar conductances of the models ($\widehat{\Sigma} = 3/6 \times 10^7$ S) are insufficient to justify the use of the TLA.

This completes our critique of the basics of torsional wave coupling to the mantle. Some of its consequences are unexpected; but most are not. Even dimensional reasoning leads to

$$\begin{aligned}&\widehat{\Gamma}_z^M = \widehat{\Gamma}_0^M(\overline{\zeta} - \widehat{\zeta}), \\ &\text{where } \widehat{\Gamma}_0^M = O\left(r_o^4\widehat{\Sigma}\mathcal{B}_o^2\right) \approx 4 \times 10^{27}\,\mathrm{Nms},\end{aligned} \tag{28a,b}$$

for $\mathcal{B} = 0.5$mT . Here $\mathcal{B}_o^2$ is an average of $\left(B_r^N\right)^2$ over the CMB, and $\overline{\zeta}$ is defined by:

$$\overline{\zeta}(t) = \frac{1}{s^2\widehat{A}(s)}\int_0^{r_o} s^2\widehat{A}(s)\zeta(s,t)\mathrm{d}s \tag{28c}$$

Perhaps unexpectedly, the boundary layer on the CMB described above reduces $\widehat{\Gamma}_0^M$ by a factor of $\Sigma/(\Sigma + \widehat{\Sigma})$, where Σ is the conductance of the boundary layer, defined by

$$\Sigma = \frac{1}{2}(1 + \imath\varsigma)\sigma d_\eta = (1 + \imath\varsigma)(\sigma/2\mu_0|\omega|)^{1/2}, \tag{28d}$$

and $\varsigma = \mathrm{sgn}(\omega)$. This gives $|\Sigma| \approx 3 \times 10^9$ S, which is comparable with $\widehat{\Sigma}$. Ignoring this factor, the magnetic interaction of mantle and FOC is governed by

$$\widehat{C}d_t\widehat{\zeta} = \widehat{\Gamma}_0^M(\overline{\zeta} - \widehat{\zeta}), \qquad Cd_t\overline{\zeta} = -\widehat{\Gamma}_0^M\left(\overline{\zeta} - \widehat{\zeta}\right), \tag{28e,f}$$

where $C = 9.08 \times 10^{36}$ kg m^2 is the polar moment of inertia of the solidly rotating FOC. These equations provide an estimate of the e-folding time, $\widehat{\tau}_\eta$, taken by mantle conduction to kill ζ and $\widehat{\zeta}$: $\widehat{\tau}_\eta = C\widehat{C}/(C + \widehat{C})\widehat{\Gamma}_0^M \approx 64$ years. This is greater than the time taken by the waves to cross the core, which is $\tau_\mathrm{A} = r_o/V_\mathrm{A} \approx 5.6$ years, for $\mathcal{B} = 2$mT.

If we take $\delta\widehat{\zeta} \approx 3 \times 10^{-12}$s^{-1}, as indicated by the LOD data of the first section above, Equations 28e, f suggest that $\overline{\zeta} - \widehat{\zeta} \approx (\widehat{C} + C)\widehat{\zeta}/C$ is about 2×10^{-11} s^{-1}, so that $r_o(\overline{\zeta} - \widehat{\zeta}) \approx 7 \times 10^{-5}$ m s^{-1}, which is less than, but comparable with, the inferred core surface flow. It also gives $\widehat{\Gamma}_z^M \approx 8 \times 10^{16}$ Nm. This increases to the target torque if we take $\widehat{\Sigma} = 1.2 \times 10^9$ S, but that reduces $\widehat{\tau}_\eta$ to 5.3 years, which is less than τ_A.

This highlights a difficulty that might be called the "magnetic coupling paradox," and quantified by a quality factor:

$$\Pi = \widehat{\tau}_\eta/\tau_\mathrm{A} = (\mathcal{B}/\mathcal{B}_o)^2\left(\mu_0\widehat{\Sigma}V_\mathrm{A}\right)^{-1}. \tag{29}$$

There is a window for $\widehat{\Sigma}$, that may be narrow or nonexistent, in which $\widehat{\Gamma}_z^M$ is large enough to explain variations in LOD by torsional waves, but simultaneously is small enough to ensure that $\Pi > 1$ so that the waves are not over

damped by mantle conduction. According to the admittedly imprecise, order of magnitude estimates made here, the window is nonexistent. For the target torque to be reached or exceeded, $\widehat{\Gamma} \gtrsim 1.2 \times 10^9$ S, but $\Pi > 1$ requires $\widehat{\Gamma} \lesssim 1.1 \times 10^9$ S. See also Dumberry and Mound (2008).

So far the existence of the SIC has been ignored, almost totally. We have however recognized that, for $s < r_i$, two Taylor cylinders exist, $\mathcal{C}^N(s)$ and $\mathcal{C}^S(s)$, in which $\zeta^N(s)$ and $\zeta^S(s)$ may be expected to differ. For simplicity, we assume here that they are equal. To evaluate $\widetilde{\Gamma}_z$, it is necessary to link $\mathbf{b}$ to $\widetilde{\mathbf{b}}$ across a boundary layer strongly influenced by Λ. An analysis of the boundary layer leads to (see RA12)

$$\widetilde{b}_\phi = \mu_0 \widetilde{\sigma} \widetilde{d}_\eta s B_r^N (\overline{\zeta} - \widetilde{\zeta}) \quad \text{on} \quad r = r_i, \tag{30a}$$

where $\widetilde{\zeta}$ is the angular velocity of the SIC, and $\widetilde{d}_\eta = \left(\frac{1}{2}|\omega|\mu_0 \widetilde{\sigma}_i\right)^{-1/2} \approx d_\eta \approx 10$ km. This leads to an expression for $\widetilde{\Gamma}_z^M$, in which the main part that couples to the TC is

$$\widetilde{\Gamma}_z^M = \widetilde{\Gamma}_0^M (\overline{\zeta} - \widehat{\zeta}), \quad \text{where} \quad \widetilde{\Gamma}_0^M = \mathrm{O}\left(r_i^4 \widetilde{\sigma} \widetilde{d}_\eta \mathcal{B}_i^2\right); \tag{30b,c}$$

cf. Equations 28a, b. For $\mathcal{B}_i = 5$ mT, $\widetilde{\Gamma}_0^M \approx 2 \times 10^{29}$ Nms. This large torque acts on the SIC whose moment of inertia is less than $10^{-3}\widehat{C}$. The coupling time, $\widetilde{\tau}_\eta$, is therefore very much less than $\widehat{\tau}_\eta$. Equations analogous to Equation 28a, b give (for $\widetilde{\sigma} = \sigma$)

$$\widetilde{\tau}_\eta \approx \widetilde{C}/|\widetilde{\Gamma}_0^M| \approx 4 \text{ days}. \tag{30d}$$

This is the time taken for a mismatch between ζ and $\widetilde{\zeta}$ to be wiped out by magnetic torques. Clearly the coupling between TC and SIC is substantial at frequencies of order ω_A. This supports the opinion, advanced several times in this review, that on semi-decadal time scales, the TC is effectively locked to the SIC in its rotation about Oz.

Synthesis

In this review we have analyzed the various ways in which Earth's core is coupled to the mantle and have presented estimates of the amplitudes of these couplings in order to show which may plausibly explain the available LOD data. In our first section, we provide observational evidence for core-mantle coupling. We show that Earth's rotation rate has a roughly semi-decadal time variability, such that the LOD fluctuates at the ms level. To explain these LOD fluctuations, an internal coupling must exist between the mantle and the core that provides torques of order 10^{18} Nm, which we named "the target torque."

In the later text, we develop estimates of the strength of the viscous, topographic, gravitational, and electromagnetic torques. Only the viscous torque, Γ_z^V, appears to be too weak to explain the LOD signal. This is true even when we allow for the enhanced coupling that turbulence can provide.

The topographic torque on the mantle, $\widehat{\Gamma}_z^T$, is created by core flow interacting with bumps on the CMB. In analyzing $\widehat{\Gamma}_z^T$, it became clear that exchange of angular momentum with the SIC is significant. Order of magnitude arguments showed that potentially the largest part of $\widehat{\Gamma}_z^T$ is $\widehat{\Gamma}_{a,z}^T$, which is produced by the gravitational field of density anomalies in the mantle and possibly SIC, including bumps on their surfaces. This part of $\widehat{\Gamma}_z^T$ is therefore intimately related to the gravitational torque $\widehat{\Gamma}_z^G$. When the two are treated together as $\widehat{\Gamma}_{a,z}^{G+T}$, there is an equal but opposite torque, $\widetilde{\Gamma}_{a,z}^{G+T}$ on the SIC. Gravitational oscillations (Buffett, 1996) occur when the system is perturbed from a state of minimum gravitational energy in which $\widehat{\Gamma}_{a,z}^{G+T} = \widetilde{\Gamma}_{a,z}^{G+T} = 0$. An oscillation period of $\tau^{\mathrm{G}} = 4.1$ years was derived. If, as seems likely, strong magnetic coupling exists between the tangent cylinder (TC) and the SIC, the gravitational oscillation period increases to $\tau_{\mathrm{TC}}^G = 8.9$ years.

The remaining part of $\widehat{\Gamma}_z^T$ is $\widehat{\Gamma}_{c,z}^T$, and is produced by core convection. Its importance is uncertain. From what is known today, $\Gamma_{c,z}^T$ may be 0 Nm or may exceed the target torque (cf. Kuang and Bloxham, 1997; Hide, 1998; Jault and Le Mouël, 1999; Kuang and Chao, 2001).

A simple model of torsional waves traversing FOC can explain oscillations of period, $\tau \approx 6$ years, but there is a paradox: The target torque cannot be attained by magnetic coupling between the waves and the mantle unless a dimensionless "paradox parameter," Π, defined in Equation 29, is large enough. If this parameter is too large, however, the waves are damped out before they can cross the core. Whether the core can evade the paradox seems uncertain.

The topographic, gravitational and magnetic torques all have significant uncertainties in their amplitudes, but the target torque falls within these uncertainties, i.e., conceivable any of them could explain the semi-decadal LOD signals. The coupling processes may be convolved. The recent model of Buffett et al. (2009) allows for this, but argues that the gravitational torque dominates. In contrast, Gillet et al. (2010) infer that torsional oscillations in the FOC can explain the LOD observations without strong gravitational coupling. Improvements in data and modeling of Earth's rotation (e.g., Gross, 2009), the geomagnetic field (e.g., Hulot et al., 2002; Jackson, 2003), core seismology (e.g., Dai and Song, 2008), and the time-variations in the gravity field (e.g., Velicogna and Wahr, 2006; Dumberry, 2010) will all prove important in testing these core-mantle coupling arguments.

This review has focussed on explaining variations in LOD owing to internal coupling between the mantle and core. This coupling produces changes primarily in the axial angular rotation rate, Ω_z, on semi-decadal time scales. Detailed measurements now exist of variations in Earth's full rotation vector on many time scales (e.g., Gross, 2007), with the different directional components

providing information on different geophysical phenomena (e.g., Mathews et al., 2002). Furthermore, rotation vector and magnetic field measurements now exist for other planets (e.g., Margot et al., 2007; Uno et al., 2009), and will improve in quality in the coming decades. Such measurements will allow the development of models of deep interior structure and dynamics in planetary bodies throughout the solar system (e.g., Tyler, 2008; Noir et al., 2009; Goldreich and Mitchell, 2010).

Acknowledgments

We thank Richard Gross, Richard Holme, and Andrew Jackson for sharing their insights and their data. We are also grateful to Bruce Buffett and the referee (Mathieu Dumberry) for giving helpful advice.

Bibliography

Abarca del Rio, R., Gambis, R., and Salstein, D. A., 2000. Interannual signals in length of day and atmospheric angular momentum, *Annales Geophysicae*, **18**, 347–364.

Alboussière, T., Deguen, R., and Melzani, M., 2010. Melting induced stratification above the Earth's inner core due to convective translation. *Nature*, **466**, 744–747.

Anufriev, A. P., and Braginsky, S. I., 1975. Influence of irregularities of the boundary of the Earth's core on the velocity of the liquid and on the magnetic field. *Geomagnetism and Aeronomy*, **15**, 754–757.

Anufriev, A. P., and Braginsky, S. I., 1977a. Influence of irregularities of the boundary of the Earth's core on the fluid velocity and the magnetic field, II. *Geomagnetism and Aeronomy*, **17**, 78–82.

Anufriev, A. P., and Braginsky, S. I., 1977b. Influence of irregularities of the boundary of the Earth's core on the fluid velocity and the magnetic field, III. *Geomagnetism and Aeronomy*, **17**, 742–750.

Braginsky, S. I., 1970. Torsional magnetohydrodynamic vibrations in the Earth's core and variations in day length. *Geomagnetism and Aeronomy*, **10**, 1–8.

Braginsky, S. I., 1984. Short-period geomagnetic secular variation. *Geophysical and Astrophysical Fluid Dynamics*, **30**, 1–78.

Braginsky, S. I., 1999. Dynamics of the stably stratified ocean at the top of the core. *Physics of the Earth and Planetary Interiors*, **111**, 21–34.

Braginsky, S. I., and Roberts, P. H., 1995. Equations governing convection in Earth's core and the Geodynamo. *Geophysical and Astrophysical Fluid Dynamics*, **79**, 1–97.

Braginsky, S. I., and Roberts, P. H., 2007. Anelastic and Boussinesq approximations. In Gubbins, D., and Herrero-Bervera, E. (eds.), *Encyclopedia of Geomagnetism and Paleomagnetism*. Heidelberg: Springer, pp. 11–19.

Brito, D., Aurnou, J. M., and Cardin, P., 2004. Turbulent viscosity measurements relevant to planetary core-mantle dynamics. *Physics of the Earth and Planetary Interiors*, **141**, 3–8.

Buffett, B. A., 1996. Gravitational oscillations in the length of day. *Geophysical Research Letters*, **23**, 2279–2282.

Buffett, B. A., 1997. Geodynamic estimates of the viscosity of the Earth's inner core. *Nature*, **388**, 571–573.

Buffett, B. A., 1998. Free oscillations in the length of day: inferences on physical properties near the core-mantle boundary. *Geodynamics*, **28**, 153–165.

Buffett, B. A., 2010. Chemical stratification at the top of Earth's core: constraints from nutation observations. *Earth and Planetary Science Letters*, **296**, 367–372.

Buffett, B. A., and Christensen, U. R., 2007. Magnetic and viscous coupling at the core-mantle boundary; inferences from observations of the Earth's nutations. *Geophysical Journal International*, **171**, 145–152.

Buffett, B. A., and Glatzmaier, G. A., 2000. Gravitational braking of inner-core rotation in geo-dynamo simulations. *Geophysical Research Letters*, **27**, 3125–3128.

Buffett, B. A., Mathews, P. M., and Herring, T. A., 2002. Modeling of nutation and precession: effects of electromagnetic coupling. *Journal of Geophysical Research*, **107**, 2070, doi:10.1029/2000JB000056.

Buffett, B. A., Mound, J., and Jackson, A., 2009. Inversion of torsional oscillations for the structure and dynamics of Earth's core. *Geophysical Journal International*, **177**, 878–890.

Dai, W., and Song, X., 2008. Detection of motion and heterogeneity in Earth's liquid outer core. *Geophysical Research Letters*, **35**, L16311.

Davidson, P. A., 2001. *An Introduction to Magnetohydrodynamics*. Cambridge, UK: Cambridge University Press.

Davidson, P. A., 2004. *Turbulence*. Oxford, UK: Oxford University Press.

de Wijs, G. A., Kresse, G., Vočadlo, I., Dobson, D. P., Alfêe, D., Gillan, M. J., and Price, G. D., 1998. The viscosity of liquid iron at the physical conditions of Earth.s core. *Nature*, **392**, 805–807.

Defraigne, P., Dehant, V., and Wahr, J., 1996. Internal loading of an inhomogeneous compressible mantle with phase boundaries. *Geophysical Journal International*, **125**, 173–192.

Deleplace, B., and Cardin, P., 2006. Viscomagnetic torque at the core-mantle boundary. *Geophysical Journal International*, **167**, 557–566.

Dobson, D. P., Crichton, W. A., Vočadlo, I., Jones, A. P., Wang, Y., Uchida, T., Rivers, M., Sutton, S., and Brodhardt, J. P., 2000. In situ measurements of viscosity of liquids in the Fe-FeS system at high pressures and temperatures. *American Mineralogist*, **85**, 1838–1842.

Dormy, E., Roberts, P. H., and Soward, A. M., 2007. Core, boundary layers. In Gubbins, D., and Herrero Bervera, E. (eds.), *Encylopedia of Geomagnetism and Paleomagnetism*. Heidelberg: Springer, pp. 111–116.

Dumberry, M., 2007. Taylor's constraint and torsional oscillations. In Cardin, P., and Cugliandolo, L. F. (eds.), *Dynamos*. Amsterdam: Elsevier, pp. 383–401.

Dumberry, M., 2010. Gravity variations induced by core flows. *Geophysical Journal International*, **180**, 635–650.

Dumberry, M., and Mound, J., 2008. Constraints on core-mantle electromagnetic coupling from torsional oscillation normal modes. *Journal of Geophysical Research*, **113**, B03102, doi:10.1029/2007JB005135.

Elsasser, W. M., 1946. Induction effects in terrestrial magnetism, II. The secular variation. *Physical Review*, **70**, 202–212.

Fearn, D. R., Loper, D. E., and Roberts, P. H., 1981. Structure of the Earth's inner core. *Nature*, **292**, 232–233.

Finlay, C. C., Dumberry, M., Chulliat, A., and Pais, M. A., 2010. Short timescale core dynamics: theory and observations. *Space Science Reviews*, **155**, 177–218, doi:10.1007/s11214-010-9691-6.

Forte, A. M., Woodward, R. J., and Dziewonski, A. M., 1994. Joint inversion of seismic and geo-dynamic data for models of three-dimensional mantle heterogeneity. *Journal of Geophysical Research*, **99**, 21857–21877.

Gargett, A. E., 1984. Vertical eddy diffusivity in the ocean interior. *Journal of Marine Research*, **42**, 359–393.

Gillet, N., Jault, D., Canet, E., and Fournier, A., 2010. Fast torsional waves and strong magnetic field within the Earth's core. *Nature*, **465**(7294), 74–77, doi:10.1038/nature09010.

Glatzmaier, G. A., and Roberts, P. H., 1995. A three-dimensional convective dynamo solution with rotating and finitely

conducting inner core and mantle. *Physics of the Earth and Planetary Interiors*, **91**, 63–75.
Goldreich, P. M., and Mitchell, J. L., 2010. Elastic ice shells and synchronous moons: implications for cracks on Europa and non-synchronous rotation on Titan. *Icarus*, doi:10.1016-/j.icarus.2010.04.013.
Gross, R. S., 2001. A combined length-of-day series spanning 1832–1997: LUNAR97. *Physics of the Earth and Planetary Interiors*, **123**, 65–76.
Gross, R. S., 2007. Earth rotation variations – long period. In Herring, T. A. (ed.), *Physical Geodesy*. Oxford: Elsevier. Treatise on Geophysics, Vol. 3, pp. 239–294.
Gross, R. S., 2009. Ocean tidal effects on Earth rotation. *Journal of Geodynamics*, **48**, 219–225.
Heimpel, M. H., and Aurnou, J. M., 2007. Turbulent convection in rapidly rotating spherical shells: a model for equatorial and high latitude jets on Jupiter and Saturn. *Icarus*, **187**, 540–557.
Hide, R., 1969. Interaction between the earth's liquid core and solid mantle. *Nature*, **222**, 1055–1956.
Hide, R., 1998. A note on topographic core-mantle coupling. *Physics of the Earth and Planetary Interiors*, **109**, 91–92.
Hide, R., and James, I. N., 1983. Differential rotation produced by potential vorticity mixing in a rapidly rotating fluid. *Geophysical Journal of the Royal Astronomical Society*, **74**, 301–312.
Hide, R., Clayton, R. W., Hager, B. H., Speith, M. A., and Voorhies, C. V., 1993. Topographic core-mantle coupling and fluctuations in Earth's rotation. In Aki, K., and Dmowska, R. (eds.), *Relating Geophysical Structures and Processes: The Jeffreys Volume*. Washington, DC: AGU. Geophysical Monograph Series, Vol. 76, pp. 107–120.
Holme, R., 1998. Electromagnetic core-mantle coupling-I. Explaining decadal changes in the length of day. *Geophysical Journal International*, **132**, 167–180.
Holme, R., and de Viron, O., 2005. Geomagnetic jerks and a high-resolution length-of-day profile for core studies. *Geophysical Journal International*, **160**, 435–439.
Hulot, G., Eymin, C., Langlais, B., Mandea, M., and Olsen, N., 2002. Small-scale structure of the geodynamo inferred from Oersed and Magsat satellite data. *Nature*, **416**, 620–623.
Jackson, A., 1997. Time-dependency of tangentially geostrophic core surface motions. *Physics of the Earth and Planetary Interiors*, **103**, 293–311.
Jackson, A., 2003. Intense equatorial flux spots on the surface of Earth's core. *Nature*, **424**, 760–763.
Jacobs, J. A., 1953. The Earth's inner core. *Nature*, **172**, 297–298.
Jault, D., 2003. Electromagnetic and topographic coupling, and LOD variations. In Jones, C. A., Soward, A. M., and Zhang, K. (eds.), *Earth's Core and Lower Mantle*. London: Taylor and Francis, pp. 46–76.
Jault, D., and Le Mouël, J. L., 1989. The topographic torque associated with a tangentially geostrophic motion at the core surface and inferences on the flow inside the core. *Geophysical and Astrophysical Fluid Dynamics*, **48**, 273–296.
Jault, D., and Le Mouël, J. L., 1999. Comment on 'On the dynamics of topographic core-mantle coupling' by Weijia Kuang and Jeremy Bloxham. *Physics of the Earth and Planetary Interiors*, **114**, 211–215.
Jault, D., Gire, C., and LeMouel, J.-L., 1988. Westward drift, core motions and exchanges of angular momentum between core and mantle. *Nature*, **333**, 353–356.
Kawai, K., and Tsuchiya, T., 2009. Temperature profile in the lowermost mantle from seismological and mineral physics joint modeling. *Proceedings of the National Academy of Sciences of the United States of America*, doi:10.1073/pnas.0905920106.
Kuang, W.-J., and Bloxham, J., 1993. The effect of boundary topography on motions in the Earth's core. *Geophysical and Astrophysical Fluid Dynamics*, **72**, 161–195.
Kuang, W.-J., and Bloxham, J., 1997. On the dynamics of topographic core-mantle coupling. *Physics of the Earth and Planetary Interiors*, **99**, 289–294.
Kuang, W.-J., and Chao, B. F., 2001. Topographic core-mantle coupling in geodynamo modeling. *Geophysical Research Letters*, **28**, 1871–1874.
Loper, D. E., 2007. Turbulence and small-scale dynamics in the core. In Olson, P. L. (ed.), *Core Dynamics*. Amsterdam: Elsevier. Treatise on Geophysics, Vol. 8, pp. 187–206.
Love, J. J., and Bloxham, J., 1994. Electromagnetic coupling and the toroidal magnetic field at the core-mantle boundary. *Geophysical Journal International*, **117**, 235–256.
Margot, J. L., Peale, S. J., Jurgens, R. F., Slade, M. A., and Holin, I. V., 2007. Large longitude libration of Mercury reveals a molten core. *Science*, **316**, 710–714.
Mathews, P. M., Herring, T. A., and Buffett, B. A., 2002. Modeling of nutation and precession: new nutation series for nonrigid Earth and insights into the Earth's interior. *Journal of Geophysical Research*, **107**, 2068, doi:10.1029/2001JB000390.
Monnereau, M., Calvet, M., Margerin, L., and Souriau, A., 2010. Lopsided growth of Earth's inner core. *Science*, **328**, 1014–1017.
Morse, S. A., 1986. Adcumulus growth of the inner core. *Geophysical Research Letters*, **13**, 1466–1469.
Mound, J. E., and Buffett, B. A., 2003. Interannual oscillations in length of day: implications for the structure of the mantle and core. *Journal of Geophysical Research*, **108**, 2334, doi:10.1029/2002JB002054.
Mound, J. E., and Buffett, B. A., 2005. Mechanisms of core-mantle angular momentum exchange and the observed spectral properties of torsional oscillations. *Journal of Geophysical Research*, **110**, B08103, doi:10.1029/2004JB003555.
Mound, J., and Buffett, B., 2006. Detection of a gravitational oscillation in length-of-day. *Earth and Planetary Science Letters*, **243**, 383–389.
Müller, U., and Bühler, L., 2001. *Magnetofluiddynamics in Channels and Containers*. Heidelberg: Springer.
Noir, J., Hemmerlin, F., Wicht, J., Baca, S. M., and Aurnou, J. M., 2009. An experimental and numerical study of librationally driven flow in planetary cores and subsurface oceans. *Physics of the Earth and Planetary Interiors*, **173**, 141–152.
Ohta, K., Onada, S., Hirose, K., Sinmyo, R., Shimizu, K., Saya, N., Ohishi, Y., and Yasuhara, A., 2008. The electrical conductivity of post-perovskite in Earth's D″ layer. *Science*, **320**, 89–91.
Olsen, N., and Mandea, M., 2008. Rapidly changing flows in the Earth's core. *Nature Geoscience*, **1**, 390–394.
Roberts, P. H., and Aurnou, J. M., 2012. On the theory of core-mantle coupling. *Geophysical and Astrophysical Fluid Dynamics* (to appear).
Roberts, P. H., and Soward, A. M., 1972. Magnetohydrodynamics of the Earth's core. *Annual Review of Fluid Mechanics*, **4**, 117–154.
Roberts, P. H., Yu, Z. J., and Russell, C. T., 2007. On the 60-year signal from the core. *Geophysical and Astrophysical Fluid Dynamics*, **43**, 321–330.
Rogers, T. M., and Glatzmaier, G. A., 2006. Angular momentum transport by gravity waves in the solar interior. *Geophysical and Astrophysical Fluid Dynamics*, **653**, 756–764.
Schubert, G., Turcotte, D. L., and Olson, P., 2001. *Mantle Convection in the Earth and Planets*. Cambridge, UK: Cambridge University Press.
Sprague, M., Julien, K., Knobloch, E., and Werne, J., 2006. Numerical simulation of an asymptotically reduced system for rotationally constrained convection. *Journal of Fluid Mechanics*, **551**, 141–174.
Stellmach, S., and Hansen, U., 2004. Cartesian convection driven dynamos at low Ekman number. *Physical Review E*, **70**, 056312.

Stix, M., and Roberts, P. H., 1984. Time-dependent electromagnetic core-mantle coupling. *Physics of the Earth and Planetary Interiors*, **36**, 49–60.
Tanaka, S., 2010. Constraints on the core-mantle boundary topography from *P4KP-PcP* differential travel times. *Journal of Geophysical Research*, **115**, B04310, doi:10.1029/2009JB006563.
Taylor, J. B., 1963. The magnetohydrodynamics of a rotating fluid and the Earth's dynamo problem. *Proceedings. Royal Society of London*, **A274**, 274–283.
Tyler, R. H., 2008. Strong ocean tidal flow and heating on moons of the outer planets. *Nature*, **456**, 770–773.
Uno, H., Johnson, C. L., Anderson, B. J., Korth, H., and Solomon, S. C., 2009. Modeling Mercury's internal magnetic field with smooth inversions. *Earth and Planetary Science Letters*, **285**, 328–339.
Velicogna, I., and Wahr, J., 2006. Acceleration of Greenland ice mass loss in spring 2004. *Nature*, **443**, 329–331.
Vočadlo, I., Alfè, D., Price, G. D., and Gillan, M. J., 2000. First principles calculation of the diffusivity of FeS at experimentally accessible conditions. *Physics of the Earth and Planetary Interiors*, **120**, 145–152.
Wahr, J., and deVries, D., 1989. The possibility of lateral structure inside the core and its implications for nutation and Earth tide observations. *Geophysical Journal International*, **99**, 511–519.
Wahr, J., Swenson, S., and Velicogna, I., 2006. Accuracy of GRACE mass estimates. *Geophysical Research Letters*, **33**, L06401, doi:10.1029/2005GL025305.
Wicht, J., and Christensen, U. R., 2010. Torsional oscillations in dynamo simulations. *Geophysical Journal International*, **181**, 1367–1380.
Yoshida, S., Sumita, I., and Kumazawa, M., 1996. Growth model of the inner core coupled with outer core dynamics and the resulting elastic anisotropy. *Journal of Geophysical Research*, **101**, 28085–28103.
Yoshino, T., 2010. Laboratory electrical conductivity measurement of mantle minerals. *Surveys in Geophysics*, **31**, 163–206, doi:10.1007/s10712-009-9084-0.

Cross-references

Core Dynamo
Earth's Structure, Lower Mantle
Energy Budget of the Earth
Geomagnetic Field, Theory

CRUSTAL REFLECTIVITY (OCEANIC) AND MAGMA CHAMBER

Satish C. Singh
Laboratoire de Géoscience Marines, Institut de Physique du Globe de Paris, Paris, France

Synonyms

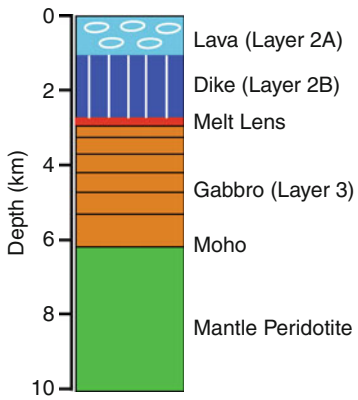
Melt lens; Spreading center and ridge

Definition

Axial magma chamber (*Melt lens*) is a thin melt lens observed at ocean spreading centers.
Layer 2A or *Lava* is the top layer of oceanic igneous crust.
Layer 2B is a dike sequence and lies above the axial melt lens.
Layer 3 (*gabbro*) forms the lower oceanic crust.
Moho is a boundary between the crust and mantle.
Pg is a seismic ray that travels through the crust.
PmP is a reflection from the crust–mantle boundary.
Pn is a ray that travels in the upper mantle.
Tomography is a technique to image the velocity structure of the earth.

Introduction

Over 70% of the earth's crust is formed by the cooling and crystallization of melt at ocean spreading centers, which represent over 55,000 km of chains of volcanoes in the middle of the oceans, called mid-ocean ridges. At ocean spreading centers, the oceanic plate separates causing the mantle to move upward, reducing the pressure and causing the melting of the mantle. Since the newly formed melt is lighter than the surrounding mantle material, it moves upward toward the surface of the earth. Part of the melt is erupted on the seafloor as lava, which cools very rapidly forming a cap of solid extrusive layer, also known as Layer 2A (Figure 1). As there is a significant amount of water present at mid-ocean ridges, the water circulates deep in the crust. Therefore, the melt stays mainly in the middle of the crust and erupts along thin dikes. When these dikes are cooled and crystallized, they form a thick intrusive layer or Layer 2B. Below the dikes, the melt could reside for a long period, forming a steady state melt lens, called axial melt lens or axial magma chamber (AMC). The magma cools and crystallizes in this melt lens, forming a crystalline lower crust. The melt lens forms the lower limit for further penetration of water, and therefore, partial melt is generally present beneath the melt lens down to the crust–mantle boundary. Sometimes hot melt ascending from the mantle may get injected in this partially molten region. Based on this basic process, the oceanic crust is divided into three layers, lava (extrusive), dikes (intrusive), and gabbroic crust. The relative



Crustal Reflectivity (Oceanic) and Magma Chamber, Figure 1 Classical model of the oceanic crust. Layers 2A and 2B form the upper crust whereas the gabbro layer corresponds to the lower crust.

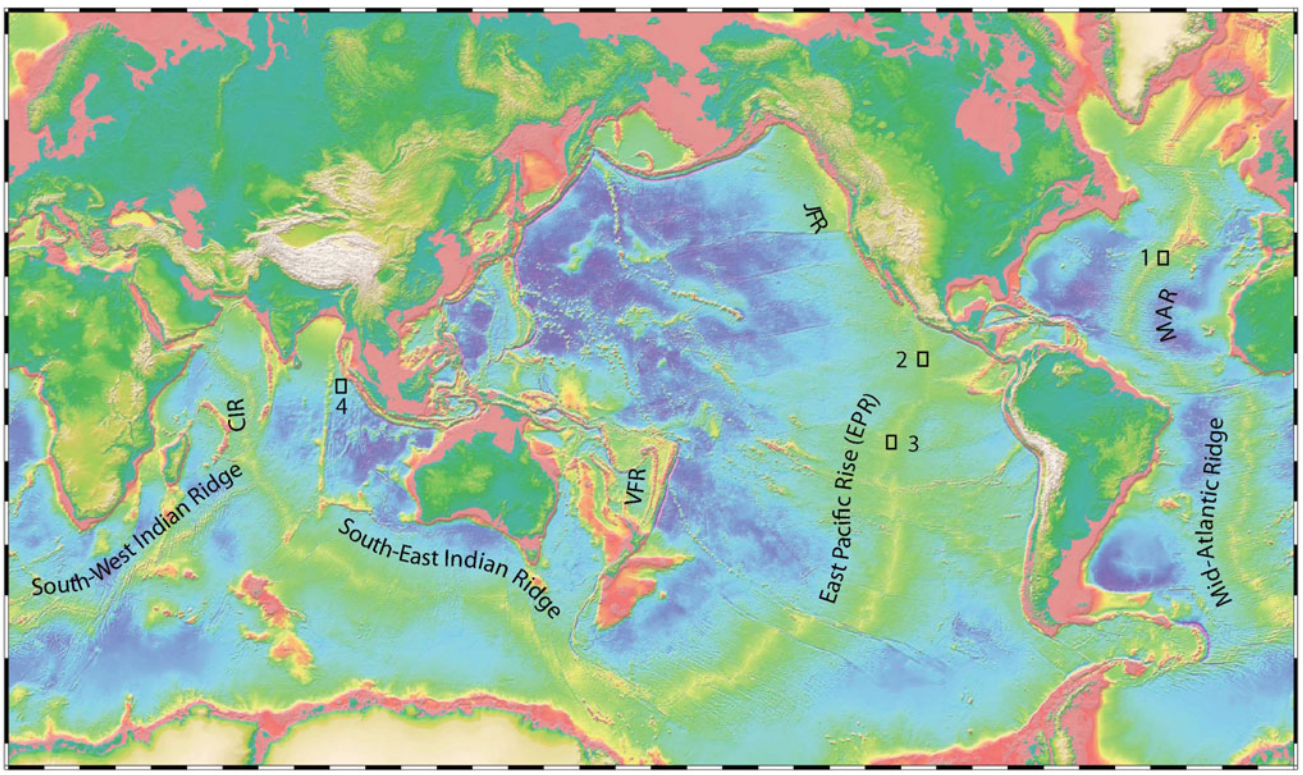
thicknesses of these layers depend on the spreading rate, which can vary from a few millimeters up to 200 mm per year.

Based on the spreading rate, mid-ocean ridges are divided into four groups: fast, intermediate, slow, and ultraslow. East Pacific Rise (EPR) is a fast spreading center, where the spreading rate varies from 70 to 180 mm per year (Figure 2). Juan de Fuca Ridge and Valu Fa Ridge are intermediate spreading centers, where the spreading rate varies from 55 to 70 mm per year. Mid-Atlantic Ridge and Central Indian Ridge are slow spreading centers with spreading rates of 20–55 mm per year. South-West Indian Ridge and Gakkel Ridge are ultraslow spreading centers with spreading rates of less than 20 mm per year. The melt production depends on the spreading rate and, therefore, the crustal thickness also varies from 3.5 to 8 km, with an average thickness of 6 km.

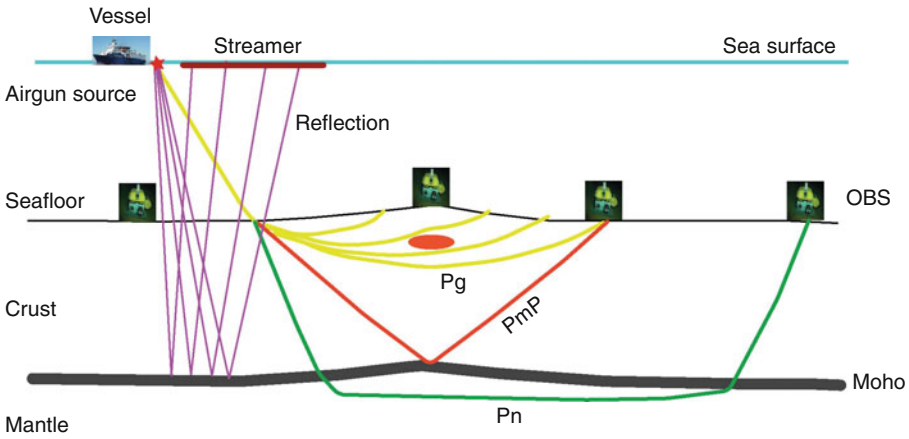
Methods

There are two main methods to study the oceanic crustal structures: seismic refraction or tomography and reflection methods. In seismic refraction method, a set of ocean bottom seismometers (OBS) are placed on the seafloor and an array of air guns is used to generate acoustic energy that travels through water column into the crust and mantle and is recorded on OBS (Figure 3). There are three types of waves that arrive on an OBS: waves that travel in the crust (Pg) and the mantle (Pn) and waves reflected from the crust–mantle boundary, Moho (PmP). The travel times of arrivals of energy on these OBS are used to determine the velocity structure of the crust and upper mantle. Since OBS are small in numbers, they are placed 5–20 km apart on the seafloor, and hence provide a very large-scale (5–10 km) velocity structure of the crust. Since the compressional (P) waves travel fast, they arrive first on OBS, and hence mainly P-wave velocities are estimated using OBS data. However, sometimes secondary (S) arrivals, converted phases Ps, are recorded on these OBS and provide information on S-wave velocity.

Seismic reflection method is most commonly used to obtain crustal reflectivity. A streamer containing pressure sensors (hydrophones) is towed behind a vessel recording seismic energy generated by an array of air guns as the vessel moves at 4.5–5 knots (Figure 3). The length of the streamer varies from a few hundred meters up to 15 km (Singh et al., 2009), and hydrophones are generally spaced at 12.5 m. For crustal imaging the air gun shot spacing is ~50 m. Depending upon the target depth, the recording length varies from 10 to 20 s (20–60 km depth). This technique is routinely used for oil and gas exploration,



Crustal Reflectivity (Oceanic) and Magma Chamber, Figure 2 Major spreading centers on the earth. JFR – Juan de Fuca Ridge; VFR – Valu Fa Ridge; CIR – Central Indian Ridge. The *rectangles* indicate the positions of data/results shown in this entry: 1 – Lucky Strike Segment at Mid-Atlantic Ridge (Figures 5, 7, 9, and 12); 2 – 9° N East Pacific Rise (Figures 8, 11, and 14); 3 – 14° S East Pacific Rise (Figures 6, 10, and 13); and 4 – Wharton basin (Figures 15 and 16).



Crustal Reflectivity (Oceanic) and Magma Chamber, Figure 3 Schematic diagram showing the seismic reflection and refraction survey to study the oceanic crust. *Red star* indicates the air gun source. OBS – ocean bottom seismometers; Pg – crustal arrival; Pn – mantle arrival; and PmP – reflection from the crust–mantle boundary. *Red ellipse* indicates melt lens.

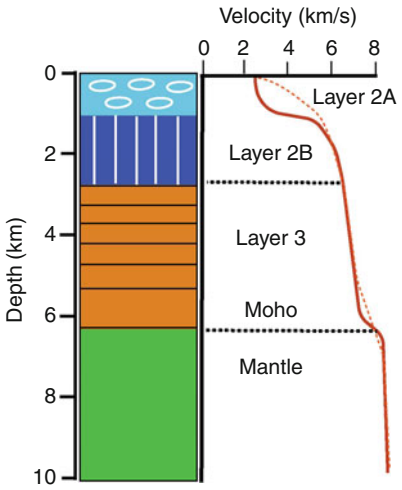
and has become a standard technique for crustal studies. There are standard tools to analyze these data. Since the source receiver distance is small as compared to the depth of the target, the data are initially presented as a function of time, and then are depth-converted using velocities obtained from the OBS study or from other sources.

Crustal structure at fast spreading centers

Based on seismic refraction and reflection studies at fast spreading centers and on ophiolites from Oman, the crust at fast spreading centers is considered as a layer cake and is divided into three distinct layers: Layer 2A (pillow lava), Layer 2B (dike sequence), and Layer 3 (gabbro) (Figure 4). The velocity in Layer 2A varies from 2.5 to 4.5 km/s, that in Layer 2B from 4.5 to 6 km/s, and in gabbro from 6 to 7 km/s. On the ridge axis, a magma chamber is generally present between the dike and the gabbro layer.

The structure of Layer 2A is inferred either using the velocity from OBS study or from the reflection method. The boundary between Layer 2A and 2B is not sharp but there is a high velocity gradient where the velocity increases from 2.5 to 4–4.5 km/s in a thin transition zone, which leads to three arrivals (triplication) from Layer 2A (Figure 5). However, since the first arrival tomography is performed on OBS data, spaced at 5–10 km distance, one can only get a smooth velocity of the crust. Therefore, a velocity contour of 4.5 km/s is generally used at the base of Layer 2A (Figure 4, *dashed curve*).

In seismic reflection data, the triplication from the Layer 2A/2B boundary lies in a narrow offset range at intermediate offsets, depending on the water depth and the thickness of Layer 2A (Figure 5) (Harding et al., 1993; Seher et al., 2010a). Since the cusp of triplication has strong amplitude, it can be used to image the base of Layer 2A, and leads to a very nice image of Layer 2A. Since the energy lies in a limited offset range, the two-way travel time for these images varies depending on the

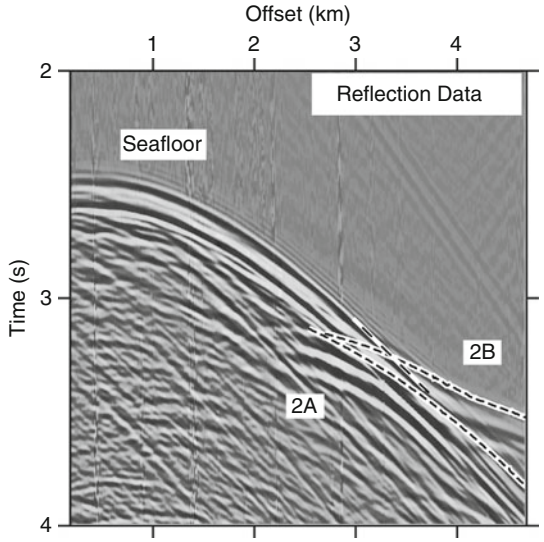


Crustal Reflectivity (Oceanic) and Magma Chamber, Figure 4 Average one-dimensional velocity of oceanic crustal layers and mantle (*solid brown line*). *Thin dashed line* indicates velocity determined using tomographic type techniques.

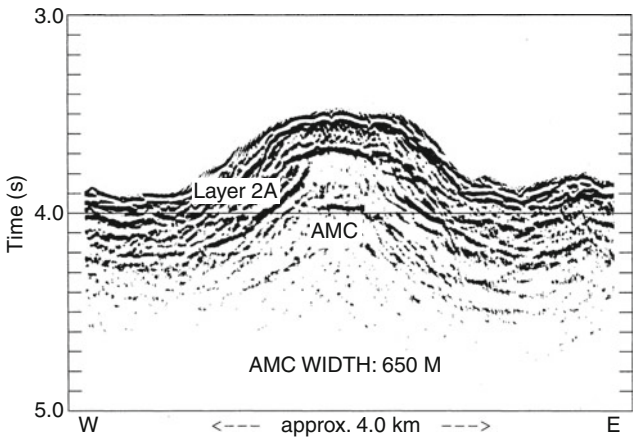
velocity used for stacking. This becomes particularly important when the thickness map needs to be prepared from 3D seismic reflection data or compared along long profiles. In order to objectively compare Layer 2A thickness, a constant velocity that provides the best image of Layer 2A is used to stack Layer 2A arrival and the two-way travel time is converted into depth using the tomographic velocity obtained using either OBS data or streamer tomography (Seher et al., 2010a).

Plot triplication

Figure 6 shows the image of Layer 2A obtained along the southern East Pacific Rise (Kent et al., 1994). The thickness of the layer varies from ~200 m on the ridge axis



Crustal Reflectivity (Oceanic) and Magma Chamber, Figure 5 Seismic reflection from the Lucky Strike Segment of Mid-Atlantic Ridge showing triplication (*dashed curve*) from Layer 2A/2B boundary (modified from Seher et al., 2010a).



Crustal Reflectivity (Oceanic) and Magma Chamber, Figure 6 Seismic reflection image of Layer 2A and axial magma chamber at 14° S East Pacific Rise (Position 3 in Figure 2) (from Kent et al., 1994). The Layer 2A thickness varies from 200 m on the ridge axis to 400 m off-axis.

and increases to 600 m off-axis (Kent et al., 1994; Harding et al., 1993). The thickening of Layer 2A off-axis could be associated with the accumulation of lavas away from the ridge axis. On slow spreading ridges, Layer 2A could be up to 1 km thick (Singh et al., 2006a; Seher et al., 2010a). Instead of thickening of Layer 2A away from the spreading center, its thickness decreases, which could be due to thinning by stretching (faulting) of the crust (Figure 7). In fact, the thinnest (300 m) Layer 2A at the Lucky Strike segment is observed near the Median bounding faults (Seher et al., 2010a).

There is still debate about the causes of Layer 2A reflection: it could be due to pillow lava and dike boundary as shown in Figure 4 or due to alteration front associated with hydrothermal circulation or pressure collapse boundary (Christeson et al., 2007).

Below Layer 2A, the velocity increases from 4.5 up to 6 km/s, corresponding to dike sequence, below which a magma chamber reflection might be present. The thickness of the dike sequence is ~1.5 km on a fast spreading center (Vera et al., 1990) and up to 2 km on a slow spreading center (Seher et al., 2010b).

Axial magma chamber (melt lens)

On a fast spreading center, the axial magma chamber (AMC) reflector marks the boundary between the upper crust (lava and dikes) and the lower crust (gabbro). It is observed along a significant part of the fast spreading axis at 1.5–2 km below the seafloor. The width of the AMC varies from a few hundred meters to 4.5 km (Kent et al., 1993). The largest melt lens is observed beneath 9° N overlapping spreading center at the East Pacific Rise (Kent et al., 2000; Singh et al., 2006a) (Figure 8). Recently, Canales et al. (2009) have imaged an off-axis melt lens. A 3D seismic reflection study of the 9° N EPR suggests that there might be an extensive presence of melt sills off-axis (Carton et al., 2009).

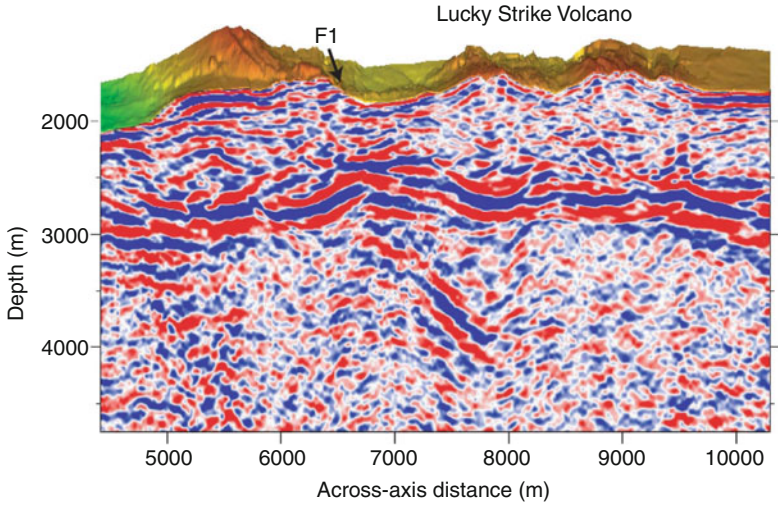
The AMC have also been observed beneath intermediate spreading centers such as Juan de Fuca Ridge (Canales et al., 2006) and Valu Fa Ridge (Collier and Sinha, 1990). They are generally observed about 3 km below the seafloor. It has been difficult to image AMC reflection on slow spreading ridges, which has been due to strong scattering on the seafloor and complex 3D bathymetry. Using 3D seismic reflection technique, Singh et al. (2006b) discovered AMC beneath the Lucky Strike segment of the Mid-Atlantic Ridge (Figure 9). The AMC is about 3 km wide and 7 km long at ~3 km below the seafloor.

The thickness of the melt lens is difficult to determine. Forward modeling of seismic reflection data from 9° N at EPR suggests that it could be 50–100 m thick (Kent et al., 1993). Seismic full waveform inversion of data from two locations at the EPR suggests that it should be 50–60 m thick (Collier and Singh, 1997; Singh et al., 1998, 1999). Using full waveform of seismic reflection data from 14° S at EPR, Singh et al. (1999) found that a 60 m thick melt lens is overlain by a 60 m thick solid roof and underlain by a 200 m thick solid floor. Above the roof layer, they find a 200 km thick low velocity layer, which they associate with the presence of hydrothermal circulation (Singh et al., 1999) (Figure 10). The roof layer could be the transition zone between the hot melt (1,200°C) below and the hydrothermal layer (400°C) above.

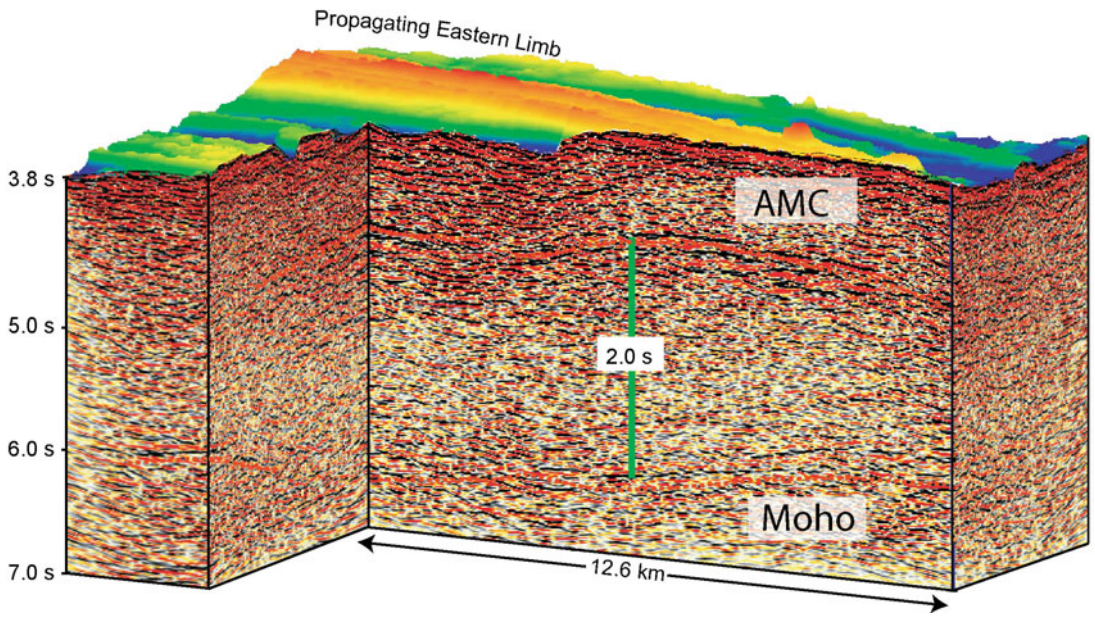
The presence of a solid floor suggests that magma cools and crystallizes in the melt lens and forms the solid floor. Using 3D seismic reflection data, we can image the base of the melt lens. Here the melt lens is about 4.5 km wide and could be up to 250 m thick (as compared to

50–100 m observed elsewhere). This could be due to its very large size (4.5 km) and the presence of low velocity underneath it (Bazin et al., 2003). Singh et al. (2006a) suggest that up to 40% melt could be present between 1.5 and 6 km depth, making it the largest observed melt lens in the crust observed on the earth so far (Figure 11).

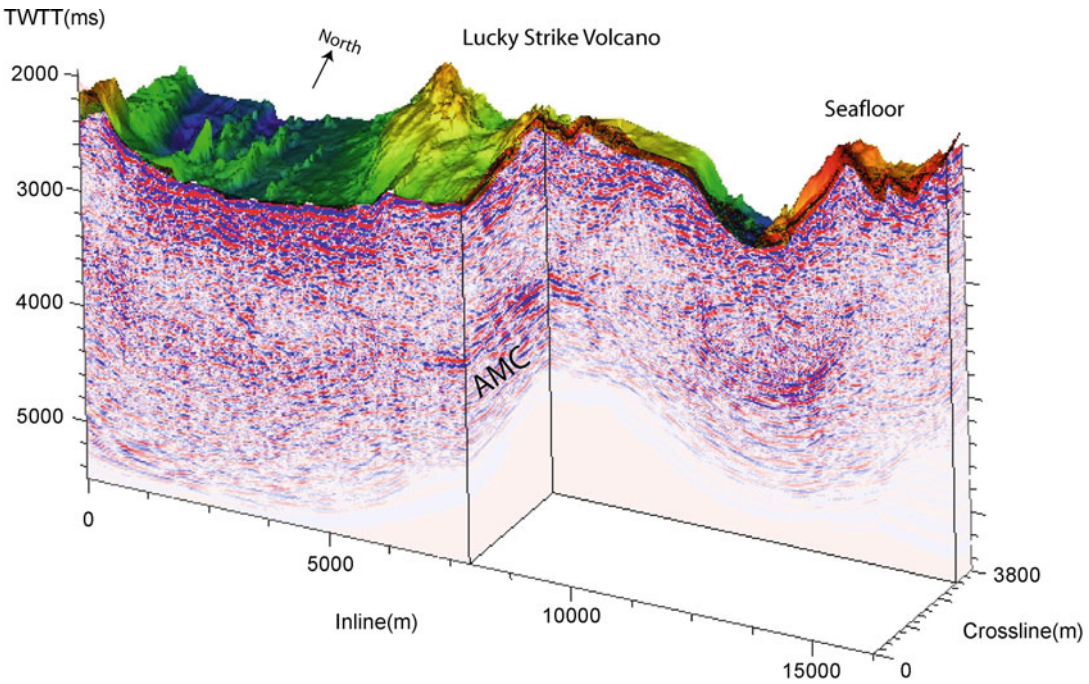
Using full waveform and partial stacking technique of seismic reflection data, Singh et al. (1998) show that the melt along 14° S EPR consists of 2–4 km long pure melt zones at 15–20 km interval along the ridge axis (Figure 10). They associate pure melt region where the shear wave velocity is nearly zero and partial melt zone where the shear wave velocity is nonzero (Taylor and Singh, 2002). They suggest that a pure melt region corresponds to fresh supply of magma from the mantle and a partial melt region corresponds to cooled and crystal-



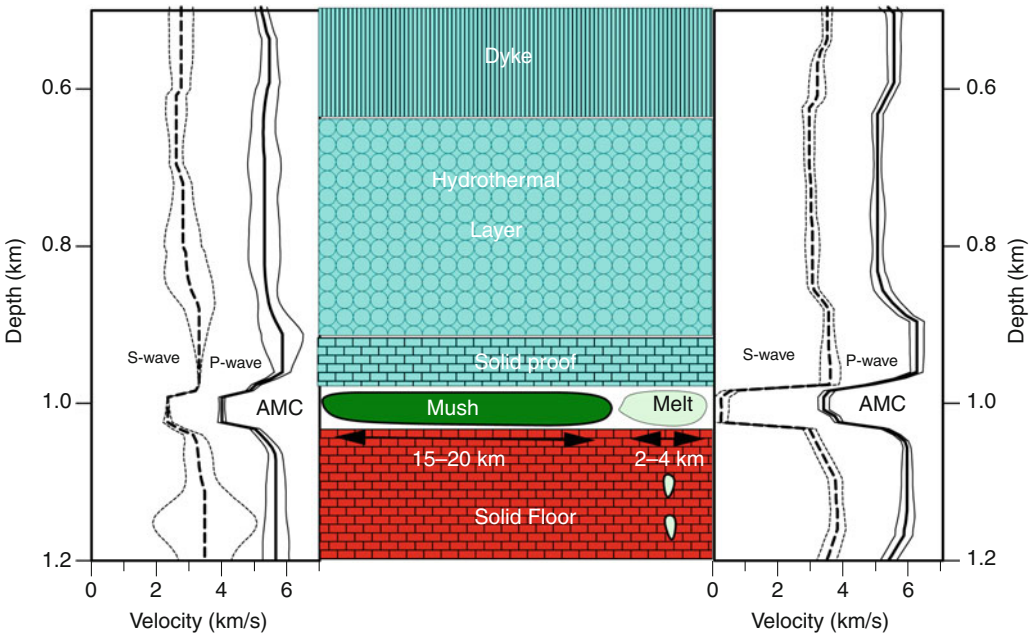
Crustal Reflectivity (Oceanic) and Magma Chamber, Figure 7 Seismic reflection of Layer 2A beneath the Lucky Strike volcano at slow spreading Mid-Atlantic Ridge (Position 1 in Figure 2). F1 indicates an active fault. Layer 2A is thick beneath the ridges and thin beneath the fault.



Crustal Reflectivity (Oceanic) and Magma Chamber, Figure 8 Seismic reflection image of axial magma chamber (AMC) and Moho reflection at 9° N overlapping spreading center at the East Pacific Rise (Position 2 in Figure 2) (modified from Singh et al., 2006a).



Crustal Reflectivity (Oceanic) and Magma Chamber, Figure 9 Axial magma chamber reflection image from the Lucky Strike Segment of the Mid-Atlantic Ridge (Position 1 in Figure 2).



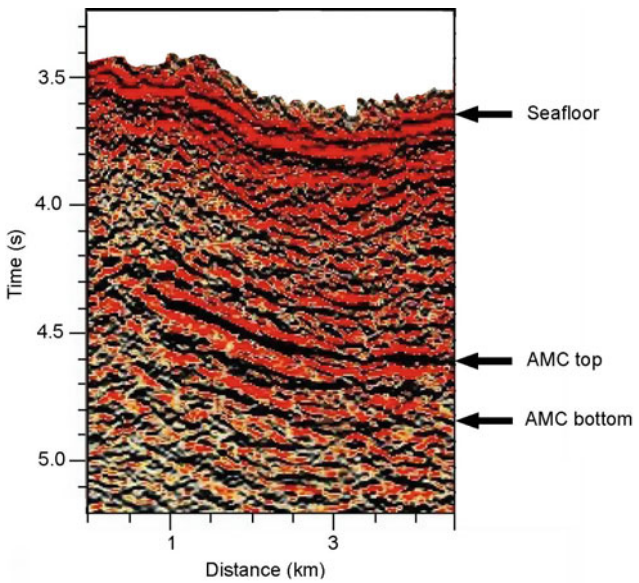
Crustal Reflectivity (Oceanic) and Magma Chamber, Figure 10 Full waveform inversion results at two different locations along the 14° S EPR (Position 3 in Figure 2). Inversion results show the presence of 60 m AMC, 60 m thick roof, and 200 m thick floor of the AMC. It also shows a hydrothermal layer above the roof of the AMC. The result on the *left panel* corresponds to a partial melt region (less than 50% of liquid melt) whereas that on the *right* corresponds to a pure melt (more than 80% of liquid) region. The depth is from the seafloor (modified from Singh et al., 1999).

lized state of melt lens. They also find that a pure melt region is associated with hydrothermal activities on the seafloor, linking the supply of magma from the mantle with hydrothermal circulation on the seafloor. In order for the melt lens to be in steady state, it would require a supply of magma from the mantle every 30 years as it will take 50 years to completely solidify a 50 m thick melt lens (Singh et al., 1999).

Lower crust

So far, no other crustal melt lenses have been observed beneath the axial melt lens. Therefore, only tomographic methods using OBS data are used to determine the velocity structure of the lower crust. Using two-ship expanding spread profile, Vera et al. (1990) and Harding et al. (1989) found that the velocity in the lower crust (5.5–6 km/s) on-axis was lower than that off-axis (6.8–7 km/s), which suggested the presence of partial melt in the lower crust. Using 3D tomography technique, Toomey et al. (1990) and Dunn et al. (2000) showed the presence of low velocity in the lower crust, suggesting the presence of partial melt. Singh et al. (2006b) found a large anomaly in two-way travel time between the Moho reflection and AMC and suggest that up to 40% of melt might be present in the lower crust (Figure 8). A low velocity anomaly has also been observed beneath the Lucky Strike segment of the Mid-Atlantic Ridge beneath the melt lens (Seher et al., 2010c; Singh et al., 2006b). These results suggest that partial melt is present in the lower crust beneath the melt lens (Figure 12).

Based on the study of Oman ophiolite, Kelemen et al. (1997) and Boudier et al. (1996) suggested that the lower crust is formed by injection of melt sills in the lower crust instead of cooling and crystallization of magma in the upper crustal melt sill. However, no melt lenses have been imaged in the lower crust so far, even using 3D seismic reflection technique (Singh et al., 2006a), which suggest that the melt in the lower crust must be in small pockets, not in large melt sills.



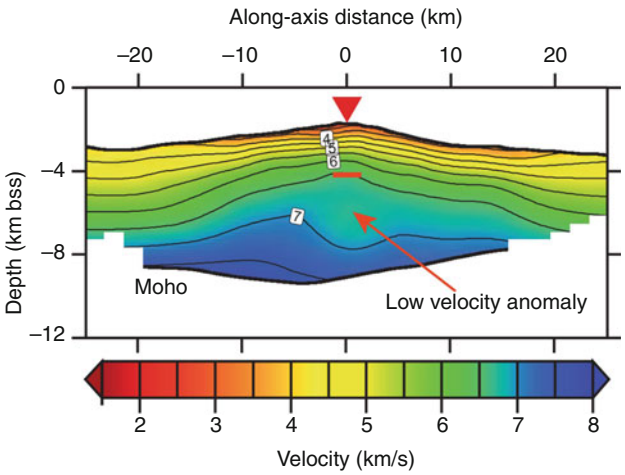
Crustal Reflectivity (Oceanic) and Magma Chamber, Figure 11 Seismic reflection image showing the top and bottom of the axial magma chamber at 9° N EPR (Position 2 in Figure 2).

Oceanic Moho

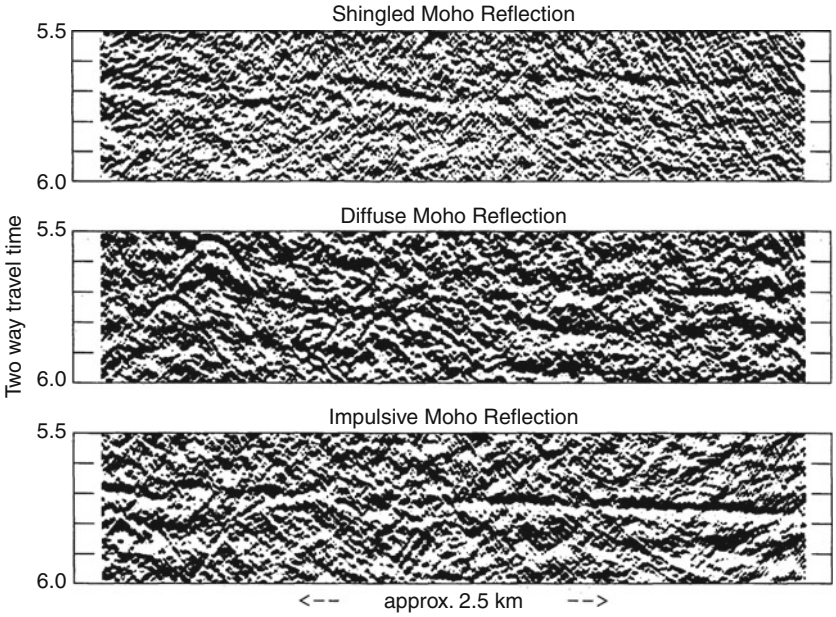
The Moho is a boundary between the crust and mantle, where the velocity changes from 6.8–7.0 km/s to 7.8–8.1 km/s. For a sharp boundary, a clear Moho reflection is observed. However, the Moho is not just a simple boundary. Kent et al. (1994) identified three types of Moho reflection: (1) Impulsive Moho where a clear single reflection is observed (Figure 13), (2) Diffuse Moho where reflection from Moho is patchy, and (3) Shingled Moho where reflection is shingled.

Moho reflections are generally observed away from the spreading center, not beneath the melt lens. This is because the upper crustal melt lens and associated lower crustal melt would attenuate seismic energy and hence, it would be difficult to image Moho using conventional seismic reflection techniques. Secondly, it was accepted that Moho is formed away from the ridge axis. However, using 3D seismic reflection technique, Singh et al. (2006a) have imaged Moho beneath the wide axial melt lens (Figure 14), suggesting the Moho is formed at zero age.

Moho reflections are observed ~2 s two-way travel time below the seafloor. The average crustal velocity is about 6 km/s, and hence the average crustal thickness is about 6 km. However, Singh et al. (2011) have observed Moho reflection 1.3 s below the seafloor offshore NW



Crustal Reflectivity (Oceanic) and Magma Chamber, Figure 12 Tomographic results from the Lucky Strike Segment of the Mid-Atlantic Ridge (Position 1 in Figure 2) showing melt lens (*red line*), low velocity in the lower crust, and Moho across the ridge axis. The inverse triangle is the center of the ridge axis (modified from Seher et al., 2010c).



Crustal Reflectivity (Oceanic) and Magma Chamber, Figure 13 Different types of reflection for Moho boundary observed at Southern East Pacific Rise (Position 3 in Figure 2) (Kent et al., 1994). *Top*: Shingled Moho, *Middle*: Diffuse Moho, *Bottom*: Impulsive Moho.

Sumatra, where the crust was formed at the fast spreading Wharton ridge about 55 Ma ago. If we assume an average velocity of 6.0 km/s, which means at the crust there is only 3.5 km thick, it would be the thinnest crust ever observed in a fast spreading environment (Figure 15).

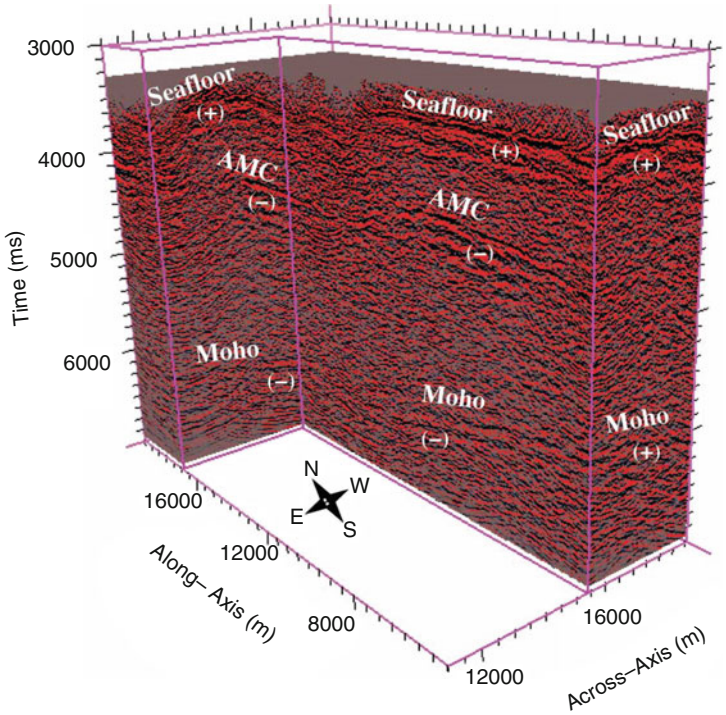
Since there is a positive velocity contrast at the crust–mantle boundary, large amplitude critical angle reflections (PmP) are generated that arrive at large offsets (15–30 km), and are recorded on OBS. These data are then used to map the Moho structure. However, there are two problems in using arrival times of PmP arrivals. The velocity gradient in the lower crust is very small or close to zero, and therefore rays that sample the lower crust do not turn in the lower crust making it difficult to estimate velocity in this zone. Since the arrival time of PmP arrival depends on the velocity in the crust, particularly the lower 3 km of the crust, it is difficult to estimate both the velocity in the lower crust and the Moho structure. Since the velocity in the lower crust can be approximated to a large extent ~6.8 km/s, these data provide a reasonable constrain on the Moho structure, and are routinely used. Secondly, the OBS spacing is generally 5–20 km and the lateral resolution of the Moho structure is very poor.

Crustal thickness

Once we determine the velocity in the crust and Moho structure, we can estimate the crustal thickness. So far, most of the crustal thickness estimations are based on wide-angle seismic studies, and there are significant uncertainties in crustal thickness estimations for the reason explained above. However, combined analyses of seismic reflection, refraction, and gravity data suggest that wide-angle estimation of the crustal thickness can be very reliable.

The crust formed at fast spreading centers is generally uniform and 6 km thick (White et al., 1992; Eittreim et al., 1994) but some variations (5–8 km) have been observed recently (Canales et al., 2003; Singh et al., 2006a; Barth and Mutter, 1996). There is no significant crustal thickness variation across the fracture zone in a fast spreading environment, where it is also 5.5–6 km thick (van Avendonk et al., 2001). The crust formed at slow spreading centers is generally thick at the center of the segment (6–7 km) and thin at segment ends (4–5 km) (Barclay et al., 1998; Seher et al., 2010b). The crust beneath fracture zones in a slow spreading environment is generally thin (3–4.5 km) (Detrick et al., 1993). The thinnest crust (3.5–4.5 km) is formed at ultraslow spreading centers (Jokat and Schmidt-Aursch, 2007). These observations suggest that the thickness of the crust depends on the spreading rate and melt distribution in the crust. On the other hand, thicker crusts are found beneath large igneous provinces, across hotspot tracks or at the interaction of plume and ridges, which are believed to be formed by higher mantle temperatures due to the presence of a plume (e.g., Parkin and White, 2008; Watts et al., 1985; Charvis et al., 1999; Grevemeyer et al., 2001). For example, the crust beneath Iceland could be up to 30 km thick, and beneath La Reunion Island it is 13 km (Charvis et al., 1999).

However, anomalously thin crust has been reported. A 5 km thick crust is observed near the South American



Crustal Reflectivity (Oceanic) and Magma Chamber, Figure 14 Seismic reflection image of Moho beneath the melt lens at the 9° N East Pacific Rise (Position 2 in Figure 2).

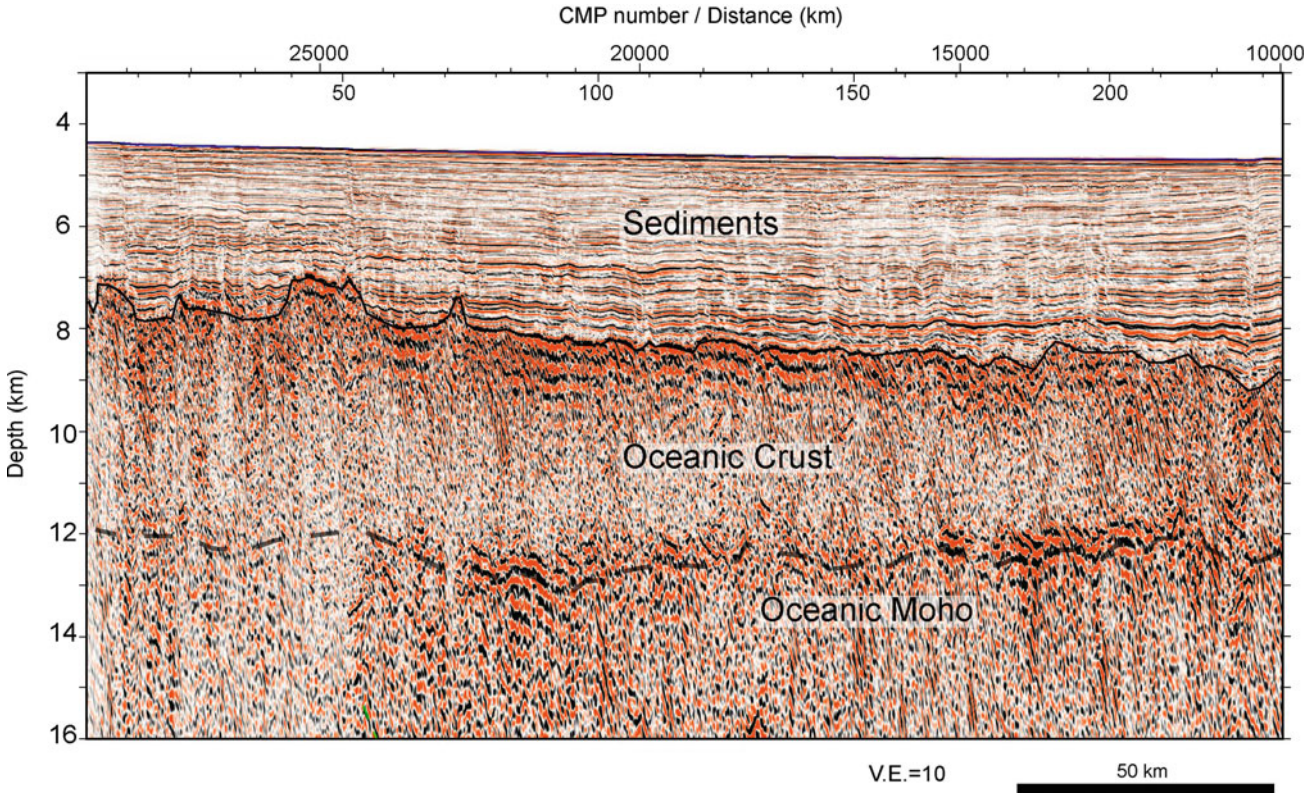
trench (Grevemeyer et al., 2007) and in the area of IODP Hole 1256 (Hallenborg et al., 2003) for crusts formed at the EPR about 20 Ma. Rodger et al. (2006) found an ultrathin crust (4 km) that was formed near the central MAR about 85 Ma ago, which they associate to be due to a reduction in the spreading rate from slow to ultraslow. The crustal study at ODP holes 504B also found thin crust (5 km), which Collins et al. (1989) associate to be due to depleted mantle. Singh et al. (2011) have observed ultrathin crust (3.5–4.5 km) in the Indian Ocean (Figure 15) that was formed at fast spreading Wharton center 55 Ma ago, which they associate to be due to the interaction of the Kerguelen plume with the Wharton Spreading center due to channeling of the cold lithosphere around a plume and its interaction with the spreading center.

Upper mantle

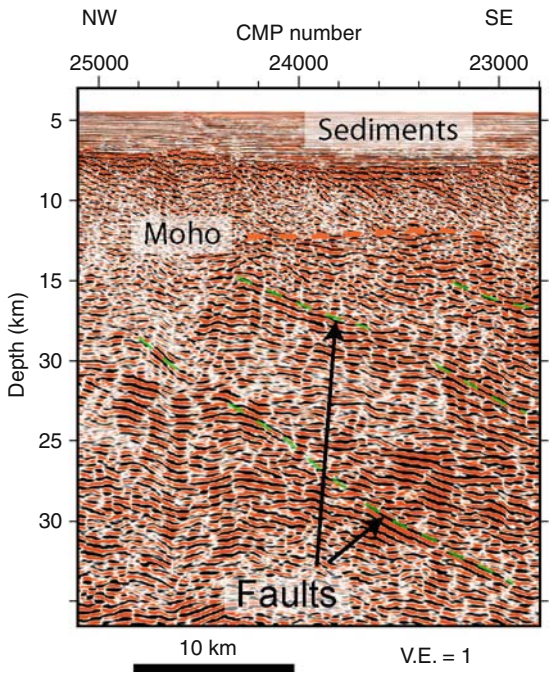
The velocity in the upper mantle is generally determined using rays that turn below the Moho, which are called Pn arrivals. The velocity in the mantle beneath ridge axis is ~7.6–7.8 km/s and is 8–8.1 km/s away from the ridge axis (Vera et al., 1990; Harding et al., 1989). An extensive 3D tomography and undershoot experiment were carried out at 9° N EPR. Using these data, Dunn et al. (2000, 2001) obtained a symmetric low velocity anomaly below the Moho in the mantle. Using the same data and some new data Toomey et al. (2007) found asymmetric low velocity anomaly at 10 km away from the ridge axis. There are serious problems with these results. First, the crustal thickness in this area varies from 6 to 8 km. Secondly, the velocity anomaly depends on the background velocity used during the inversion. Dunn et al. (2000) use a starting mantle velocity of 8.2 km/s whereas one year later Dunn et al. (2001) use 7.6 km/s. Although the crustal thickness is up to 8 km (Canales et al., 2003), Dunn et al. (2001) interpret velocity at 7 km below the seafloor; the part of the anomaly could be due to crustal variations. Toomey et al. (2007) show that the ray only penetrates down to 8 km below the seafloor but interpret the velocity at 9 km below the seafloor. These conflicting results and the inaccuracy in inversion led Singh and Macdonald (2009) to suggest that these results are not reliable. The only robust solution we have is that the velocity in the mantle is ~7.6–7.8 km/s below the ridge axis and 8–8.1 km/s away from the ridge axis (Vera et al., 1990; Harding et al., 1989). However, it is possible that melt sills get trapped in the mantle as shown by Nedimovic et al. (2005) and may lead to a low velocity anomaly in the mantle.

Faults in the oceanic crust and upper mantle

Along with magmatic process, tectonic processes also play an important role in shaping the oceanic crust. Singh et al. (2006b) have shown that median valley bounding faults could be imaged down to 3 km, close to the melt lens. They also found extensive faults above the melt lens. It is possible that some of these faults penetrate down to the crust–mantle boundary (Dusunur et al., 2009).



Crustal Reflectivity (Oceanic) and Magma Chamber, Figure 15 Seismic reflection image of extremely thin crust in the Indian Ocean (Position 4 in Figure 2) (modified from Singh et al., 2011).



Crustal Reflectivity (Oceanic) and Magma Chamber, Figure 16 Seismic reflection of deep penetrating faults in the mantle in the Indian Ocean (Position 4 in Figure 2).

In a subduction zone environment, there are two forces that act on the incoming oceanic plate: plate bending leading to normal faulting and compressive forces leading to thrust faulting. Water can enter into the mantle leading to serpentinization of the oceanic mantle. Normal faulting down to 5–8 km below the Moho, related to bending of the subducting plate and the associated serpentinization, has been reported for the Central American trench (Ranero et al., 2003). Faults have also been imaged in the mantle offshore Sumatra down to 35 km depths (Figure 16) suggesting that a significant part of the oceanic lithosphere is involved during oceanic earthquakes.

Summary

The reflectivity of the oceanic crust and axial magma chambers at ocean spreading centers are described in detail using real seismic images. The oceanic crust consists of three distinct layers: Layer A is 200–1,000 m thick and consists mainly of pillow lavas. Layer B is 1–2 km thick and consists mainly of cooled basalts in dikes. Layer 3 is 2–3 km thick and is formed by the cooling and crystallization of magma in the melt lens. Axial melt lens has been imaged on fast and intermediate spreading centers and recently on a slow spreading center (Singh et al., 2006b). Melt lenses are 50–60 m and have a roof and a floor. Between the melt and Moho, lower crust,

partial melt is generally present. The oceanic crust is about 6 km thick, which could vary from 3.5 to 8 km. The P-wave velocity in the mantle lie between 7.6 and 8.1 km/s, but the detailed nature of the oceanic mantle is poorly constrained and requires further investigations. Faults due to earthquakes have been observed down to 35 km depth.

Bibliography

Barclay, A. H., Toomey, D. R., and Solomon, S. C., 1998. Seismic structure and crustal magmatism at the Mid-Atlantic Ridge, 35° N. *Journal of Geophysical Research*, **103**, 17827–17844.

Barth, G. A., and Mutter, J. C., 1996. Variability in oceanic crustal thickness and structure: multichannel seismic reflection results from the northern East Pacific Rise. *Journal of Geophysical Research*, **101**, 17951–17975.

Bazin, S., Harding, A. J., Kent, G. M., Orcutt, J. A., Singh, S. C., Tong, C. H., Pye, J. W., Barton, P. J., Sinha, M. C., White, R. S., Hobbs, R. W., and van Avendonk, H. J. A., 2003. A three-dimensional study of axial low velocity beneath the 9° 03′ N overlapping spreading center. *Geophysical Research Letters*, **30**, 1–4.

Boudier, F., Nicolas, A., and Ildefonse, B., 1996. Magma chambers in the Oman ophiolite: fed from the top and the bottom. *Earth and Planetary Science Letters*, **144**, 238–250.

Canales, J.-P., Detrick, R. S., Toomey, D. R., and Wilcock, W. S. D., 2003. Segment-scale variations in the crustal structure of 150–300 kyr old fast spreading oceanic crust (East Pacific Rise, 8° 15′ N–10° 5′ N) from wide-angle seismic refraction profiles. *Geophysical Journal International*, **152**, 766–794.

Canales, J.-P., Singh, S. C., Detrick, R. S., Carbotte, S. M., Harding, A. J., Kent, G. M., Diebold, J. B., Babcock, J., and Nedimovic, M. R., 2006. Seismic evidence for variations in axial magma chamber properties along the southern Juan de Fuca Ridge. *Earth and Planetary Science Letters*, **246**, 353–366.

Canales, J. P., Nedimovic, M. R., Kent, G. M., Carbotte, S. M., and Detrick, R. S., 2009. Seismic reflection images of a lear-axis melt sill within the lower crust at the Juan de Fuca Ridge. *Nature*, **460**, 89–94.

Carton, H. D., Carbotte, S. M., Mutter, J. C., Canales, J., Nedimovic, M. R., Marjanovic, M., Aghaei, O., Xu, M., Han, S., Stowe, L., 2009. Characteristics of the crustal magma body in the 2005–2006 eruption area at 9° 50′ N on the East Pacific Rise from 3D multi-channel seismic data. *AGU*, Fall meeting, Abstract OS11B-02.

Charvis, P., Laesanpura, A., Gallart, J., Hirn, A., Lepine, J., de Voogd, B., Minshull, T. A., Hello, Y., and Pontoise, B., 1999. Spatial distribution of hotspot material added to the lithosphere under La Reunion, from wide-angle seismic data. *Journal of Geophysical Research*, **104**, 2875–2893.

Christeson, G., McIntosh, K., and Karson, J., 2007. Inconsistency in correlation of layer 2A and lava layer thickness in oceanic crust. *Nature*, **445**, 418–421.

Collier, J. S., and Singh, S. C., 1997. Detailed structure of the top of the melt body beneath the East Pacific Rise at 9°40′ N from waveform inversion of seismic reflection data. *Journal of Geophysical Research*, **102**, 20287–20304.

Collier, J. S., and Sinha, M., 1990. Seismic images of a magma chamber beneath the Lau Basin back-arc spreading centre. *Nature*, **346**, 646–648.

Collins, J. A., Purdy, G. M., and Brocher, T. M., 1989. Seismic velocity structure at Deep Sea Drilling Site 504B, Panama Basin: evidence for thin oceanic crust. *Journal of Geophysical Research*, **94**, 9283–9302.

Detrick, R. S., Buhl, P., Vera, E., Mutter, J., Orcutt, J., Madsen, J., and Brocher, T., 1987. Multi-channel seismic imaging of a crustal magma chamber along the East Pacific Rise. *Nature*, **326**, 35–41.

Detrick, R. S., White, R. S., and Purdy, G. M., 1993. Crustal structure of North Atlantic Fracture Zone. *Reviews of Geophysics*, **31**, 439–458.

Dunn, R. A., Toomey, D. R., and Solomon, S. C., 2000. Three-dimensional seismic structure and physical properties of the crust and shallow mantle beneath the East Pacific Rise 9° 30′ N. *Journal of Geophysical Research*, **105**, 23537–23555.

Dunn, R. A., Toomey, D. R., Detrick, R. S., and Wilcock, W. S. D., 2001. Continuous mantle melt supply beneath an overlapping spreading centre on the East Pacific Rise. *Science*, **291**, 1955–1958.

Dusunur, D., Seher, T., Combier, V., Escartin, J., Crawford, W., Cannat, M., Singh, S. C., Matias, L. M., and Miranda, J. M., 2009. Seismological constraints on the thermal structure and interaction of tectonic and magmatic processes around the Lucky Strike magma chamber, slow spreading Mid-Atlantic Ridge. *Marine Geophysical Research*, **30**, 105–120.

Eittreim, S. L., Gnibidenko, H., Helsley, C. E., Sliter, R., Mann, D., and Ragozin, N., 1994. Oceanic crustal thickness and seismic character along a central Pacific transect. *Journal of Geophysical Research*, **99**, 3139–3145.

Grevemeyer, I., Flueh, E. R., Reichert, C., Bialas, J., Klaschen, D., and Kopp, C., 2001. Crustal architecture and deep structure of the Ninetyeast Ridge hotspot trail from active-source ocean bottom seismology. *Geophysical Journal International*, **144**, 414–431.

Grevemeyer, I., Ranero, C. R., Flueh, E. R., Klaschen, D., and Bialas, J., 2007. Passive and active seismological study of bending-related faulting and mantle serpentinisation at the Middle American trench. *Earth and Planetary Science Letters*, **258**, 528–542.

Hallenborg, E., Harding, A. J., and Kent, G. M., 2003. Seismic structure of 15 Ma oceanic crust formed at an ultra-fast spreading East Pacific Rise: evidence for kilometer-scale fracturing from dipping reflectors. *Journal of Geophysical Research*, **108**, 2532, doi:10.1029/2003JB002400.

Harding, A. J., Kappus, M. E., Orcutt, J. A., Vera, E. E., Buhl, P., Mutter, J. C., Detrick, R., and Brocher, T., 1989. The structure of young oceanic crust 13° N on the East Pacific Rise from expanding spread profiles. *Journal of Geophysical Research*, **94**, 12163–12196.

Harding, A. J., Kent, G. M., and Orcutt, A. J., 1993. A multichannel seismic investigation of upper crustal structure at 9° N on the East Pacific Rise: implications for crustal accretion. *Journal of Geophysical Research*, **98**, 13925–13944.

Jokat, W., and Schmidt-Aursch, M. C., 2007. Geophysical characteristics of the ultra-slow spreading Gakkel Ridge, Arctic Ocean. *Geophysical Journal International*, **168**, 983–998.

Kelemen, P. B., Koga, K., and Shimizu, N., 1997. Geochemistry of gabbro sills in the crust-mantle transition zone of the Oman ophiolite: implications for the origin of the oceanic lower crust. *Earth and Planetary Science Letters*, **146**, 475–488.

Kent, G. M., Harding, A. J., and Orcutt, J. A., 1993. Distribution of magma beneath the East Pacific Rise between the Clipperton Transform and the 9° 17′ N Deval from forward modelling of common depth point data. *Journal of Geophysical Research*, **98**, 13945–13969.

Kent, G. M., Harding, A. J., Orcutt, J. A., Detrick, R. S., Mutter, J. C., and Buhl, P., 1994. Uniform accretion of oceanic crust south of the Garrett transform at 14° 15′ S on the East Pacific Rise. *Journal of Geophysical Research*, **99**, 9097–9116.

Kent, G. M., Singh, S. C., Harding, A. J., Sinha, M. C., Tong, V., Barton, P. J., Hobbs, R., White, R., Bazin, S., and Pye, J.,

2000. Evidence from three-dimensional reflectivity images for enhanced melt supply beneath mid-ocean-ridge discontinuities. *Nature*, **406**, 614–618.
Nedimovic, M. R., Carbotte, S. M., Harding, A. J., Detrick, R. S., Canales, J. P., Diebold, J. B., Kent, G. M., Tischer, M., and Babcock, J. M., 2005. Frozen magma lenses below the oceanic crust. *Nature*, **436**, 1149–1152.
Parkin, C. J., and White, R. S., 2008. Influence of the Iceland mantle plume on the oceanic crustal generation in the North Atlantic. *Geophysical Journal International*, **173**, 168–188.
Ranero, C. R., Reston, T. J., Belykh, I., and Gnibidenko, H., 1997. Reflective oceanic crust formed at a fast spreading centre in the Pacific. *Geology*, **25**, 499–502.
Ranero, C. R., Phipps Morgan, J., McIntosh, K., and Reichert, C., 2003. Bending-related faulting and mantle serpentinization at the Middle America trench. *Nature*, **425**, 367–373.
Rodger, M., Watts, A. B., Greenroyd, C. J., Peirce, C., and Hobbs, R. W., 2006. Evidence for unusually thin oceanic crust and strong mantle beneath the Amazon Fan. *Geology*, **34**, 1081–1084.
Seher, T., Crawford, W., Singh, S. C., and Cannat, M., 2010a. Seismic layer 2A variations in the Lucky Strike segment at the Mid-Atlantic Ridge from reflection measurements. *Journal of Geophysical Research*, **115**, B07107, doi:10.1029/2009JB006783.
Seher, T., Singh, S. C., Crawford, W., and Escartin, J., 2010b. Upper crustal velocity structure beneath the central Lucky Strike Segment from seismic refraction measurements. *Geochemistry, Geophysics, Geosystems*, **11**, 5, doi:10.1029/2009GC002894.
Seher, T., Crawford, W., Singh, S. C., Cannat, M., Combier, V., Dusunur, D., and Canales, J.-P., 2010c. Crustal velocity structure of the Lucky Strike segment of the Mid-Atlantic Ridge (37°N) from seismic refraction measurements. *Journal of Geophysical Research*, **115**, B03103, doi:10.1029/2009JB006650.
Singh, S. C., and Macdonald, K., 2009. Mantle skewness and ridge segmentation. *Nature*, **458**, E11–E12.
Singh, S. C., Midenet, S., and Djajadihardja, Y., 2009. Sesimic evidence of the locked and unlocked Sumatra subdction zone. EOS, **90**, 471–472.
Singh, S. C., Kent, G. M., Collier, J. S., Harding, A. J., and Orcutt, J. A., 1998. Melt to mush variations in crustal magma properties along the ridge crest at the southern East Pacific Rise. *Nature*, **394**, 874–878.
Singh, S. C., Collier, J. S., Kent, G. M. J. S., Harding, A. J., and Orcutt, J. A., 1999. Seismic evidence for a hydrothermal layer above the solid roof of axial magma chamber at the southern East Pacific Rise. *Geology*, **27**, 219–222.
Singh, S. C., Harding, A., Kent, G., Sinha, M. C., Combier, V., Hobbs, R., Barton, P., White, R., Tong, V., Pye, J., and Orcutt, J. A., 2006a. Seismic reflection images of Moho underlying melt sills at the East Pacific Rise. *Nature*, **442**, 287–290.
Singh, S. C., Crawford, W., Carton, H., Seher, T., Combier, V., Cannat, M., Canales, J., Dusunur, D., Escartin, J., and Miranda, M., 2006b. Discovery of a magma chamber and faults beneath a hydrothermal field at the Mid-Atlantic Ridge. *Nature*, **442**, 1029–1033.
Singh, S. C., Carton, H., Chauhan, A., et al., 2011. Extremely thin crust in the Indian Ocean possibly resulting from Plume-Ridge interaction. *Geophysical Journal International*, doi:10.1111/j.1365-246X.2010.04823.x.
Taylor, M., and Singh, S. C., 2002. Composition and microstructure of magma bodies from effective medium theory. *Geophysical Journal International*, **149**, 15–21.
Toomey, D. R., Purdy, G. M., Solomon, S. C., and Wilcock, W. S. D., 1990. The three-dimensional seismic velocity structure of the East Pacific Rise near latitude 9°30′ N. *Nature*, **347**, 639–645.
Toomey, D. R., Jousselin, D., Dunn, R. A., Wilcock, W. S. D., and Detrick, R. S., 2007. Skew of mantle upwelling beneath the East Pacific Rise governs segmentation. *Nature*, **446**, 409–414.
Van Avendonk, H. J. A., Harding, A. J., Orcutt, J. A., and McClain, J. S., 2001. Contrast in crustal structure across the Clipperton transform fault from travel time tomography. *Journal of Geophysical Research*, **106**, 10961–10981.
Vera, E. E., Mutter, J. C., Buhl, P., Orcutt, J. A., Harding, A. J., Kappus, M. E., Detrick, R. S., and Brocher, T. M., 1990. The structure of 0 to 0.2 m.y. old oceanic crust at 9° N on the East Pacific Rise from expanding spread profiles. *Journal of Geophysical Research*, **95**, 15529–15556.
Watts, A. B., ten Brink, U. S., Buhl, P., and Brocher, T. M., 1985. A multi-channel seismic study of lithospheric flexure across the Hawaiian Emperor seamount chain. *Nature*, **315**, 105–111.
White, R. S., McKenzie, D., and O'Nions, R. K., 1992. Oceanic crustal thickness from seismic measurements and rare earth element inversions. *Journal of Geophysical Research*, **97**, 19683–19715.

Cross-references

Deep Seismic Reflection and Refraction Profiling
Earth's Structure, Continental Crust
Lithosphere, Oceanic
Ocean Bottom Seismics
Ocean, Spreading Centre
Seismic Data Acquisition and Processing
Seismic Structure at Mid-Ocean Ridges
Single and Multichannel Seismics

CURIE TEMPERATURE

Vincenzo Pasquale
Dipartimento per lo Studio del Territorio e delle sue Risorse, Settore di Geofisica, Università degli Studi di Genova, Genova, Italy

Synonyms

Curie point

The critical temperature above which any material loses its permanent magnetism.

A rock containing magnetic minerals loses its permanent magnetism when heated up to reach a critical temperature. This is the Curie temperature or Curie point, ϑ_c, above which the rock is paramagnetic. On the atomic level, below ϑ_c, the magnetic moments are aligned in their respective domains and even a weak external field results in a net magnetization. As the temperature increases to ϑ_c and above, however, fluctuations due to the increase in thermal energy destroy that alignment.

Strongly magnetic minerals are all of ferrimagnetic type, in which neighboring magnetic moments are aligned antiparallel, as in antiferromagnetism, but unequal numbers or strengths give a net magnetization. Among them, magnetite, one of the members of the titanomagnetite solid-solution series, is the most important magnetic mineral to geophysical studies of crustal rocks (Stacey, 1992). There are many other magnetic minerals, but they are rare

(iron sulfide), unstable (maghemite) or having a weak spontaneous magnetization due to the canting of its nearly equal and opposite atomic moments (canted antiferromagnetism), as hematite occurring in sedimentary rocks often in solid solution with ilmenite.

Variation of Curie temperature of titanomagnetite, $Fe_{3-x}Ti_xO_4$, is approximately given by $\vartheta_c = 580\,(1-1.26\,x)$, where magnetite (with chemical composition $x = 0$) and ulvöspinel ($x = 1$) are end members. In pure magnetite, ϑ_c is 580°C, but titaniferous inclusions can reduce ϑ_c, which approximates room temperature for $x = 0.77$ (Hunt et al., 1995). Saturation magnetization is thus a function of temperature, and it disappears above ϑ_c; at room temperature it decreases from 92 A m^2 kg^{-1} for $x = 0$ to about zero for $x = 0.8$. ϑ_c increases with pressure, but the change is quite small. At the crust–mantle boundary, it should not be more than a few degrees of the experimental values at normal pressure.

The crustal rocks lose their magnetization when they are at a depth where the temperature is greater than ϑ_c and their ability to generate detectable magnetic anomalies disappears. A comparison of the crustal temperature with ϑ_c may give information on the depth of compositional change where magnetic rocks are replaced with nonmagnetic material. Transforming regional magnetic anomalies data sets into the Fourier domain and analyzing their spectra allow to obtain information about deep magnetic discontinuities (Blakely, 1996). An example of deep magnetic discontinuities and their relation with Moho depth, temperature, and surface heat flux in central and southern Europe is given by Chiozzi et al. (2005) (Table 1).

The depth of the magnetic layer bottom corresponds to the Moho in the Variscan units and the expected temperature is close to ϑ_c of magnetite. Beneath the Alpine units, this depth is within the crust and its average temperature of 550°C indicates the presence of Ti content. In the Ligurian basin and the Tuscany-Latium geothermal area, the deep magnetic discontinuity sometimes lies a few kilometers below the Moho. This may indicate that the uppermost mantle also contributes to the geomagnetic field. The observed maximum temperature of 650°C, decidedly larger than ϑ_c of magnetite, is due to the fact that the local thermal anomalies are not discernible by the spectral analysis, which is a kind of low-pass filter depending on the window size of the magnetic data investigated.

Curie Temperature, Table 1 Crustal magnetic discontinuities with corresponding parameters

Area	Magnetic layer bottom depth (km)	Moho depth (km)	Temperature at magnetic layer bottom (°C)	Surface heat flux (mW m^{-2})
Variscan units (central Europe, Corsica–Sardinia block)	29–33	29–33	540–580	60–70
Alpine units (Alps, Apennines, Molasse and Po basins)	22–28	25–50	500–600	50–80
Ligurian basin, Tuscany–Latium geothermal area	17–28	17–20	500–650	80–120

Bibliography

Blakely, R. J., 1996. *Potential Theory in Gravity and Magnetic Applications*. Cambridge: Cambridge University Press.

Chiozzi, P., Matsushima, J., Okubo, Y., Pasquale, V., and Verdoya, M., 2005. Curie-point depth from spectral analysis of magnetic data in central-southern Europe. *Physics of the Earth and Planetary Interiors*, **152**, 267–276.

Hunt, C. P., Moskowitz, B. M., and Banerjee, S. K., 1995. Magnetic properties of rocks and minerals. In Ahens, T. J. (ed.), *Rock Physics and Phase Relations: A Handbook of Physical Constants*. Washington: American Geophysical Union, pp. 189–204.

Stacey, F. D., 1992. *Physics of the Earth*. Brisbane: Brookfield Press.

Cross-references

Earth's Structure, Continental Crust
Geomagnetic Field, Global Pattern
Geomagnetic Field, Theory
Heat Flow, Continental
Magnetic Domains
Remanent Magnetism

D

DEEP SCIENTIFIC DRILLING

Ulrich Harms[1], Harold J. Tobin[2]
[1]Scientific Drilling, GFZ German Research Centre for Geosciences, Potsdam, Germany
[2]Department of Geoscience, University of Wisconsin-Madison, Madison, WI, USA

Definition

Scientific drilling. Applying drilling methods developed in the hydrocarbon and mineral exploration industry to dig holes and retrieve samples of rock and fluid for scientific, not-for-profit purposes.

Deep scientific drilling. Drilling for scientific research goals beyond near-surface geological and shallow groundwater exploration to depths greater than several hundred meters subsurface, through land, sea ice, or below the sea/lake floor.

Introduction

Geoscientific insight into the deep Earth is mostly based on observations from the surface, indirect evidence through deep geophysical investigations or, increasingly, modeling. However, several critical Earth system processes and records of environmental and paleoclimate change are difficult to observe from ground level only. Hence, despite the great progress achieved in solid Earth science of the past decades, truly ground-truthed knowledge of the dynamics of Earth's crust is limited.

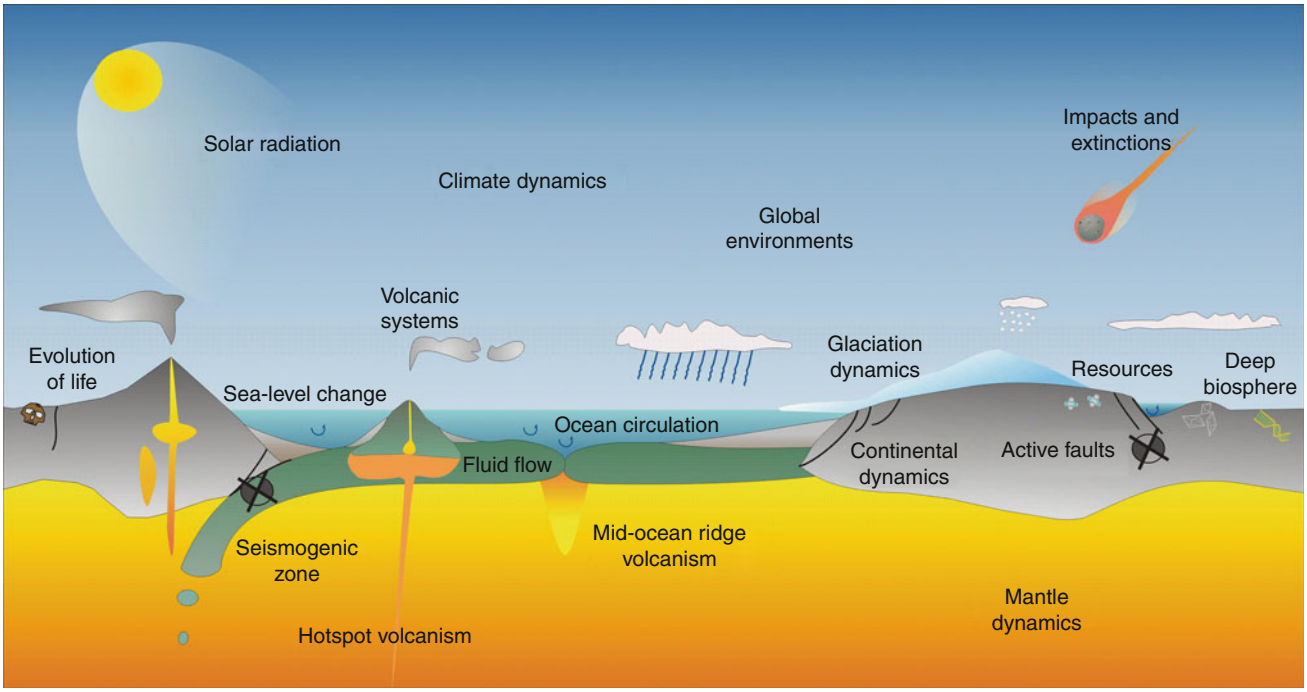
Today, direct access to the interior of the Earth's crust is still confined almost exclusively to sedimentary basins where hydrocarbon resources are recovered through deep industry-financed wells, if typically shallow exploration and mining for groundwater and minerals are neglected. Deep drilling for mainly scientific reasons has remained so far a relatively rare exception because the very high cost of drilling is a critical barrier in the design of research projects. Nevertheless, this threshold can be overcome if societal interests are pressing, if political and financial support is gained, and if international long-term cooperation is achieved. A small number of international Earth science programs devoted to scientific drilling have been established with widely agreed strategic scientific goals, international co-funding, and a necessary minimum of infrastructure and operational capabilities. While the overarching research goals are almost identical, the drilling programs are separated, based on largely technological grounds, into oceanic (Integrated Ocean Drilling Program; Scientific Planning Working Group, 2001) and continental (International Continental Scientific Drilling Program, Harms and Emmermann, 2007; Harms et al., 2007).

Goals of scientific drilling

The international programs for scientific drilling address topics of broad international interest, which can contribute significantly to solving important Earth science themes of high societal relevance (Figure 1). The research themes of these programs can be summarized in three overarching branches:

- The evolution and dynamics of the Earth's crust and lithosphere, especially in subduction zones and orogenic belts, with special emphasis on the physicochemical background of geological hazards such as earthquakes, volcanic eruptions, landslides, and meteorite impacts
- Environmental, climate, and paleoceanographic changes as recorded in geological archives of sediments in the oceans, in lakes, in ice shields, and in other depositional environments, including the resulting biological consequences, as well as the exploration of the widely unknown microbial biosphere living at depth
- Basic science of unconventional resources such as methane hydrates or geothermal energy, and of novel

Harsh K. Gupta (ed.), *Encyclopedia of Solid Earth Geophysics*, DOI 10.1007/978-90-481-8702-7,



Deep Scientific Drilling, Figure 1 Sketch of research themes addressed by major international scientific drilling programs.

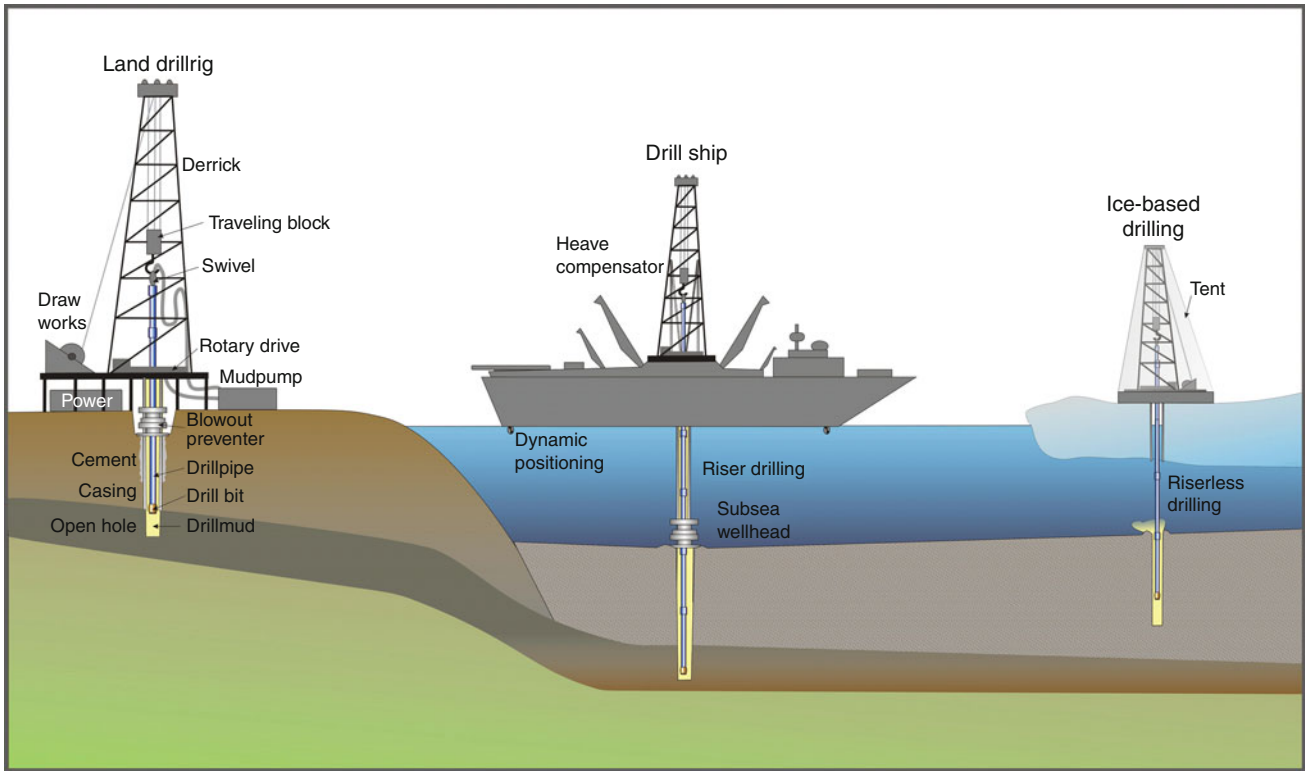
utilization of the deep underground environment such as CO_2-sequestration

This entry addresses mainly the first topic on lithospheric dynamics, hazards, and evolution of the Earth because most deep scientific drilling projects are dealing to a large extent with questions about internal forcing while climate, environmental, and subsurface biosphere research primarily employs shallow drilling techniques. Due to this limited depth extent, paleo-environmental scientific drilling projects are not included here.

Drilling, sampling, and monitoring techniques

A standard in most deep drilling operations (e.g., hydrocarbon exploration) is the rotary drilling technique. A cutting bit is lowered into the ground on a steel tube rotated through a rotary drive in a drilling derrick. The propelled drill string consists of connected pipe elements through which a drilling fluid is pumped down the well. The drill mud, usually water with clay minerals and other added solids to provide viscosity, cools the bit and flushes out chips of the destroyed volume of rock (drill cuttings) to the surface through the annulus between the borehole wall and the drill string. The weight-on-bit is controlled through variable lifting of the pipe, which is pulled down and kept in tension by heavyweight sections of drill pipe installed right above the drill bit. The drilling progress is supported by extending the string in the derrick with additional sections, or stands, of pipe. When drilling from a barge, ship or floating platform, the annulus extends only to the sea/lake floor, so mud and cuttings do not return to the drill rig. In this setting drilling can be performed with water in place of drilling mud, and cuttings spill out on sea or lake bottom around the well. However, if pressure control and mud return is required, an outer second pipe, a so-called "riser," is put in place so the mud and cuttings can be pumped back to the deck. Furthermore, in waves or swell, a heave compensation device is required to ensure constant weight-on-bit. Drilling through tidally lifted ice necessitates defrosting of the pipe in the ice section. Figure 2 summarizes basic drilling methods and important elements used in scientific drilling.

In normal rotary drilling technique, coring is performed with a hollow core bit that leaves a central column of rock that slides into the pipe while drilling progresses. After a few meters of coring, the whole assembly has to be pulled back out of the hole (called "pipe tripping") to get the core to surface. In scientific drilling projects, by contrast, continuous coring is often desirable. To avoid time-consuming pipe tripping runs, wireline coring techniques are applied. An inner core barrel is lowered through the drill string and latches above the core bit to retrieve the barrel with drilled-out rock column by a winch-drawn steel cable, obviating the need for a complete pipe trip with each core. The actual formation-cutting method varies depending on the type of rock or sediment present. Typically, thin-kerfed diamond core bits with high-rotation speed are used for hard rock drilling, roller cone abrasion bits are used for softer sedimentary rock, and nonrotating sharp edged hollow metal pistons of several meters length are hydraulically shot (forced) into soft sediments to collect cores and advance the borehole.



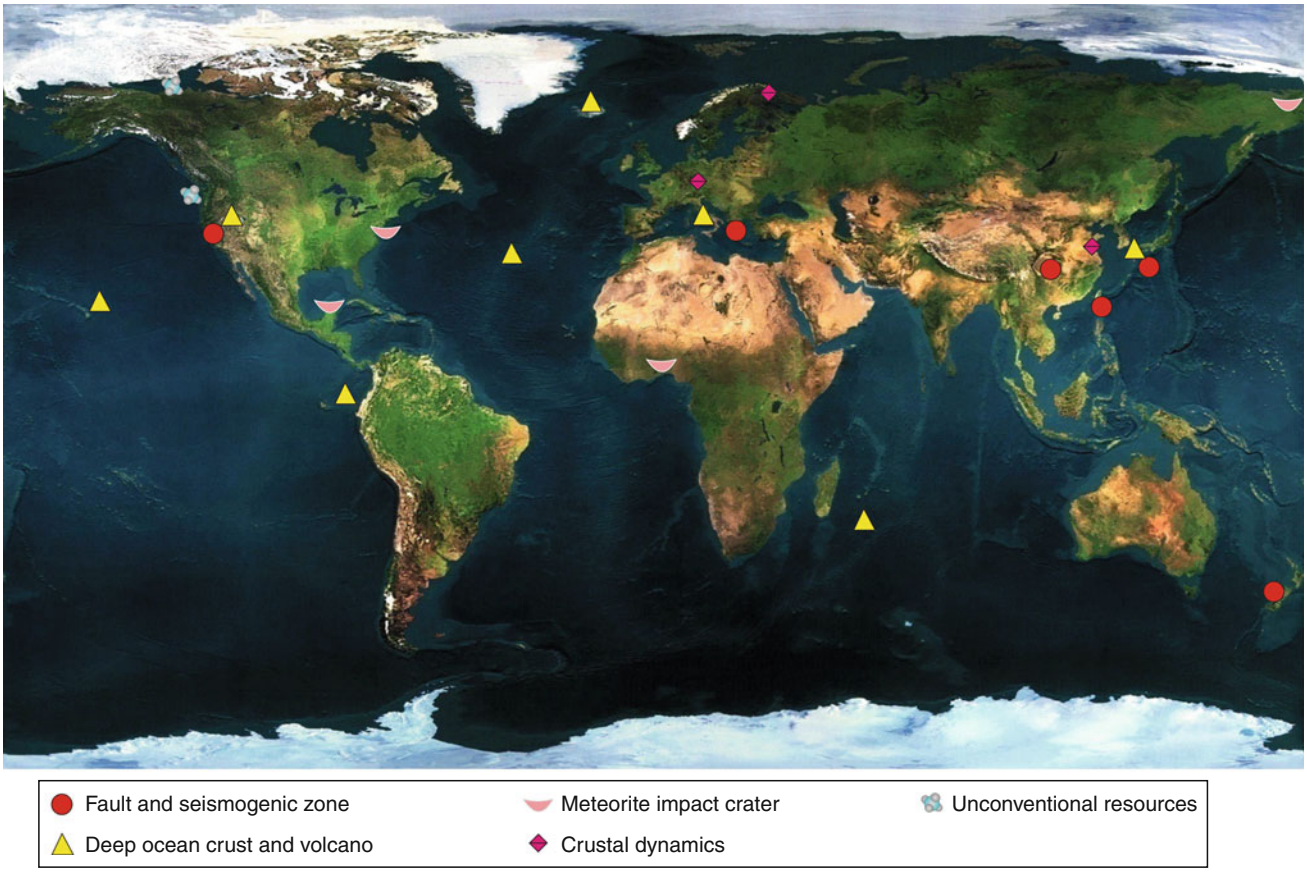
Deep Scientific Drilling, Figure 2 Principle technical components and styles of drilling operations in different environments.

After research wells are drilled, in situ measurements and monitoring of active processes in rocks and fluids allow for unprecedented data acquisition undisturbed by surface processes. A classic example is seismometer recordings in boreholes whose resolution and noise floor is improved by orders of magnitudes in comparison to surface measurements (Abercrombie, 1997). Sealed long-term borehole observatories have been installed in many places on land and at sea that are designed to record changes in temperature, pressure, strain, tilt, acceleration, and hydrologic parameters, and to sample fluids or microbiology (Fisher et al., 2008; Kastner et al., 2006; Prevedel and Kück, 2006).

Dynamics of continental crust

An unprecedented and still standing depth record in ultradeep scientific drilling was achieved in 1989 in the former Soviet Union when the Kola Superdeep Borehole reached 12,261 m after 25 years of drilling (Kozlovsky, 1987; Fuchs et al., 1990). Temperatures of less than 200°C, low tectonic stress, and special lightweight aluminum tools were key to this success. The well intersected Paleoproterozoic to Archean meta-volcano-sedimentary sequences and plutonic rocks in the Kola Peninsula of the northeastern Baltic Shield. The main scientific targets were the origin and extent of ore deposits as well as the principal structure and composition of the deeper crust. Important findings include – besides the falsification of a hypothesized basaltic layer in an onion-skin-like deep crust – the discovery of high-permeability and circulating fluids throughout the crust.

The German Continental Deep Drilling Program (KTB) aimed at principal physical and chemical processes and the geological evolution of the mid-European continental crust in an amalgamated collision zone (Emmermann and Lauterjung, 1997) (Figure 3). After 4 years of drilling through high-grade metamorphic rocks, a final depth of 9,101 m was reached in 1994 at formation temperatures of ~270°C. A truncated Paleozoic suture zone was strongly deformed by post-orogenic Mesozoic fault zones in an antiformal stack. Down to 9 km, faults contain saline, gas-rich free Ca-Na-Cl fluids, the geothermal gradient increases from about 21 to 28 K/km, and fluids percolate freely along fracture pathways. A detailed correlation of geophysical data with petrophysical properties at depth allowed for the calibration of surface geophysics to in situ conditions. Furthermore, for the first time, the state of stress and the rheological behavior of a deep crustal profile were analyzed in situ. Determinations of the minimal and maximum horizontal principal stress between 3 and 9 km depth yielded a uniform stress direction except for a major fault at 7 km depth (Zoback and Harjes, 1997). In conjunction with the deep main well, a 4,000-m-deep pilot hole just 200 m away allowed for the installation of a deep laboratory for fluid level (Schulze et al., 2000) and induced seismicity experiments (Baisch et al., 2002).



Deep Scientific Drilling, Figure 3 World map with important scientific deep drilling sites.

Ultrahigh pressure metamorphic (UHPM) rocks form the 1,000-km-long Qinling-Dabie-Sulu Belt in China, which was formed in a Triassic continental plate collision zone. From 2002 to 2005 a 5,158-m-deep corehole of the Chinese Continental Scientific Drilling project served to retrieve a vertical sample profile of the protracted multiphase deep (>150 km) subduction and rapid exhumation history of UHPM rocks (e.g., Liu et al., 2007; Zhang et al., 2006). The complex metamorphic path and its timing could be determined specifically in high-pressure minerals, which survived late-stage fluid-enhanced retrogression and overprint encased in zircons and other refractory accessories.

Volcanic systems

Volcanism is one of the most spectacular expressions of the dynamics of tectonic processes and the gigantic scale of recycling working in the Earth. The scientific investigation of volcanic processes is also crucial for societal security and thus an important mission for science. Since most of the history of volcanism is encapsulated in rocks that are overlain by sediments, younger lava flows or beneath the sea, drilling is in many cases a critical tool for the investigation of the different types of magmatism reaching surface.

Mid-ocean ridge magmatism

As defined by Conference Participants (1972), ophiolites represent a suite of oceanic crust plus upper mantle. Beneath a sediment cover, lava flows and pillow basalts overlay sheeted dike complexes, gabbroic rocks and, as part of the upper mantle, serpentinized peridotites and similar ultramafic rocks. Several holes drilled by the Deep Sea Drilling Program 1969 to 1985 and the Ocean Drilling Program 1985 to 1996 confirmed this structure but at the same time provided a much more complex picture. After several meters of lava flows were encountered in the early years of these programs, deeper parts of the oceanic plutonic rock succession were drilled in composite sections in tectonic windows and contributed much to the understanding of the ocean crust and mantle. Several drillholes in the Mid-Atlantic Ridge provided insight into slow-spreading modes with emplacement of ultramafic rocks. A 1,508-m-deep hole (735B) penetrated a long section of gabbroic rocks of the slow-spreading Southwest Indian Ridge off Madagascar (Dick et al., 2000).

The first lava-dike transition was drilled in 6 Ma age crust of the Nazca Plate off Ecuador (Hole 504B) down to 2,111 m below seafloor (Alt et al., 1996). And the first complete section of the upper oceanic crust was recovered by the Integrated Ocean Drilling Program at Site 1256 in

2005 and 2006 when the lava-sheeted dyke (1,061 m) and the sheeted dyke-gabbro (1,407 m depth) transitions of the Cocos Plate off Costa Rica were penetrated. The 15 Ma crust was formed in the East Pacific Rise by superfast spreading producing a thick succession (750 m) of lava flows, a relatively thin-sheeted dyke complex, and an upper gabbro section of still unknown extent (Alt et al., 2007). As the drilled fractionated gabbros (1,407–1,507 m) are not the residues of the overlying volcanics, the full sequence of plutonic ocean crust and the Moho transition to the upper mantle remains to be detected in future ocean drilling (Ildefonse et al., 2007).

Active volcano drilling and subduction-induced volcanism

The most common and typically most dangerous types of volcanoes on Earth are aligned along the subduction-related arc systems. They are typically fed by highly viscous andesitic to rhyolithic magmas whose degassing behavior is critical for their eruption mode, either effusive or explosive. While erosion and tectonics provide detailed insight into solidified and cooled volcanic structures, little is known about the physical and chemical conditions or permeability close to and in feeder fissures and magma conduits of active volcanoes. Despite challenging engineering and safety issues, some attempts have been made to drill into active volcanism. Still notable because of temperatures of above 1,400°C is the shallow drilling of the Kilauea Iki lava lake on Hawaii in the 1970s (Hardee et al., 1981). Furthermore, in Eastern California the recent, rhyolithic Inyo Lava Domes were sampled to understand degassing mechanisms (Eichelberger et al., 1986).

A series of devastating eruptions with frequently occurring pyroclastic flows and vulcanian explosions in early 1990 at Mount Unzen Volcano in southeastern Japan was accompanied by volcano-tectonic earthquakes and isolated tremor allowing for a geophysical pinpointing of the ascent path and surface appearance of magma. The identified conduit locale was within depth reachable by drilling, and justified drilling to sample in situ conditions after eruptions ended. Within the framework of the International Continental Scientific Drilling Program, the Unzen Volcano Drilling Project 2004 hit the conduit through an almost 2,000 m long but 1,500-m-deep well (Sakuma et al., 2008). Deviated drilling steered by downhole motors from the northern flank of Unzen was utilized to meet environmental and safety restrictions. Large-diameter casings were cemented in place to allow mud circulation for cooling of high temperatures, which were modeled to reach up to 600°C. However, maximum temperatures were as low as 160°C, and no gas kicks appeared due to enhanced permeability and 9 years of convection and groundwater circulation after the last eruptions. At 1,996 m depth, rocks were cored that could be identified as feeder dyke material (Nakada et al., 2005).

In the next stage of scientific drilling into volcanic features, calderas will be explored because they form through collapse during super-eruptions – the most explosive volcanism on Earth. An outstanding example of an actively inflating caldera is the Campi Flegrei volcanic field in Naples, Italy (Troise et al., 2007). A shallow crustal magma chamber, strong geothermal activity with enormous release of volcanic gases, and recent episodes of unrest call for drilling as a tool to investigate the driving mechanisms. Similar to the Unzen drilling, a deviated well is planned to be drilled from the land under the Bay of Naples toward the center of seismic unrest in the heart of the structure at 3.5 km depth.

Mantle plume volcanism

Hot spots truncate lithospheric oceanic and continental plates and build up giant volcanic edifices in Large Igneous Provinces (LIP) and Ocean Island Volcanoes. Their buildup requires vast melt extraction from the deep mantle and provides insight into mantle compositions, transport mechanisms, and the evolution of volcanic structures over space and time. Several boreholes of the Ocean Drilling Program opened windows into the Ontong-Java Plateau of the Western Pacific (Fitton et al., 2004) and the Kerguelen Plateau in the Southern Indian Ocean (Frey et al., 2000) to confirm their LIP origin.

In the Hawaii Scientific Drilling Project, the Mauna Kea volcano was cored with 95% recovery to a total depth of 3,520 m (3,508 m below sea level) to sample the evolution of plume melts over about 700,000 years as the volcanoes have drifted with the Pacific Plate over the Hawaii Plume (Stolper et al., 2009). The formation temperatures increase only from about 10°C to 50°C in the borehole (19°C km^{-1} gradient below 2,000 m) due to deep circulation of seawater in a complex hydrological regime including freshwater inflow as deep as 3,000 m. The circa 850 m subaerial and 2,440 m submarine layered lava and hyaloclastite accumulations represent almost 400 lithological units such as individual lava flows. They show geochemical and isotopic heterogeneity with tholeiitic versus alkaline magma transitions. This bimodal distribution of source components indicates a radial zoning within the melting region with a hot inner zone, which does not contain entrained mantle material and must be based in the lowermost mantle region (Bryce et al., 2005).

A further drilling target is supercritical fluids at 4–5 km depth in the Iceland Deep Drilling Project (Fridleifsson and Elders, 2007). Through coupled geothermal industry research and scientific investigation of in situ fluid-magma exchange, the first well at Krafla, Northern Iceland, hit rhyolithic magma at about 2,000 m depth before the well was cased for testing (Fridleifsson et al., 2010). In the late stages of planning is also drilling of the track of the Yellowstone plume in the Snake River Plain of southern Idaho (DePaolo and Weis, 2007).

Scientific drilling into volcanic rocks of the ocean crust provided the most important test and confirmation of the seafloor-spreading hypothesis. Samples retrieved from thickened ocean crust confirmed another type of seafloor

volcanism forming Large Iigneous Provinces, which contribute large magma volumes to the continental crust as well. These LIPs seem to represent transient plume heads while the plume tails persist in ocean island volcanoes. Hawaii served as pivotal example to drill deep into a single volcano to sample different parts of the plume and the volcanic history. Furthermore, recent penetrations into active volcanoes demonstrated the technical feasibility of such operations at very high temperatures. This access to deep volcanic processes is opening novel possibilities for monitoring.

Impact craters

Almost 180 craters on Earth are known currently that have been formed by astrophysical chance when celestial bodies collided with the Earth. The effects on life and the environmental consequences of the destructive forces as well as the input of extraterrestrial material, including organic compounds, is an overarching question in Earth systems history. Together with geophysical illumination, deep drilling into impact structures provides crucial data to study cratering, informing models of the impactor's size, impact angle, the resulting energy release through melting, evaporation, ejection and, most importantly, provides data for charting the aftermaths of such dramatic events. The best studied example to date is the Chicxulub crater (Dressler et al., 2003) in which the approximately 10-km-large impactor penetrated the whole crust and caused upper mantle rebound and an exemplary drastic biotic turnover 65.5 Ma ago (Schulte et al., 2010). The impact ejected a spherule-rich ash layer over the whole globe and released vast amounts of carbonate and sulfate gases previously bound in sedimentary rocks at the site of impact. The resulting darkness, cooling, geochemical cycle perturbations, and sudden change of environment interrupted the food chain and led to the major extinction event at that time.

In the past decade, four main impact crater studies including scientific drilling have been accomplished: the 200-km-wide Cretaceous-Paleogene Chicxulub Crater in Mexico (Hecht et al., 2004), the 1 Ma age 12-km-wide Bosumtwi Crater in Ghana (Ferrière et al., 2008), the late Eocene, 60 km Chesapeake Bay Crater in the Eastern USA (Gohn et al., 2008), and the 14 km Lake Elgygytgyn Crater in Chukotka, Siberia (Brigham-Grette and Melles, 2009). One of the important results is an improved understanding of the cratering process. Giant km-sized megablocks of target rocks were drilled in the lower sections of the huge Chesapeake and Chicxulub craters at 800–1,700 m depth. The blocks were shattered by shock waves into the transient crater and embedded in suevititic and lithic impact breccias by rapid mass transport in the crater within a few minutes. Unsorted marine debris flows driven by resurge and finally sorted airfall layers buried the blocks, which appear at a first glance as in situ target rocks. Furthermore, for large craters, a quasi-fluidization of target rocks and impactors leads to the formation of a central uplift by rebounded upper lithosphere, whereas in smaller craters such as Bosumtwi a central peak at about 500 m depth is formed by brittle deformation processes.

Active fault zones

Since the late 1990s, researchers have undertaken a wide range of ambitious projects aimed at drilling into active faults in a range of tectonic settings, both on land and at sea. These projects have been motivated by the recognition that fault zone processes are poorly understood from surface data, and that a combination of (a) samples of fault materials from "live" systems in the subsurface and (b) access through boreholes to in situ measurements and monitoring of ambient conditions in active faults, can provide key new observations for learning about these processes. Our understanding of the mechanics and dynamics of plate boundary faulting is severely limited by a lack of information on ambient conditions and mechanical properties of active faults at depth. A major goal in earthquake mechanics research is thus direct in situ sampling and instrumentation by drilling into the seismogenic zone of active faults, including both inter- and intra-plate fault systems in many tectonic settings.

Understanding of the complex physics of tectonic faulting, earthquakes, and the generation of tsunami in the Earth's crust is one of the grand challenges of the geosciences in general and seismology in particular (Lay, 2009). Because earthquakes take place deep below the Earth's surface, direct observation of processes acting in the faults themselves has generally not been possible; what we know about earthquake processes is derived from seismologic observations, geodetic deformation, and studies of exhumed rocks containing products of slip at depth long ago. Researchers have learned and inferred a great deal from these studies, but they are inherently limited. Much of fault zone drilling is aimed at understanding earthquake slip, but in fact fault physics encompasses a spectrum of slip and slip rates ranging from purely aseismic behavior, such as steady creep, through events such as slow earthquakes and episodic creep, to true seismicity. It is not yet clear how similar or distinct the physics of these modes of fault slip may be (Ide et al., 2007).

Hence, some of the key questions addressed through fault zone drilling projects include the following: What are the deformation mechanisms of fault slip in both aseismic and seismic conditions? Why is slip sometimes seismic, sometimes aseismic, and sometimes in an intermediate state, and what governs the transitions among these processes? How do earthquakes nucleate, and how does rupture grow and stop? Are there precursory phenomena in the fault or near field that indicate a preparatory phase to earthquake or slip nucleation? If so, can a prediction or early warning strategy based on borehole observation be developed? What are the processes governing tsunami-genic slip? What is the role of fluid content and fluid pressure in modulating faulting processes, especially during rapid seismic slip? How does

permeability of the fault zone evolve interseismically and coseismically? What is the stress tensor in and around a fault zone throughout the earthquake cycle or other slip? What controls localization of slip versus distributed strain in faulting?

Pursuing these questions, a number of major national and international efforts have been carried out or planned in recent years. These prominently include the San Andreas Fault Observatory at Depth (SAFOD), the Taiwan Chelungpu Fault Drilling Project (TCDP), and the Gulf of Corinth drilling on land, and the Nankai Trough Seismogenic Zone Experiment (NanTroSEIZE) at sea.

Fault zone drilling projects are different in many respects from most other scientific drilling efforts. Rather than the complete stratigraphic intervals or sampling of broader rock volumes targeted by most other deep drilling, fault zone targets include *discontinuities* and small anomalous intervals. The targets of interest are often the exact locations of poorest drilling conditions, such as a highly fractured damage and gouge zone with potential for excess pore pressure above hydrostatic condition, anomalous stress state, weak rock, and other drilling hazards. The scientific approach commonly focuses on obtaining very extensive downhole measurements, logs, cores, and seismic structure in and around these fault zone environments. Measurements at the time of drilling as well as long-term observations are emphasized.

Nojima fault projects, Japan

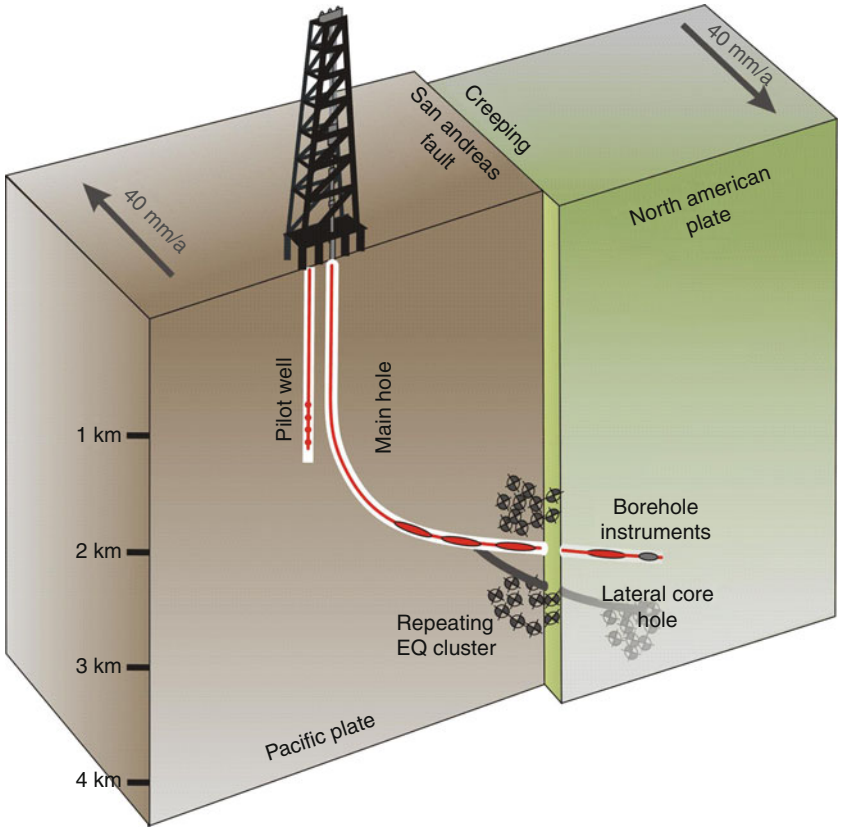
Soon after the Kobe earthquake in 1995 (M6.9), several research teams drilled multiple boreholes into and across the shallow portion of the Nojima fault (the locus of that event) from ~750 to ~1,800 m depth (Ito et al., 1999). Detailed compositional and structural analyses identified a narrow fault core of gouge and surrounding damage zone, with asymmetric structure in the footwall and hanging wall (Boullier et al., 2001; Fujimoto et al., 2001; Ito et al., 1999; Ohtani et al., 2001; Tanaka et al., 2001). Shear wave splitting studies of aftershocks showed a rapid evolution of fast direction after the earthquake, interpreted as evidence of fracture healing during a 12 month postseismic period (Tadokoro and Ando, 2002). From repeated injection tests, permeability change with time was observed (Tadokoro et al., 2000). The structural properties (fault zone width, shear wave velocity, and Qs) of the Nojima fault were studied using Love-wave-type fault zone trapped waves (LTWs) recorded at two borehole seismic stations.

SAFOD, California

The SAFOD borehole was drilled into the creeping section of the San Andreas Fault in central California close to the northern end of the locked Parkfield segment (Williams et al., 2010; Zoback et al., 2010). A vertical pilot hole was drilled in 2002 to circa 2 km depth (Hickman et al., 2004). The main hole was drilled in three phases from 2004 to 2007 to a depth of 4 km (3 km true vertical depth), the first 1,800 m vertical, then deviated toward the steeply dipping San Andreas Fault Zone, and finally crossing the fault at an angle of 55° (Figure 4; Hickman et al., 2007). High-resolution seismic observations in the well and at the surface allowed to steer the well directly into a zone of repeating micro-earthquakes. The initial main borehole was logged and cased; repeated casing diameter logs over time showed that two intervals of casing several meters wide were actively deforming. Together with drastically decreasing values of v_p, v_s, and resistivity, the deforming intervals were interpreted as a manifestation of creep in two discrete fault zones (Hickman et al., 2007). In 2007, the hole was reentered and sidetracks were drilled to obtain core samples (Williams et al., 2010). A core interval of approximately 40 m was obtained from one of the two identified major fault zones, and additional cores were collected from adjacent wall rock intervals. Borehole seismometers installed in the well detected target earthquakes as small as M ~ 2 within a few tens of meters of the source, allowing unprecedented high-frequency detection of source processes (Ellsworth et al., 2007). A notable result from SAFOD has been the absence of evidence for either a thermal anomaly from frictional heating around the fault core or an excess pore fluid pressure substantially above hydrostatic anywhere in or around the faults, as well as geochemical evidence that the fault zone is a permeability barrier to cross-fault flow (Hickman et al., 2007; Wiersberg et al., 2007), simultaneously adding evidence for absolute fault weakness in the San Andreas and opening new questions about the source of that weakness. The actively deforming zones consist of about 2 m wide foliated fault gouge showing dissolution–precipitation reactions, serpentinite porphyroclasts and smectite–illite coatings on minerals lubricating the fault in the creeping section. A permanent borehole monitoring package for seismic and tilt observations was installed in 2008. As this instrument set failed due to hostile fluid conditions, a new observatory system is in the design phase and will be installed for long-term measurements. The long-term observation of in situ physical parameters over several seismic cycles is the key goal of SAFOD.

Corinth rift laboratory, Greece

The Gulf of Corinth is a location of back-arc extension within the Hellenic Arc system and is among the most seismically active areas of Europe (Cornet, 2007). The Corinth Rift Laboratory (CRL) project drilled into the Aigion fault at 760 m depth and continued to 1,000 m, encountering cataclastic fault rocks in a karstic carbonate host-rock environment. The fault zone forms a strong hydrologic barrier with an associated pressure differential across it. Transients in pore pressure have been detected associated with remote earthquakes (Cornet et al., 2004). Since 2007, pore pressure measurements have been made in the borehole, but only in the hanging wall of the fault zone. Plans for upcoming activities at CRL include



Deep Scientific Drilling, Figure 4 Sketch of drilling design in the San Andreas Fault Zone Observatory at Depth (SAFOD).

near-term deployment of a string of high-frequency three-component geophones, packers, pressure transducers, tiltmeters, and thermistance meters across the fault penetration, and also a proposed 4.5-km-deep borehole to monitor pore pressure and other features of the seismogenic zone (Cornet, 2007).

Taiwan Chelungpu drilling project, Taiwan

The devastating 1999 M7.6 Chi-Chi earthquake in Taiwan was a thrust event in a collisional tectonic setting that exhibited very large slip (>10 m), including large surface displacements, on the Chelungpu fault. The principal objective of drilling was to sample the main slip surface or zone, if one could be identified, at a relatively short post-seismic interval. The fault zone was drilled in 2004 and 2005 in an area of high-velocity large slip during the main shock, and crossed the fault zone and several hundred meters of the footwall in two adjacent holes (Ma et al., 2006). The main slip zone and several subsidiary (or older) fault cores were identified in the borehole at ~1,100 m depth (Hung et al., 2007). Studies of the fault gouge particle size, thickness, and composition have yielded quantitative estimates for the work done in forming gouge and damage (fracture energy), contributing to the understanding of energy budgets during earthquakes (Ma et al., 2006). Borehole hydrologic tests, temperature logging, and sample-based frictional experiments contributed to a tight integration of borehole-derived information with seismological observations.

Wenchuan fault scientific drilling program, Sichuan, China

The devastating Wenchuan earthquake (M7.9) of May 12, 2008 caused at least 68,000 deaths. It occurred on a reverse fault in crystalline basement rocks, and slip propagated to the near surface. Only 178 days after the earthquake, the Wenchuan Fault drilling program was begun to access the coseismic slip zone for sampling and in situ observations, in up to four planned boreholes. The first borehole was completed in 2008, but results have not been reported internationally as of this writing.

Subduction zone fault drilling at sea

For over 20 years, drilling efforts have targeted active large-offset thrusts and décollement faults in subduction zone settings, but these have mostly been in areas shallower than the depth of seismogenic strain accumulation and release. Beginning in the 1980s with the Barbados Accretionary Wedge, at the Atlantic Caribbean plate boundary, drilling was aimed at understanding the faulting conditions, ambient pore fluid pressure, and mechanics in a very low angle (~2° dip) décollement thrust fault at the

base of an accretionary wedge (Moore, 1989). Drilling and logging measurements made while drilling (LWD) there demonstrated that the fault zone was a mechanically very weak interval with high excess pore fluid pressure and patchily distributed areas of anomalously high porosity (Moore et al., 1998). Installation of a borehole seal (CORK), downhole testing, and other studies at ODP Sites 948 and 949 have shown that the fault zone has effective permeability that is strongly pressure dependent and varies over ~4 orders of magnitude with fluid pressure (Fisher, 2005). Drilling at Costa Rica off the Nicoya peninsula also elucidated the structure of a shallow décollement zone. Structural studies and geochemical tracers of hydrological flow paths suggest that two strongly decoupled systems exist above and below the ~20-m-thick main fault zone, which acts simultaneously as a low-permeability barrier to cross-fault flow and a high-permeability channel for along-fault flow (Tobin et al., 2001). Installation of a sealed borehole monitoring system there documented an apparently weakly overpressured décollement zone that exhibits temporal pressure excursions tied to apparent aseismic transients recorded geodetically on land (Davis and Villinger, 2006). Similarly, instrumented boreholes at the Nankai Trough off Shikoku Island in Japan show pressure transients associated temporally and spatially with swarms of very low frequency earthquakes, suggesting a long-distance (tens of kilometers) poroelastic connection between the earthquake events and the boreholes (Davis et al., 2006). The same data show long-term pressure trends interpreted as a response to gradual interseismic strain accumulation.

Beginning in 2007, IODP embarked on its most ambitious fault zone drilling project to date, to drill a transect of boreholes, including two deeper than 1,500 m below the seafloor, spanning the up-dip transition from seismic to aseismic fault mechanics. The Nankai Trough Seismogenic Zone Experiment (NanTroSEIZE) project is the first attempt to drill, sample, and instrument the seismogenic portion of a plate boundary fault or megathrust within a subduction zone. The fundamental goal of NanTroSEIZE science plan (Tobin and Kinoshita, 2006) is to access, sample, and then instrument the faults in aseismic and seismogenic regions of the megathrust system. This is pinpointed in a region where recently discovered very low frequency (VLF) earthquakes have been shown to occur (Ito and Obara, 2006), as well as possible shallow low frequency tremor (Obana and Kodaira, 2009). NanTroSEIZE involves drilling of faults in the active plate boundary system at several locations off the Kii Peninsula of Honshu Island, Japan, from the very shallow onset of the plate interface at the trench to ~7 km depths where earthquake slip and locking takes place. At this location, the plate interface and active mega-splay faults implicated in causing tsunami are accessible to drilling within the region of coseismic rupture in the 1944 Tonankai (magnitude 8.1) great earthquake (Baba et al., 2006). In 2007, an initial transect of eight sites was drilled to depths from 400 to 1,400 m below the seafloor (Tobin et al., 2009a, b). Sampling of major large-offset thrust faults imaged in 3D seismic data (Moore et al., 2007) identified strongly localized fault slip zones even at positions just a few hundred meters below the surface where the host sediments are poorly consolidated. New evidence from detailed vitrinite reflectance measurements (Sakaguchi et al., 2009) suggests that frictional thermal heating may have affected even these presumed aseismic faults. In a second stage of drilling in 2009, the riser system was deployed for the first time to drill a cased observatory hole to 1,600 m (Saffer et al., 2009), and the incoming plate geology and fluids were sampled on the incoming plate beyond the subduction front (Underwood et al., 2009), to define a baseline and mass balance for material delivered to the subduction interface. Borehole observatories to monitor seismic activity, strain, tilt, temperature, and pore fluid pressure were planned for installation in late 2010. The main objective of NanTroSEIZE, however, is slated to take place in 2011 and 2012, with planned deep riser drilling to 7 km below the seafloor, across the full plate interface. The objective is to access and instrument the full Nankai plate interface within the seismogenic zone, sampling the fault and wall rocks and obtaining measurements for years to come as the interseismic period passes.

Unconventional resources

Among the unexplored, unconventional resources requiring drilling for basic scientific research are gas hydrates, in which methane and other gaseous hydrocarbons combine with water to form crystalline, ice-like structures. Their stability field varies only within a few MPa and °C but they can be stable even at temperatures above 0°C in deepwater or buried sediments. Two widespread environments of their occurrence have been known for a long time: in sediments of the continental slopes where they simulate bottom reflectors in seismic images, and below permafrost in polar regions (Kvenvolden, 1988; Buffett and Archer, 2004). Although estimates of the global abundance of gas hydrates vary widely, the total volume appears to be in the same order of magnitude as conventional hydrocarbon resources. At the same time, they bear a high potential of generating continental slope instabilities causing giant submarine landslides and contributing significantly to the atmospheric greenhouse effect when released into atmosphere (Dickens, 2003).

Little is known about the distribution of gas hydrates, and even less about the physical and chemical parameters that affect the formation and stability of gas hydrates. At sea, the best studied, drilled areas are in the Eastern Pacific off Vancouver Island and at Blake Ridge in the western Atlantic Ocean off Georgia (Tréhu et al., 2006). Continental gas hydrate accumulations were drilled in the Mackenzie River delta in northern Canada (Dallimore and Collett, 2005). Two observation wells and one central production well dug to 1,200 m depth were used for an

extensive field program including coring, downhole logging, cross-hole experiments, and production testing. All three wells show massive thermogenic methane gas hydrate accumulations between 890 and 1,100 m depth with the highest concentration at the base of the gas hydrate stability zone (Takahashi et al., 2005). Similarly, in marine sediments the hydrates are concentrated in coarse-grained, high-permeability layers. A series of controlled depressurization and thermal-stimulation experiments yielded important baseline data on phase equilibria as well as on petrophysical and hydraulic parameters. Warm drill mud caused dissociation of gas hydrate in a 13-m-thick hydrate concentration zone at about 920 m depth (80% pore saturation) and produced 50–350 m^3/day. Existing fluid pathways such as in faults appear to enhance the gas production. Concentrated gas flow along such pathways seems to be also critical for the gas hydrate formation, in sediments and especially on seafloor.

Summary

This paper summarizes the most significant, but not all deep scientific drilling projects that have been carried out in recent years by international research teams. Key goals of these projects encompass topics of high societal relevance such as a better understanding of the physics driving geological disasters and a wise utilization of resources. Various several kilometer deep drilling ventures allowed for unprecedented sampling and in situ monitoring of dynamic processes governing formation and restructuring of the Earth's crust such as collision events, volcanism, faulting, or impact cratering. Despite grand engineering challenges, deep Earth access techniques are now available for novel burgeoning scientific exploration. Scientific drilling promises to be an integral tool of research generating new observations for the time to come.

Bibliography

Abercrombie, R. E., 1997. Near surface attenuation and site effects from comparison of surface and deep borehole recordings. *Bulletin of the Seismological Society of America*, **87**, 731–744.

Alt, J. C., Laverne, C., Vanko, D. A., Tartarotti, P., Teagle, D. A. H., Bach, W., Zuleger, E., Erzinger, J., Honnorez, J., Pezard, P. A., Becker, K., Salisbury, M. H., and Wilkens, R. H., 1996. Hydrothermal alteration of a section of upper oceanic crust in the eastern equatorial Pacific: a synthesis of results from Site 504 (DSDP Legs 69, 70, and 83, and ODP Legs 111, 137, 140, and 148). In Alt, J. C., Kinoshita, H., Stokking, L. B., and Michael, P. (eds.), *Proceedings of the Ocean Drilling Program: Scientific Results*, Vol. 148, pp. 417–434.

Alt, J. C., Teagle, D. A. H., Umino, S., Miyashita, S., Banerjee, N. R., Wilson, D. S., and the IODP Expeditions 309 and 312 Scientists, 2007. IODP Expeditions 309 and 312 drill an intact section of upper oceanic basement into gabbros. *Scientific Drilling,* **6**, 4–10, doi:10.0/iodp.sd.01.2007.

Baba, T., Cummins, P. R., Hori, T., and Kaneda, Y., 2006. High precision slip distribution of the 1944 Tonankai earthquake inferred from tsunami waveforms: possible slip on a splay fault. *Tectonophysics*, **426**(1–2), 119–134, doi:10.1016/j.tecto.2006.02.015.

Baisch, S., Bohnhoff, M., Ceranna, L., Tu, Y., and Harjes, H.-P., 2002. Probing the crust to 9 km depth: fluid injection experiments and induced seismicity at the KTB superdeep drilling hole, Germany. *Bulletin of the Seismological Society of America*, **92**, 2369–2380.

Boullier, A.-M., Ohtani, T., Fujimoto, K., Ito, H., and Dubois, M., 2001. Fluid inclusions in pseudotachylytes from the Nojima fault, Japan. *Journal of Geophysical Research*, **106**, 21965–21977.

Brigham-Grette, J., and Melles, M., 2009. Complex drilling logistics for Lake Elgygytgyn. *Scientific Drilling*, **7**, 38–39, doi:10.2204/iodp.sd.7.05.2009.

Bryce, J. G., DePaolo, D. J., and Lassiter, J. C., 2005. Geochemical structure of the Hawaiian plume: Sr, Nd, and Os isotopes in the 2.8 km HSDP-2 section of Mauna Kea volcano. *Geochemistry Geophysics Geosystems*, **6**, Q09G18, doi:10.1029/2004GC000809.

Buffett, B., and Archer, D., 2004. Global inventory of methane clathrate: sensitivity to changes in the deep ocean. *Earth and Planetary Science Letters*, **227**, 185–199.

Conference Participants, 1972. Penrose field conference on ophiolites. *Geotimes*, **17**, 24–25.

Cornet, F. H., 2007. The Corinth Rift Laboratory or an in situ Investigation on Interactions between Fluids and Active Faults. *Scientific Drilling Special Issue*, **1**, 35–38.

Cornet, F. H., Bernard, P., and Moretti, I., 2004. The Corinth Rift Laboratory. *Compte Rendu Geosci*, **336**, 235–241, doi:10.1016/j.crte.2004.02.001.

Dallimore, S. R., and Collett, T. S., 2005. Summary and implications of the Mallik 2002 Gas Hydrate Production Research Well Program. In Dallimore, S. R., and Collett, T. S. (eds.), *Scientific Results from the Mallik 2002 Gas Hydrate Production Research Well Program, Mackenzie Delta, Northwest Territories, Canada*. Canada: Geological Survey of Canada, p. Bulletin 585.

Davis, E. E., and Villinger, H. W., 2006. Transient formation fluid pressures and temperatures in the Costa Rica forearc prism and subducting oceanic basement: CORK monitoring at ODP Sites 1253 and 1255. *Earth and Planetary Science Letters*, **245**(1–2), 232–244, doi:10.1016/j.epsl.2006.02.042.

Davis, E. E., Becker, K., Wang, K., Obara, K., Ito, Y., and Kinoshita, M., 2006. A discrete episode of seismic and aseismic deformation of the Nankai trough subduction zone accretionary prism and incoming Philippine Sea plate. *Earth and Planetary Science Letters*, **242**, 73–84.

DePaolo, D. J., and Weis, D., 2007. Hotspot volcanoes and large igneous provinces. In Harms, U., Koeberl, C., and Zoback, M. D. (eds.), *Continental Scientific Drilling: A Decade of Progress and Challenges for the Future*. Heidelberg: Springer, pp. 259–288.

Dick, H. J. B., Natland, J. H., Alt, J. C., Bach, W., Bideau, D., Gee, J. S., Haggas, S., Hertogen, J. G. H., Hirth, G., Holm, P. M., Ildefonse, B., Iturrino, G. J., John, B. E., Kelley, D. S., Kikawa, E., Kingdon, A., LeRoux, P. J., Maeda, J., Meyer, P. S., Miller, J. D., Naslund, H. R., Niu, Y., Robinson, P. T., Snow, J., Stephen, R. A., Trimby, P. W., Worm, H. U., and Yoshinobu, A., 2000. A long in-situ section of the lower ocean crust: Results of ODP Leg 176 drilling at the Southwest Indian Ridge. *Earth and Planetary Science Letters*, **179**, 31–51.

Dickens, G. R., 2003. A methane trigger from global warming? *Science*, **299**, 1017–1020.

Dressler, B. O., Sharpton, V. L., Morgan, J., Buffler, R., Moran, D., Smit, J., Stöffler, D., and Urrutia, J., 2003. Investigating a 65-Ma-old smoking gun: deep drilling of the Chicxulub impact structure. *Eos*, **84**(14), 125–130, doi:10.1029/2003EO140001.

Eichelberger, J. C., Carrigan, C. R., Westrich, H. R., and Price, R. H., 1986. Nonexplosive silicic volcanism. *Nature*, **323**, 598–602.

Ellsworth, W., Malin, P. E., Imanishi, K., Roecker, S. W., Nadeau, R., Oye, V., Thurber, C. H., Waldhauser, F., Boness, N. N.,

Hickman, S. H., and Zoback, M. D., 2007. Seismology inside the fault zone: applications to fault-zone properties and rupture dynamics. *Scientific Drilling*, **1**, 85–88. Special issue.

Emmermann, R., and Lauterjung, J., 1997. The German Continental Deep Drilling Program KTB: overview and major results. *Journal of Geophysical Research*, **102**, 18179–18201, doi:10.1029/96JB03945.

Ferrière, L., Koeberl, C., Ivanov, B. A., and Reimold, W. U., 2008. Shock metamorphism of Bosumtwi Impact Crater rocks, shock attenuation, and uplift formation. *Science*, **12**(322), 1678–1681, doi:10.1126/science.1166283.

Fisher, A. T., Davis, E. E., and Becker, K., 2008. Borehole-to-borehole hydrologic response across 24 km in the upper oceanic crust: implications for crustal scale properties. *Journal of Geophysical Research*, **113**, B07106, doi:10.1029/2007JB005447.

Fisher, A. T., 2005. Marine hydrogeology: recent accomplishments and future opportunities. *Hydrogeology Journal*, **13**, 69–97, doi:10.1007/s10040-004-0400-y.

Fitton, J. G., Mahoney, J. J., Wallace, P. J., and Saunders, A. D., 2004. Leg 192 synthesis: origin and evolution of the Ontong Java Plateau. In Fitton, J. G., Mahoney, J. J., Wallace, P. J., and Saunders, A. D. (eds.), *Proceedings of the ODP, Scientific Results*, 192, doi:10.2973/odp.proc.sr.192.101.2004.

Frey, F. A., Coffin, M. F., Wallace, P. J., Weis, D., Zhao, X., Wise, S. E., Wähnert, V., Teagle, D. A. H., Saccocia, P. J., Reusch, D. N., Pringle, M. S., Nicolaysen, K. E., Neal, C. R., Müller, R. D., Moore, C. L., Mahoney, J. J., Keszthelyi, L., Inokuchi, H., Duncan, R. A., Delius, H., Damuth, J. E., Damasceno, D., Coxall, H. K., Borre, M. K., Boehm, F., Barling, J., Arndt, N., and Antretter, M., 2000. Origin and evolution of a submarine large igneous province: the Kerguelen Plateau and Broken Ridge, Southern Indian Ocean. *Earth and Planetary Science Letters*, **176**, 73–89.

Fridleifsson, G. O., and Elders, W. E., 2007. Progress report on the Iceland deep drilling project (IDDP). *Scientific Drilling*, **4**, 26–29, doi:10.2204/iodp.sd.4.04.2007.

Fridleifsson, G. O., Pálsson, B., Stefánsson, B., Albertsson, A., Gunnlaugsson, E., Ketilsson, J., Lamarche, R., and Andersen, P. E., 2010. Iceland Deep Drilling Project. The first IDDP Drill Hole Drilled and Completed in 2009. *Proceedings World Geothermal Congress 2010*. Indonesia, Bali, PP. 25–29.

Fuchs, K., Kozlovsky, Y. A., Krivtsov, A. I., and Zoback, M. D., 1990. *Super-Deep Continental Drilling and Deep Geophysical Sounding*. Heidelberg: Springer.

Fujimoto, K., Tanaka, H., Higuchi, T., Tomida, N., Ohtani, T., and Ito, H., 2001. Alteration and mass transfer inferred from the Hirabayashi GSJ drill penetrating the Nojima fault, Japan. *The Island Arc*, **10**, 401–410.

Gohn, G. S., Koeberl, C., Miller, K. G., Reimold, W. U., Browning, J. V., Cockell, C. S., Horton, J. W., Kenkmann, T., Kulpecz, A. A., Powars, D. S., Sanford, W. E., and Voytek, M. A., 2008. Deep drilling into the Chesapeake Bay impact structure. *Science*, **27**(320), 1740–1745, doi:10.1126/science.1158708.

Hardee, H. C., Dunn, J. C., Hills, R., and Ward, R. W., 1981. Probing the melt zone of Kilauea Iki lava lake, Kilauea volcano, Hawaii. *Geophysical Research Letters*, **8**, 1211–1214.

Harms, U., and Emmermann, R., 2007. History and status of the international continental scientific drilling program. In Harms, U., Koeberl, C., and Zoback, M. D. (eds.), *Continental Scientific Drilling. A Decade of Progress, and Challenges for the Future*. Berlin/Heidelberg: Springer, pp. 1–53.

Harms, U., Koeberl, C., and Zoback, M. D., 2007. *Continental Scientific Drilling. A Decade of Progress, and Challenges for the Future*. Berlin: Springer.

Hecht, L., Wittmann, A., Schmitt, R. T., and Stöffler, D., 2004. Composition of impact melt particles and the effects of post-impact alteration in suevitic rocks at the Yaxcopoil-1 drill core, Chicxulub crater, Mexico. *Meteoritics & Planetary Science*, **39**, 1169–1186.

Hickman, S., Zoback, M. D., and Ellsworth, W. E., 2004. Introduction to special section: preparing for the San Andreas Fault Observatory at Depth. *Geophysical Research Letters*, **31**, L12S01, doi:10.1029/20040GL20688.

Hickman, S., Zoback, M., Ellsworth, W., Boness, N., Malin, P., Roecker, S., and Thurber, C., 2007. Structure and properties of the San Andreas fault in central California: recent results from the SAFOD experiment. *Scientific Drilling*, **Special issue No. 1**, 29–32, doi:10.2204/iodp.sd.s01.39.2007.

Hung, J., Ma, K., Wang, C., Song, S., Ito, H., Lin, W., and Yeh, E., 2007. Subsurface structure, fault zone characteristics and stress state in scientific drill holes of Taiwan Chelungpu fault drilling project. *Scientific Drilling*, **1**, 55–58. Special issue.

Ide, S., Beroza, G. C., Shelly, D. R., and Uchide, T., 2007. A scaling law for slow earthquakes. *Nature*, **447**, 76–79.

Ildefonse, B., Christie, D. M., and the Mission Moho Workshop Steering Committee, 2007. Mission Moho workshop: drilling through the oceanic crust to the mantle. *Scientific Drilling*, **4**, 11–17, doi:10.2204/iodp.sd.4.02.2007.

Ito, Y., and Obara, K., 2006. Very low frequency earthquakes within accretionary prisms are very low stress-drop earthquakes. *Geophysical Research Letters*, **33**, L09302, doi:10.1029/2006GL025883.

Ito, H., Fujimoto, K., Tanaka, H., and Lockner, D., 1999. Proceedings of the international workshop on the Nojima fault core and borehole data analysis. *Geological Survey of Japan Interim Report, No. EQ/00/1*, USGS Open-File 895 Report 000-129.

Kastner, M., Becker, K., Davis, E. E., Fischer, A. T., Jannasch, H. W., Solomon, E. A., and Wheat, G., 2006. New insights into the hydrogeology of the oceanic crust through long-term monitoring. *Oceanography*, **19**(4), 46–57.

Kozlovsky, Y. A., 1987. *The Superdeep Well of the Kola Peninsula*. Berlin: Springer.

Kvenvolden, K. A., 1988. Methane hydrate – a major reservoir of carbon in the shallow geosphere? *Chemical Geology*, **71**, 41–51.

Lay, T. (ed.), 2009. *Seismological Grand Challenges in Understanding Earth's Dynamic Systems*. Report to the National Science Foundation, IRIS Consortium, 76 pp.

Liu, F. L., Xu, Z. Q., Liou, J. G., Dong, H. L., and Xue, H. M., 2007. Ultrahigh-pressure mineral assemblages in zircons from the surface to 5158 m depth in cores of the main drill hole, Chinese Continental Scientific Drilling project, SW Sulu belt, China. *International Geological Review*, **49**, 454–478.

Ma, K.-F., Tanaka, H., Song, S.-R., Wang, C.-Y., Hung, J.-H., Tsai, Y.-B., Mori, J., Song, Y.-F., Yeh, E.-C., Soh, W., Sone, H., Kuo, L.-W., and Wu, H.-Y., 2006. Slip zone and energetics of a large earthquake from the Taiwan Chelungpu-fault Drilling Project. *Nature*, **444**, 473–476, doi:10.1038/nature05253.

Moore, J. C., 1989. Tectonics and hydrogeology of accretionary prisms: role of the decollement zone. *Journal of Structural Geology*, **11**, 95–106.

Moore, J. C., Klaus, A., Bangs, N. L., Bekins, B., Buecker, C. J., Brueckmann, W., Erickson, S. N., Hansen, O., Horton, T., Ireland, P., Major, C. O., Moore, G. F., Peacock, S., Saito, S., Screaton, E. J., Shimeld, J. W., Stauffer, P. H., Taymaz, T., Teas, P. A., and Tokunaga, T., 1998. Consolidation patterns during initiation and evolution of a plate-boundary décollement zone: Northern Barbados accretionary prism. *Geology*, **26**, 811–814, doi:10.1130/0091-7613(1998) 026<0811:CPDIAE>2.3.CO;2.

Moore, G., Bangs, N., Taira, A., Kuramoto, S., Pangborn, E., and Tobin, H., 2007. Three-dimensional splay fault geometry and implications for tsunami generation. *Science*, **318**, 1128–1131.

Nakada, S., Uto, K., Sakuma, S., Eichelberger, J. C., and Shimizu, H., 2005. Scientific results of conduit drilling in the Unzen Scientific Drilling Project (USDP). *Scientific Drilling*, **1**, 18–22.

Obana, K., and Kodaira, S., 2009. Low-frequency tremors associated with reverse faults in a shallow accretionary prism. *Earth and Planetary Science Letters*, **287**, 168–174.

Ohtani, T., Tanaka, H., Fujimoto, K., Higuchi, T., Tomida, N., and Ito, H., 2001. Internal structure of the Nojima fault zone from the Hirabayashi GSJ drill core. *The Island Arc*, **10**, 392–400.

Prevedel, B., and Kück, J., 2006. New developments in long-term downhole monitoring arrays. *Scientific Drilling*, **3**, 34–37, doi:10.2204/iodp.sd.2.08.2006.

Saffer, D., McNeill, L., Araki, E., Byrne, T., Eguchi, N., Toczko, S., Takahashi, K., and the Expedition 319 Scientists, 2009. NanTroSEIZE Stage 2: NanTroSEIZE riser/riserless observatory. *IODP Preliminary Report,* 319, doi:10.2204/iodp.pr.319.2009.

Sakaguchi, A., Chester, F., Fabbri, O., Goldsby, D., Li, C., Kimura, G., Tsutsumi, A., Ujiie, K., Yamaguchi, A., and Curewitz, D., 2009. Paleo-thermal condition of the shallow mega-splay fault based on vitrinite reflectance: core analysis of IODP NanTroSEIZE stage 1. *Eos, Transactions, AGU*, **90**(52), Fall Meet. Suppl., Abstract T12A-08.

Sakuma, S., Kajiwara, T., Nakada, S., Uto, K., and Shimizu, H., 2008. Drilling and logging results of USDP-4 – penetration into the volcanic conduit of Unzen Volcano, Japan. *Journal of Volcanology and Geothermal Research*, **175**, 1–12, doi:10.1016/j.jvolgeores.2008.03.039.

Schulte, P., Alegret, A. I., Arz, J. A., Barton, P. J., Bown, P. R., Bralower, T. J., Christeson, G. L., Claeys, P., Cockell, C. S., Collins, G. S., Deutsch, A., Goldin, T. J., Goto, K., Grajales-Nishimura, J. M., Grieve, R. A. F., Gulick, S. P. S., Johnson, K. R., Kiessling, W., Koeberl, C., Kring, D. A., MacLeod, K. G., Matsui, T., Melosh, J., Montanari, A., Morgan, J. V., Neal, C. R., Nichols, D. J., Norris, R. D., Pierazzo, E., Ravizza, G., Rebolledo-Vieyra, M., Reimold, W. U., Robin, E., Salge, T., Speijer, R. P., Sweet, A. R., Urrutia-Fucugauchi, J., Vajda, V., Whalen, M. T., and Willumsen, P. S., 2010. The Chicxulub asteroid impact and mass extinction at the cretaceous-paleogene boundary. *Science*, **327**, 1214–1218, doi:10.1126/science.1177265.

Schulze, K., Kümpel, H. J., and Huenges, E., 2000. In-situ petrohydraulic parameters from tidal and barometric analysis of fluid level variations in deep wells: some results from KTB. In Bucher, K., and Stober, I. (eds.), *Hydrogeology of Crystalline Rocks*. Rotterdam: Kluwer, pp. 79–104.

Scientific Planning Working Group., 2001, Earth, oceans, and life: scientific investigation of the earth system using multiple drilling platforms and new technologies. *Integrated Ocean Drilling Program Initial Science Plan 2003–2013*, 120 pp., http:www.iodp.org/isp.

Stolper, E. M., DePaolo, D. J., and Thomas, D. M., 2009. Deep drilling into a Mantle Plume Volcano: The Hawaii Scientific Drilling Project. *Scientific Drilling*, **7**, 4–14, doi:10.2204/iodp.sd.7.02.2009.

Tadokoro, K., and Ando, M., 2002. Evidence for rapid fault healing derived from temporal changes in S wave splitting. *Geophysical Research Letters*, **29**(4), 1047, doi:10.1029/2001GL013644.

Tadokoro, K., Ando, M., and Nishigami, K., 2000. Induced earthquakes accompanying the water injection experiment at the Nojima fault zone, Japan: Seismicity and its migration. *Journal of Geophysical Research*, **105**, 6089–6104, doi:10.1029/1999JB900416.

Takahashi, H., Fercho, E., and Dallimore, S. R., 2005. Drilling and operations overview of the Mallik 2002 Production Research Well Program. In Dallimore, S. R., and Collett, T. S. (eds.), *Scientific results from the Mallik 2002 Gas Hydrate Production Research Well Program, Mackenzie Delta, Northwest Territories, Canada*. Canada: Geological Survey of Canada. Bulletin.

Tanaka, H., Fujimoto, K., Ohtani, T., and Ito, H., 2001. Structural and chemical characterization of shear zones in the freshly activated Nojima fault, Awaji Island, southwest Japan. *Journal of Geophysical Research*, **106**, 8789–8810.

Tobin, H., and Kinoshita, M., 2006. NanTroSEIZE: the IODP Nankai Trough seismogenic zone experiment. *Scientific Drilling*, **2**, 23–27, doi:10.2204/iodp.sd.2.06.2006.

Tobin, H., Vannuchi, P., and Meschede, M., 2001. Structure, inferred mechanical properties, and implications for fluid transport in the décollement zone, Costa Rica convergent margin. *Geology*, **29**, 907–910, doi:10.1130/0091-7613(2001)029<0907:SIMPAI>2.0.CO;2.

Tobin, H., Kinoshita, M., Ashi, J., Lallemant, S., Kimura, G., Screaton, E., Thu, M., Masago, H., Curewitz, D., and IODP Expeditions 314/315/316 Scientific Party, 2009a. NanTroSEIZE Stage 1 Expeditions 314, 315, and 316: first drilling program of the Nankai trough seismogenic zone experiment. *Scientific Drilling*, **8**, 4–17.

Tobin, H., Kinoshita, M., Ashi, J., Lallemant, S., Kimura, G., Screaton, E.J., Moe, K.T., Masago, H., Curewitz, D., and the Expedition 314/315/316 Scientists, 2009b. NanTroSEIZE Stage 1 expeditions: introduction and synthesis of key results. In Kinoshita, M., Tobin, H., Ashi, J., Kimura, G., Lallemant, S., Screaton, E. J., Curewitz, D., Masago, H., Moe, K. T., and the Expedition 314/315/316 Scientists (eds.), *Proceedings of the IODP, 314/315/316*, Integrated Ocean Drilling Program Management International, Washington, DC, doi:10.2204/iodp.proc.314315316.101.2009.

Tréhu, A. M., Ruppel, C., Holland, M., Dickens, G. R., Torres, M. E., Collett, T. S., Goldberg, D., Riedel, M., and Schultheiss, P., 2006. Gas hydrates in marine sediments: lessons from scientific ocean drilling. *Oceanography*, **19**(4), 124–142.

Troise, C., De Natale, G., Pingue, F., Obrizzo, F., De Martino, P., Tammaro, U., and Boschi, E., 2007. Renewed ground uplift at Campi Flegrei caldera (Italy): new insight on magmatic processes and forecast. *Geophysical Research Letters*, **34**, L03301, doi:10.1029/2006GL028545.

Underwood, M. B., Saito, S., Kubo, Y., and the Expedition 322 Scientists, 2009. NanTroSEIZE Stage 2: subduction inputs. *IODP Preliminary Report*, 322, doi:10.2204/iodp.pr.322.2009.

Wiersberg, T., and Erzinger, J., 2007. Hydrogen anomalies at seismogenic depths of the San Andreas Fault. In *Abstracts of the 17th Annual V. M. Goldschmidt Conference: Geochimica et Cosmochimica Acta*, **71**,15S, A1110.

Williams, M. L., Fischer, K. M., Freymueller, J. T., Tikoff, B., Tréhu, A. M., and others, 2010. *Unlocking the Secrets of the North American Continent: An EarthScope Science Plan for 2010–2020*, February, 2010, 78 pp.

Zhang, Z. M., Xiao, Y. L., Xu, Z. Q., Hoefs, J., Yang, J. S., Liu, F. L., Liou, J. G., and Simons, K., 2006. UHP metamorphic rocks from the Chinese continental scientific drilling project: I. petrology and geochemistry of the main hole (0–2050 m). *Contributions to Mineralogy and Petrology*, **152**, 421–441.

Zoback, M. D., and Harjes, H. P., 1997. Injection-induced earthquakes and crustal stress at 9 km depth at the KTB deep drilling site, Germany. *Journal of Geophysical Research*, **102**, 18,477–18,491.

Zoback, M. D., Hickman, S., and Ellsworth, W. S., 2010. Scientific drilling into the San Andreas Fault Zone. *Eos*, **91**(22), 197–204.

Cross-references

Deep Scientific Drilling
Earthquakes and Crustal Deformation
Geophysical Well Logging
Impact Craters on Earth

Lithosphere, Continental
Lithosphere, Oceanic
Mantle Plumes
Seismic Instrumentation
Slow Earthquake
Vertical Seismic Profiling

DEEP SEISMIC REFLECTION AND REFRACTION PROFILING

Kabir Roy Chowdhury
Department of Earth Sciences, Utrecht University, Utrecht, The Netherlands

Synonyms

Active source seismology; Controlled source seismology; Deep seismic sounding; Explosion seismology; Wide-angle reflection/refraction profiling

Definition

Multi-channel recording, along a measurement line, of seismic waves, artificially generated using large energy sources, after these have travelled deep through the earth's crust (and upper mantle).

Deep-reflection profiling is done using vibrators (on land) or air guns (in water) at near-vertical distances (8–12 km) to image the structure of the crust and upper mantle. Wide-angle reflection/refraction profiling uses large explosions and recording distances (200–300 km), primarily to obtain velocity information down to upper mantle.

Introduction

Vibrations caused by earthquakes are known to travel great distances, and provide valuable information about the internal structure of the earth. Explosions and other artificial sources could also be used to generate such waves. Man-made *earthquake* signals have several advantages of their own, e.g., their location – both in space and in time – is precisely known. Further, due to relatively shorter distances of observation, higher frequencies are better preserved, and allow a better resolution of the subsurface structure. Development of *marine seismics* has extended the coverage of this technique to offshore regions too.

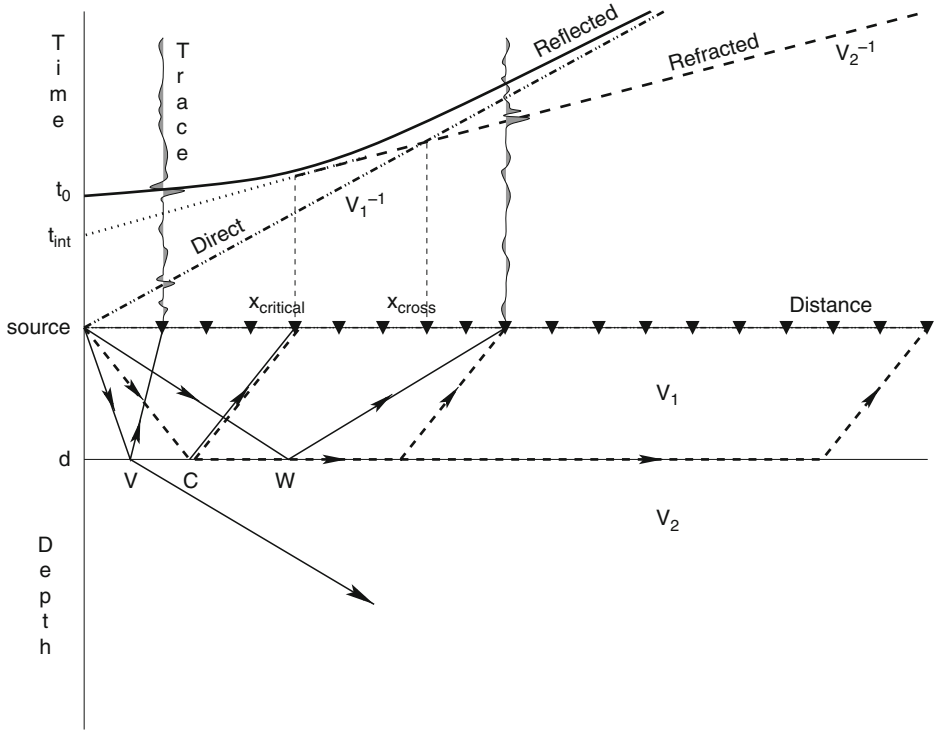
As a science, *Deep Seismics* is positioned between exploration seismics and seismological studies. The former uses man-made waves to investigate structures in the sedimentary basins (down to about 5 km) to locate likely occurrences of oil and gas, whereas the latter uses earthquake-generated waves to study earth's deep interior down to the core. With the shared theory of wave propagation in (visco)elastic media, similar instrumentation and computer-intensive processing/interpretation techniques (see *Seismic Data Acquisition and Processing* for a comparative table), there is considerable overlap between these two approaches. Improvements in data-acquisition and processing have extended the depth-range of exploration seismics – covering the entire continental crust down to *Moho* and the upper-mantle region immediately below it (50–100 km). Such studies are called by different names (deep seismic sounding, long-range reflection/refraction, wide-angle crustal-seismics, etc.) depending upon their focus and regional usage. Here, the term *DSRRP,* derived from the title, will be used to denote this developing branch of Earth Sciences, concerned with studying the (continental) crust and upper mantle of the earth, using man-made sources. The article will include studies of continental margins, but not those covering oceanic crust (See also *Earth's Structure, Continental Crust*).

There are several good resources about the general theory of wave propagation in elastic media (e.g., Aki and Richards, 2002) and exploration seismics (e.g., Sheriff and Geldart, 1995 and "*Seismic Data Acquisition and Processing*"), special issues pertaining to the latter in the context of DSRRP will be elaborated whenever needed. Below, all capitals will be used for acronyms and phrases within double quotes will refer to articles elsewhere in *this* volume.

Basic principles

Figure 1 illustrates the principle of the seismic experiment using a simplified earth model. Here, a horizontal boundary separates the top layer of velocity (actually wave-speed) v_1 and thickness d from a half-space of velocity v_2, with $v_1 < v_2$, mostly valid within the earth. After the generation of the vibration at the source, elastic energy travels through the media, and is recorded by many (usually hundreds) of receivers on the surface as *seismic traces,* the collection being a *seismic record*. Plot of such a record – amplitudes as a function of recording distance vs. travel-time – shows distinct spatio-temporal *alignments,* representing different wave-types (*phases*), three of which have been sketched in the figure. Depending upon their travel path, the phases show up in the record as *first,* or, later arrivals, and have information about the structure of the medium. In addition, the amplitudes of these arrivals – and their variation with the recording distance (*offset*) – yield information about the material properties of the medium, e.g., *density*, *elasticity*, etc.

The subject matter of this article forms an extension of exploration seismics, which is a broad field, with a substantial literature. Sheriff and Geldart (1995) and Liner (2004) provide good coverage of its various aspects, e.g., theory, data-acquisition, -processing, and -interpretation. Aki and Richards (2002) (theory of wave propagation), Vermeer (2002) (data-acquisition), and Yilmaz (2001) (data-processing) provide more focused treatments of individual topics. DSSRP-specific modifications/extensions to the standard field- and processing-procedures will be described later; see *Seismic Data*



Deep Seismic Reflection and Refraction Profiling, Figure 1 Schematics of a multi-channel seismic experiment to image the earth (layer of wave-speed v_1 overlying a half-space of wave-speed v_2). Energy travels from the source through the medium and is recorded by the receivers (filled triangles), planted in the ground, as *seismic traces*. Lines with different patterns represent typical travel paths (rays). Dash-dots indicate energy *directly* travelling from the source to the receivers along the surface, continuous lines show *reflections* from the layer boundary, and dashes represent energy *critically refracted* along the boundary. The upper-half uses the same patterns to show the *travel-times* of these three arrivals. The refracted phase is first recorded at the critical distance $x_{critical}$ and overtakes the direct wave as the first arrival from the crossover distance x_{cross} onward. Letters *v*, *c*, and *w* are points representing (near) vertical, critical, and wide-angle incidence of energy on the boundary, see text for further explanation.

Acquisition and Processing for a general introduction. Based on Figure 1, a brief overview of some key concepts from *seismics* – relevant to DSRRP – follows.

Linear arrivals represent the travel-times of both *direct* and *refracted* arrivals, if v_1 and v_2 are constant, the slopes (dt/dx) being inversely proportional to the corresponding velocities. Using these, and the *intercept time* (t_{int}, Figure 1) of the refracted arrival, the layer thickness d can be calculated.

Critical distance is the minimum recording distance for observing the refracted wave. According to Snell's law, only when the energy is incident on the boundary at the critical angle $\left(= \sin^{-1} \frac{v_1}{v_2}\right)$, will it be refracted *along* the boundary, will propagate with the velocity v_2 of the lower medium, and will keep sending energy back into the upper medium according to the Huygen's principle.

Crossover distance denotes the distance, beyond which the faster refracted phase overtakes the direct wave ($v_2 > v_1$) and becomes the first arrival; the latter can also be used to determine the thickness d. In a layered earth with velocities increasing with depth, refraction arrivals from deeper boundaries become the first arrival after their respective crossover distances. Note, that A. Mohorovicic used the overtaking of the direct P-phase by a faster (refracted) one, which he called P_n, to infer the thickness of the earth's crust and the presence of a higher velocity medium underneath.

Reflections represent that part of wave energy incident upon a layer boundary, say at the point "*V*" in Figure 1, which will be transmitted back into the upper medium, according to Snell's law. The *reflected* arrival is hyperbolic in shape, with its apex at time t_0, corresponding to the vertical two-way time ($= 2d/v_1$) from the source down to the boundary and back. Reflection amplitude is sensitive to the (contrast of) material properties across the boundary, and the incidence angle, and provides valuable additional information regarding the media. The importance of this phase for crustal imaging was realized following corresponding advances in exploration seismics.

Super-critical incidence occurs, as the recording distance increases, with the incidence angle changing in Figure 1 from (near) vertical (e.g., point "*V*") through the critical incidence (e.g., point "*C*") to super-critical (e.g., point "*W*"). Note, that the reflection hyperbola becomes asymptotic to the linear arrival belonging to the layer above,

but never crosses it. This relationship holds in a multi-layered situation, with increasing velocities, too.

Wide-angle incidence corresponds to large recording distances (compared to the target-depth), and sees a sharp increase of *reflection coefficient.* Reflections from crustal boundaries, even with small contrasts, may thus be observable in this *wide-angle* range. Also, the quantity $t_x - t_0$ associated with the reflection hyperbola (Figure 1), called normal moveout (NMO), can be measured well in this region and helps velocity estimation. The seismic signal window gets compressed though at this range due to the geometry (kinematics) of the travel-time curves; note the *squeezing* in Figure 1 (top) between the direct and the reflected branches with distance.

Modern DSRRP represents an integration of the *near-vertical* and *wide-angle* modes of seismic investigation. While near-vertical multi-channel reflection seismic (NMRS) provides powerful tools for structural imaging developed for the exploration industry, wide-angle reflection/refraction (WARR) is valuable for constraining the velocities of the crustal units and thus provides the geological framework.

Data acquisition

Seismic field measurements need three components, namely, man-made *source* to produce seismic energy (usually at the earth surface, or, within water), which then propagates within the earth, *receivers* (usually also at the surface) to pick up a part of this energy returning back after reflection/refraction/diffraction (scattering in general) from within, and *recorder* to store the received signals (nowadays after digitization), see *Seismic Data Acquisition and Processing* for details.

Sources used in DSRRP have to be much stronger than those typical in exploration, as the distances travelled by the seismic waves – both vertically, but especially horizontally – are much larger. Earlier, large explosions – several kilotons of TNT strong – were routinely set off for this purpose. Such explosions – sometimes under water – were recorded over long distances, sometimes across international boundaries (Steinhart and Meyer, 1961; György, 1972; Kaila et al., 1978). Sometimes quarry-blasts were used for crustal seismic studies, provided good arrangements could be made for the recording of the time of explosion. Calibrating and monitoring of nuclear explosions necessitated several long-distance (mostly) refraction profiles, with nuclear explosions as sources (see also *Seismic Monitoring of Nuclear Explosions*). Later, very long-range crustal seismic measurements were carried out in the Soviet Union using dedicated *peaceful nuclear explosions* (Pavlenkova and Pavelenkova, 2006). Although chemical explosions continue to be used for WARR studies, powerful hydraulic-driven mechanical vibrators on land and large compressed-air sources (air guns) under water are preferred for NMRS; these have sufficient energy to return signals back to the surface from Moho and upper mantle.

Receivers used in DSRRP are similar to those used in seismology and in exploration seismics, but need some special characteristics. Long measurement profiles imply deployments over a large distance/area, often necessitating stand-alone capabilities, especially for power supply and storage capacity, see Mereu and Kovach, 1970, in which mention is also made of an early German effort (MARS-66) using a portable four-channel analog tape recorder. Modern DSRRP receivers digitize real-time and can often be programmed to record at certain prearranged time-windows (see Texan in the references). For near-vertical land/marine deployments, standard equipment from the exploration industry is used; nowadays even these are frequently cable-free, and use radio communication for data-transfer.

Time-of-shot transfer, with high fidelity, from the source to the recorder was the weakest link in DSRRP – especially for WARR deployment, with distances of several hundred kilometers. Early experiments used short-wave radio-link, or, the parallel recording of a continually broadcasting radio transmitter. Availability of GPS-based time-signals, and high-quality internal clocks, has mitigated this problem.

Field layouts

Receiver arrays were utilized early on for long-range refraction-seismic studies for their effectiveness in recording signals from preferred directions (possible nuclear-test sites), and gave rise to permanent installations in several countries, e.g., NORSAR (Norway), GBA (India), and WKA (Australia). More recently, dynamic arrays have also been deployed. The Program for Array Seismic Studies of the Continental Lithosphere (PASSCAL), an initiative under the Incorporated Research Institutions for Seismology (IRIS), has made hundreds of identical instruments available for temporary deployment (see PASSCAL in the references) at many places around the world. Using such receivers, a 15-year static/dynamic observation of seismic signals is currently under way in the USA (see USARRAY in the references), which uses natural sources (*passive seismics*) too.

Multi-channel near-vertical land/marine studies use standard industry equipments and deployments (see *Seismic Data Acquisition and Processing* for details), the main difference being a much longer recording time (15–30 s, or longer). When using vibratory sources on land, long sweep- and listening-times are needed too. If the continental crust is covered by a shallow water layer (e.g., British Isles), marine seismic data acquisition offers twin advantages of speed and better signal-to-noise ratio, leading to better imaging.

DSS – DSRRP – onshore–offshore in the early days recorded from the source (near-vertical) outward and continued at least till the P_n-phase (refraction below Moho) was recorded as the first arrival (Gamburtsev, 1952).

Outside Soviet Union, this technique was extensively used in India starting 1972, to study the tectonic

Deep Seismic Reflection and Refraction Profiling, Table 1 Some (inter)national DSRRP efforts

Acronym	Location	Period	Remarks
COCORP	USA	1975–	Pioneered near-vertical imaging
BIRPS	UK	1981–1998	Marine-seismic imaging
DEKORP	Germany	1983–1997	+ Deep drilling (KTB)
LITHOPROBE	Canada	1984–2003	+ Multidisciplinary
ECORS	France	1983–1991	+ IFREMER (marine)
INDEPTH	China	1992–	+ US and other countries
ANCORP	Chile	1996	+ Germany and other countries
KRISP	Kenya	1985–1994	+ European and US universities
BEST	Russia	2002	+ Denmark and Poland
BABEL	Baltic Sea	1989	European groups, on-/offshore
EAGLE	Ethiopia	2003	+ European and US universities

framework of the subcontinent (Reddy et al., 1999), and was later adapted for sub-basalt (*Deccan Trap*) exploration. Many DSRRP investigations nowadays combine the two modes somewhat differently, an industry standard *common mid-point* (CMP) -profile (NMRS) is frequently interspersed with a few widely spaced explosions, observed at large distances (WARR). Variations include two-ship marine recording (see below) and simultaneous onshore–offshore recording, particularly useful for investigating continental/accretionary margins.

(Inter)national consortia in DSRRP started in 1975, when an industry crew measured a reflection (CMP) profile in Texas, using a vibratory source, with a recording time of 15 s (Finlayson, 2010b). For the next two decades, a large amount of data was collected and analyzed in several national academic–industrial collaborations: Table 1 provides a partial overview and IGCP-559 (2010b) is a good internet resource with many more details.

More recently, international DSRRP efforts are studying specific geological problems, e.g., Himalaya and Tibetan Plateau (Zhao et al., 1993), or, active continental margin in Central Andes (ANCORP working group, 2003).

Special processing and interpretation tools

Processing schemes for DSRRP data followed, at first, the standard industry scheme, even as the latter was undergoing change, e.g., transition from 2-D to 3-D coverage and true-amplitude processing. Soon, however, the special needs of processing deep seismic data were realized. As a result, new schemes were established, including modules newly developed by the industry and academia, and older data were frequently reprocessed, often resulting in *better* images (Cook and Vasudevan, 2006).

Line-drawings sketched on paper plots were used at first, to prepare DSRRP data for further processing, e.g., migration and interpretation. This is because, unlike basin-reflections, deep-crustal seismic signals are generally characterized by weak, laterally discontinuous backscattered energy – embedded in a noisy background. Lines were therefore manually drawn on the seismic section to represent more-or-less continuous alignments, which were taken to represent boundaries meaningful for interpretation. Clearly, the process was strongly subjective. This approach was also linked to the tradition of interpreting long-offset data using long refraction segments; naturally, it neglected a large part of the recorded signals with limited lateral continuity.

Coherency filter was proposed to mitigate this. Kong et al. (1985) formulated a procedure to *automatically* identify *signals* present in the seismic section. Based upon the concept of *phase-coherency,* it uses a few user-defined (tunable) parameters, to yield *repeatable* results, and can detect weak but laterally continuous signals. Some form of coherency filtering has nowadays become a part of the standard processing flow for deep-reflection data.

Statistical processing of DSRRP data, for objective high-grading using a coherency criterion, has also received attention. Hansen et al. (1988) proposed a method based upon statistical hypothesis testing, which provided some estimate of the robustness of the results, albeit at the cost of additional computation time. Vasudevan and Cook (1998) introduced the concept of *skeletonization* to delineate regions of the deep crust based upon their seismic signature, Baan (2000) included local signal statistics for high-grading.

Attempts have also been made to treat the entire deep-reflection wave-field as the backscattering from a random medium (Hurich, 1996; Hurich and Kocurko, 2000; Pullammanappallil et al., 1997), and analyze it to extract parameters describing the medium – this will be further discussed below.

Vertical- vis-a-vis horizontal-tectonics used preferentially could lead to different interpretations of the same data. DSS-profiles in György, 1972 and later literature (e.g., Kaila et al., 1979) frequently contain, e.g., intra-crustal normal faults. With the hindsight of the paradigm shift associated with the formulation of plate tectonics, some of these may need to be revisited. Gibbs (1986) illustrated this intriguing possibility of an alternate interpretation by using a DSS-section sans the earlier interpretive lines.

Main results

NMRS recordings of 12–15 s contain coherent reflected energy from down to 35–45 km, depending upon the presence/absence of sedimentary cover. Crystalline crust

appears to be much more reflective than assumed earlier, although there are transparent zones too. In general, the mature continental middle crust is less reflective (more transparent) than the lower crust – probably indicating differences in their rheology (brittle vs. ductile). Intra-crustal dipping reflective zones – of both thrust- and normal-fault types – are encountered frequently. At times, these cut through the Moho into upper mantle as *frozen* evidence for palaeo-subduction. Moho appears frequently as the termination of *diffuse* reflectivity, the boundary itself occasionally showing sharp displacement (sign of strength).

Data from WARR recordings can be modeled in terms of longer boundaries separating tectonic/velocity blocks at a regional scale (see below). These also provide occasional evidence of *sharp* offsets in crustal boundaries including Moho, the latter seems, in some cases, to be *doubled.*

The role of deeper structures – especially *deep-faults* – in controlling the evolution of shallower geology, e.g., deposition, deformation, fluid-movement, etc., is being increasingly appreciated. The knowledge gained is of economic significance, both for understanding *systems* associated with economic accumulation of hydrocarbons and minerals, and to help steer search for them.

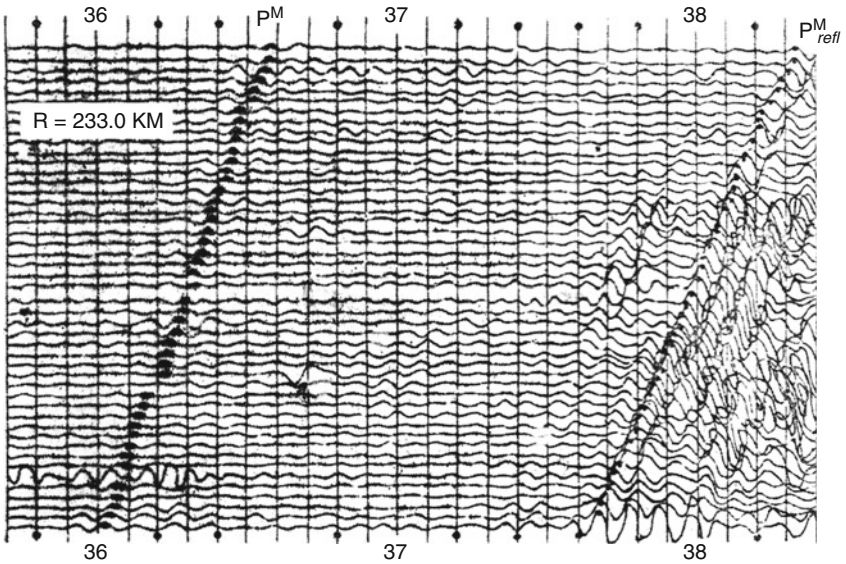
International Biennial symposia have been organized (roughly) every 2 years since 1984, to showcase the data and discuss the results from DSRRP surveys. The publications resulting from these meetings provide a historical record of the progress of DSRRP – both with respect to technological advancement and scientific knowledge (Barazangi and Brown, 1986a, b; Matthews and Smith, 1987; Leven et al., 1990; Meissner et al., 1991; Clowes and Green, 1994; White et al., 1996; Klemperer and Mooney, 1998a, b; Carbonell et al., 2000; Thybo, 2002; Davey and Jones, 2004; Snyder et al., 2006; Thybo, 2010; IGCP-559, 2010a).

(*Re*)*processing*, *synthesis*, *and interpretation* of the vast amount of DSRRP data – near-vertical and wide-angle – is not an easy task. The data quality depends mostly upon the respective geological settings and the data-acquisition technique and parameters used. Uniform processing of this dataset of mixed quality/vintage is necessary though for regional syntheses and interpretation, to understand the internal architecture of the continental crust (e.g., Phinney and Roy Chowdhury, 1989; Cook and Vasudevan, 2006).

The results from DSRRP are incorporated in several articles in this volume (please see the cross-references).

Through the years, DSRRP has made significant advances – technological in data-acquisition and methodological in processing/interpretation; its scientific results have included several surprising features – some of which are still being interpreted – and have yielded new insights into the processes that shape the continental crust. Below are a few of these highlights; the acronyms referring to the consortia/projects are explained in Table 1.

Early DSS(*RP*) *in Eastern Europe* has been nicely summarized in György (1972), from which Figure 2 is taken. It shows a part of a long-offset (≈233 km) DSS profile recorded in the Ukraine in the 1960s, as part of extensive DSS-surveys in the East-European countries mentioned earlier, which had established the observability of both



Deep Seismic Reflection and Refraction Profiling, Figure 2 DSRRP data recorded in Ukraine in the 1960s. Both refracted energy and reflected energy are clearly visible at a distance of about 233 km, and is correlatable trace-to-trace with a geophone spacing of 100 m. The first arrival, here marked P^M, is the head-wave from Moho (P_n), the strong later phase, marked P^M_{refl}, is the wide-angle reflection from the base of the crust. For further details, see György (1972, p. 50) (Courtesy *Geophysical Transactions*).

refracted- and reflected-phases in the wide-angle range. High apparent velocity of the refracted P_n-phase (first arrival at those distances), and its relationship with reflected (later) phase, help identify the base of the crust as their origin. Some figures in the above reference also show near-vertical reflections from the Moho, albeit with less clarity, not surprising in view of the older acquisition technology.

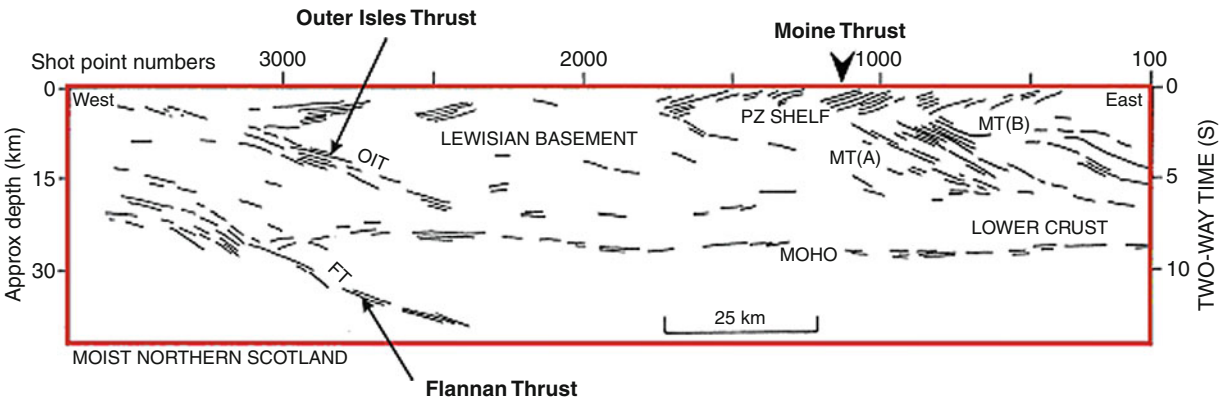
György (1972, p. 44–68) includes on page 66 a contour map of the Moho-depth below Ukraine, the detail of which, although debatable, is impressive, especially considering its vintage. It is based on an astounding 6,000 line-km of DSS profiling with a dense network, following the methodology described in Gamburtsev (1952). The contours indicate depths between 22.5 and 55 km, with rapid lateral variations at places. The latter were interpreted as *deep-faults* displacing the Moho.

Imaging deeper with multi-channel exploration seismics began in 1975. The COCORP consortium – the catchy acronym is reported to have been coined past midnight at a bar in Mexico – pioneered the use of industry standard sources (vibrators), recording layout (NMRS), and processing for deep seismic profiling on land. It was shown that under favorable geological and recording conditions, useful signals could be received back from depths of 40–50 km, by using four to five vibrators simultaneously and extending the recording time (Oliver et al., 1976). Later, similar studies confirmed that the crust underlying the basement possesses variable reflectivity, including some *transparent* regions. Moho, the base of the crust, often showed up on such images as the termination of a zone of diffuse reflectivity, and not as a long and sharp boundary inferred from earlier (refraction) studies.

One of the early surprises of the COCORP lines was the discovery of a mid-crustal zone of strong reflectivity below southern Georgia, USA; the *Surrency Bright Spot* was reconfirmed during a later survey (Pratt et al., 1993). Zones of strong reflectivity have since been reported in other surveys too, e.g., the *Quebrada Blanca Bright Spot* (QBBS) in Andean subduction zone (ANCORP working group, 2003) and the *Nyinzhong Bright Spot* (NBS) under Tibet (Makovsky and Klemperer, 1999). The possible cause(es) of this strong reflectivity remain controversial (see below).

Marine-seismic imaging of the continental crust was seized upon to (partly) alleviate the unfavorable *signal-to-noise ratio* (SNR) for deep seismic data-acquisition on land (noise from traffic, industry, etc.). Offshore surveys in waters underlain by continental crust were expected to provide better images of the latter (although marine seismics has its noise-sources too). Phinney (1986) confirmed this using data from a 48-channel marine survey (recorded by USGS during 1973–1979) over the Long Island Platform. The original stack-sections of Lines 36 and 23, processed to 12 s, showed clear evidence of a rich crustal architecture, with half-grabens, wedges, and other tectonic features indicating both the compressional- and extensional phases of a Wilson-cycle.

Existence of known and expected hydrocarbon-bearing basin structures had already attracted marine-seismic exploration activity in the 1970s and 1980s to the waters around the British Isles. The latter, surrounded by North Sea and the northeast margin of the Atlantic Ocean, are a part of the European continental shelf. Starting 1981, this situation was utilized to great advantage by the BIRPS consortium – essentially by extending the marine-seismic exploration recording time to 15 s. The very first profile, MOIST (recorded by the preceding BIRPS group), contained strong reflections from lower crust, Moho and upper mantle (see Figure 3), which could be connected



Deep Seismic Reflection and Refraction Profiling, Figure 3 Annotated line-drawing interpretation of the MOIST-profile data, showing coherent reflections from throughout the crustal column. Moho shows up as a more-or-less continuous boundary. Several dipping thrust-like features (e.g., Outer Isles Thrust) can be seen at all depth-ranges, which also include transparent zones. One dipping reflective zone (Flannan Thrust) is seen to penetrate through the Moho into the upper mantle. (Figure from Finlayson 2010a, see Brewer and Smythe 1984 for details).

to surface-geology on land (Smythe et al., 1982), and even prompted correlation of tectonic evolution across the Atlantic (Brewer and Smythe, 1984).

Later BIRPS profiles, e.g., WINCH (Brewer et al., 1983), continued to impress with very good data quality and evidence for shallower/younger tectonics being controlled by older/deep-crustal structures. The DRUM profile extended the recording to 30 s. The density of coverage and the quality of data allowed Flack and Warner (1990) to map deep reflections in 3-D, and enabled Chadwick and Pharaoh (1998, p. 268) to produce a contour map of Moho beneath the UK, which may be compared with a similar map below Ukraine (György, 1972, p. 66) mentioned earlier.

Integrated transects, using additional geophysical tools, e.g., magneto-tellurics, electro-magnetics, and geochrology, characterized the LITHOPROBE programme in Canada. It took advantage of the geology to investigate both ancient processes, e.g., assembly of continents, and modern crustal dynamics of active subduction, detachment, and imbrication. For example, the Kapuskasing uplift, one of few lower crustal exposures on earth surface, was imaged in the KSZ-transect (Percival et al., 1989), whereas the transects SNORCLE and SC examined younger mountain building processes (e.g., Clowes et al., 1983).

Relating to the ground truth, the strength of exploration seismics, is the Achilles' heel of DSRRP (and seismology). Super-deep drill holes provide the only opportunities of directly correlating observations to the rock-properties; the German DEKORP programme was able to utilize this in a symbiotic manner.

Besides WARR recordings of quarry-blasts, Germany had an early start in near-vertical recordings of deep reflections and statistical evaluation of their amplitudes (e.g., Dohr, 1957; Dohr and Fuchs, 1967). The DEKORP programme was formulated to study the evolution of the European continent through *Variscan* and later orogenic episodes. More recently, their efforts have included international investigations across active collisional zones, e.g., Alps (Gebrande et al., 2006) and the Andes (ANCORP working group, 2003).

DEKORP played an important role in the site-selection phase of the German super-deep drilling programme, KTB, which was set up to drill ≈ 10,000 m down through an ancient geodynamic suture (see Emmermann and Lauterjung (1997) for an overview of KTB). Later, besides providing a reference point for the seismics at the drill-site, the KTB programme has yielded direct evidence of shear-zones, anisotropy, and substantial amounts of fluids deep in the crust; the analysis of the latter has provided new insights into their origin and role in controlling/influencing geodynamic processes.

The Russian super-deep drilling programme in the Kola peninsula (Smythe et al., 1994) reached the record depth of 12,200 m, and also provided valuable *ground truth* regarding the macrostructure of the mature continental crust, and the origin of crustal reflectivity.

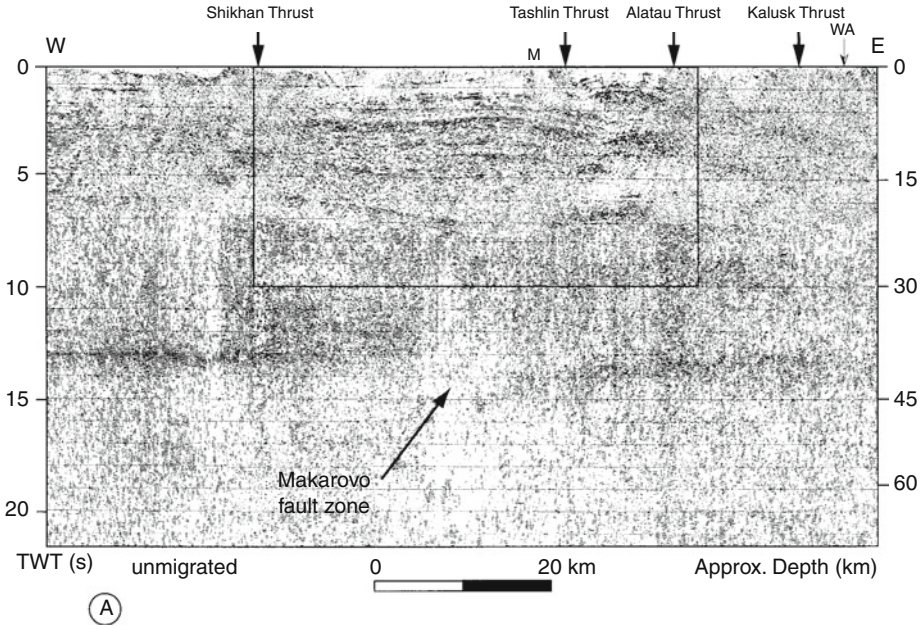
Upper-mantle reflectivity – frozen subduction has also seen a paradigm shift as a result of DSRRP investigations. Reflection signals from the upper mantle had already been reported in the wide-angle data (György, 1972). DSRRP has provided vivid proof that reflective zones may extend into the upper mantle too. Analysis of wave propagation through this region has helped us understand their structure and the possible role of scattering (see below). Dipping reflectivity in this region has provided information regarding palaeo-tectonics and may – in some cases (see Figure 3) – be evidence for palaeo-subduction. Such evidence has been reported from DSRRP experiments in different parts of the world (e.g., BABEL working group, 1990; Morgan et al., 1994).

Rheology of crust, *Moho*, *and upper mantle* depends upon their composition and in situ physical conditions. Intra-crustal faulting, affecting even the Moho, was already inferred from DSS data in Ukraine (György, 1972) and also from reflection/refraction in Southern Alberta (Kanasewich et al., 1969). DSS data from the Indian shield (Kaila et al., 1979) was used by Roy Chowdhury and Hargraves (1981) to infer constraints about the thermo-tectonic evolution of the early earth. With improved acquisition/processing, such features are now regularly reported from different parts of the world, and are no longer considered exceptions.

(Near) vertical offset in the Moho, inferred at many places from teleseismic, gravimetric, and other geophysical observations, is now often imaged in NMRS. This has important bearings regarding the *nature* of Moho. This first-order seismic boundary is often thought to be a surface that re-equilibrates after tectonic events above (e.g., thrusting), or, below (e.g., underplating). DSRRP images contain examples that contradict this view, showing that Moho topography can survive later orogenic cycles, e.g., BABEL working group (1990); Diaconescu et al. (1998) – Figure 4 is taken from the latter. The reason for this behavior is not well understood.

The thermal- and the stress-regimes in the lower crust and upper mantle determine the interaction between the two during the formation of Moho as the boundary layer. Chadwick and Pharaoh (1998) interpret a DSRRP line by associating increased reflectivity of the Moho *there* to its being a detachment surface resulting from low-angle shear (Figure 5). Local *doubling* of Moho has also been seen at places, especially in the WARR data – its evolutionary mechanism remains unclear though.

WARR had taken a back seat for a while with the increasing use of NMRS. With progress on some of the processing issues (see below), it has made a comeback though following the adage *structures from reflection and velocities from refraction.* Many recent DSRRP experiments have specifically included both the components. WARR is, sometimes, a part of onshore–offshore measurements, sometimes three-component receivers are used to study crustal anisotropy, etc. Mooney and Brocher (1987) provide a review of earlier coincident WARR studies. A more recent example is BABEL working group



Deep Seismic Reflection and Refraction Profiling, Figure 4 Unmigrated seismic section from URSEIS profile across southern Urals (Figure from Diaconescu et al., 1998, with permission from *Geology*). Moho (older than 1 billion years) is bright on the left and is offset sharply in the middle of the figure (Makarovo fault zone) by ≈5 km. Migrated image of the boxed upper part (not shown), contains laterally continuous stratification over this fault zone.

(1993). Stand-alone WARR measurements remain useful at places where continuous reflection profiling may be difficult, an example being the recent KRISP experiment in the Kenyan Rift Valley (Khan et al., 1999).

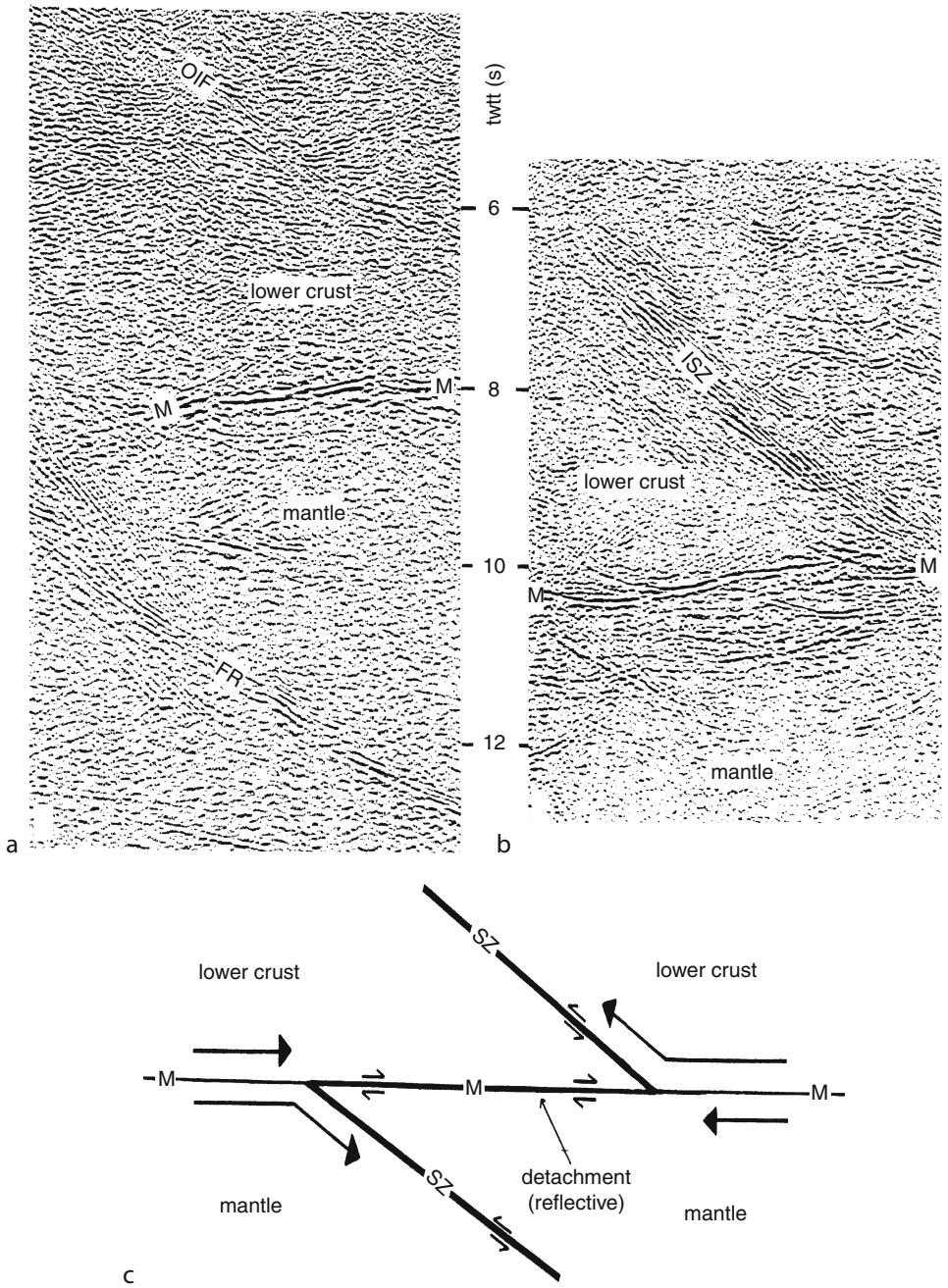
Deep seismic investigations using nuclear explosions: During 1965–1988 (before the comprehensive (nuclear) test ban treaty), the-then USSR had carried out a series of very long-range seismic experiments using *peaceful nuclear explosions* (PNEs), supplemented with chemical explosions placed every 100–150 km. (Benz et al., 1992). Employing three-component receivers, typically 10 km apart, and recording out to 1,550–3,000 km (Pavlenkova and Pavelenkova, 2006), these datasets will provide an invaluable basis for current and future research as they become widely available. QUARTZ, the profile studied most, 3,850 km long, used 400 three-component receivers, 3 PNEs, and 51 chemical explosions (Morozov et al., undated IRIS communication, see references); analysis of its data has already generated new ideas about wave propagation in lower crust and upper mantle (see below).

Analysis/Modeling of WARR data typically starts by picking *travel-times* of the phases (first and later arrivals) that are interesting, the choice being decided by the data quality and the geological aim. Usually, some prominent (mid) crustal reflected/refracted arrivals are identified and *picked,* along with arrivals from Moho, and possibly deeper ones. Methods of deriving crustal models from this data include ray-tracing (Zelt and Smith, 1992), tomography (Hole, 1992), computation of synthetics, etc. (see also *Traveltime Tomography Using Controlled-Source Seismic Data*). Frequently, a preliminary *velocity model* is iteratively fine-tuned to obtain a desired level of fit to the observed travel-times. There are several schemes for *ray-tracing,* see *Seismic, Ray Theory* and Zelt (1995) for overviews. Recently, amplitude information has also been incorporated in such schemes.

Maguire et al. (2006) contains examples of the different aspects of the procedure in the context of the EAGLE experiment (Figure 6). Different modeling/inversion schemes result, of course, in different models, even with the same input data. The DOBREfraction'99 Working Group (2003) provides an example of this, along with discussion of likely causes.

Analyzing amplitude of DSRRP signals was identified early on as a crucial step, to differentiate between competing models of the crust. Improvements in the data quality, and careful processing, allow attempts to quantify the material properties that influence the strength of the recorded deep seismic signals. Properties that are of primary interest include: reflection coefficients (RC) across various crustal boundaries, their variation with the angle of incidence (AVO), and the Q(uality) factor along the path (see *Seismic Data Acquisition and Processing* for definitions). These properties are especially important while considering the probable cause(s) of the bright spots, e.g., fluids, intrusions, layering, etc.

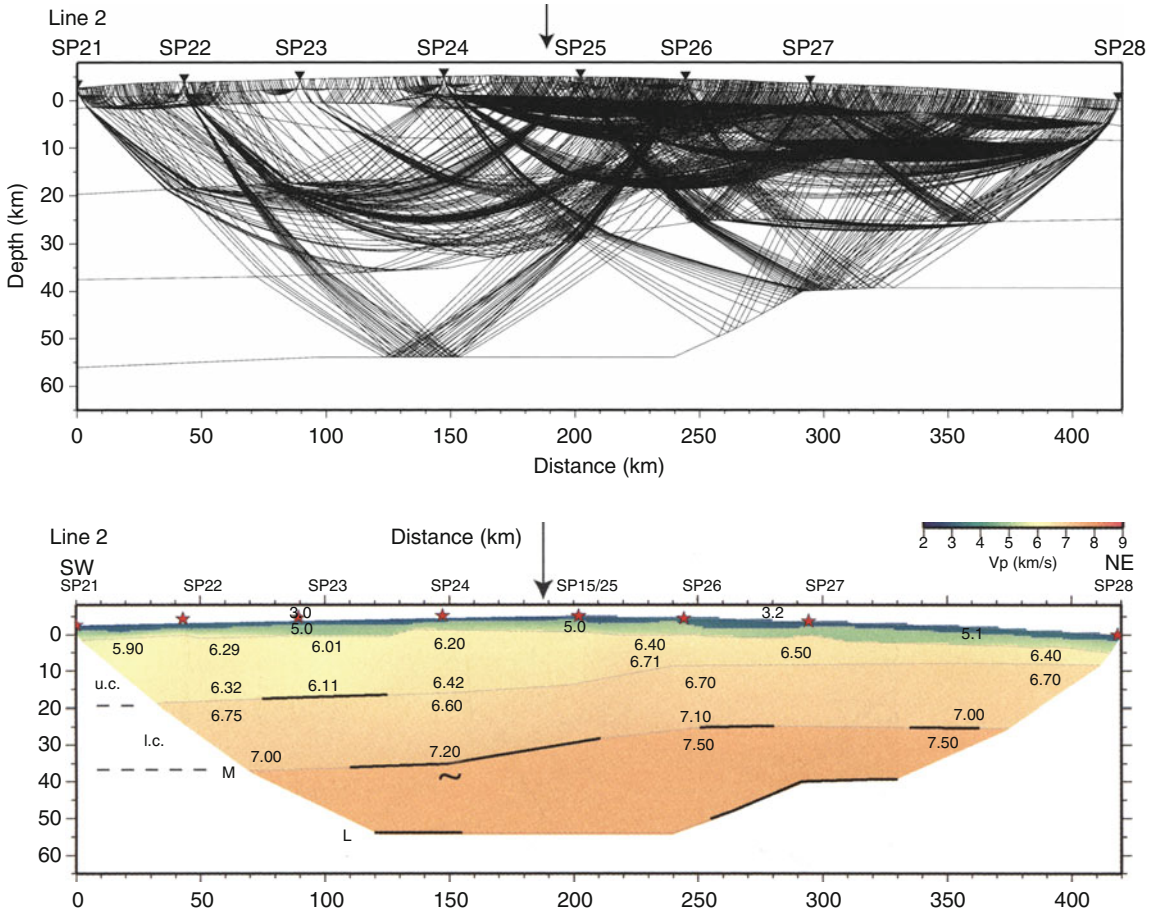
Warner (1990) reported RC values of about 0.1 for lower crustal layers and about 0.15 for Moho – both derived from the WAM profile (Peddy et al., 1989). The



Deep Seismic Reflection and Refraction Profiling, Figure 5 Schematic representation of low-angle shear near Moho (**c**) with parts of BIRPS seismic sections GRID (**a**) and NEC (**b**). For details, see Chadwick and Pharaoh (1998) (Courtesy *Tectonophysics*).

polarity (sign) of RC is important in differentiating between likely causes of bright spots sometimes seen in DSRRP images. However, even using polarity-preserving processing, it is often only possible to determine relative values of RC. ANCORP working group (2003) discusses the issues involved for a couple of strong reflectors – *Nazca and QBBS* – while reporting RC > 0.2 for the latter. Combining NMRS and WARR data, Makovsky and Klemperer (1999) report RC values between -0.3 and -0.4 for the NBS bright spot in the Tibetan middle crust.

Obtaining estimates of Q along the travel path of NMRS signals (relatively high-frequency body waves) is also difficult – the effects due to scattering, conversion,



Deep Seismic Reflection and Refraction Profiling, Figure 6 Velocity modeling by ray-tracing for the EAGLE project (in the Ethiopian rift). Note the variable coverage of the ray-tracing (*above*), and the *smooth* result, with a few long segments, in the final P-wave velocity model (*below*) – both these characteristics are typical for WARR data analysis. For details, see Maguire et al. (2006) (Courtesy *Geol. Soc. London*, with permission from the author).

etc. (apparent-Q), cannot be readily separated from the intrinsic attenuation (see also *Seismic, Viscoelastic Attenuation*). Combining the two, Hobbs (1990) obtained a value of 500 ± 200 for effective-Q for the lower crust below the WAM profile. Morozov et al. (1998) used the data from the QUARTZ profile mentioned above to obtain a 2-D Q-model for the upper mantle below this PNE-profile down to a depth of 500 km.

Research problems

The massive amount of DSRRP data collected during the past half century has not only produced many new insights, but has also brought up some (yet unsolved) problems.

Origin of crustal reflectivity, clearly visible in DSRRP images of mature continental crust, especially at greater depths, cannot be explained uniquely. Surface exposures of basement rocks mostly show steep dips; these could explain relatively transparent zones in reflection images at shallow depths. Starting at intermediate depths though, the images reveal sub-horizontal layering including strongly reflective zones. Smithson and Brown (1977, Figure 5) had already proposed a *complex* crustal model with a threefold subdivision based on geo-scientific data. DSRRP has since provided frequent evidence for both low- and high-angle intra-crustal deformation, sometimes even affecting Moho and upper mantle. The processes and the materials needed to explain these fully, remains a challenge; some of the possibilities will be briefly discussed below.

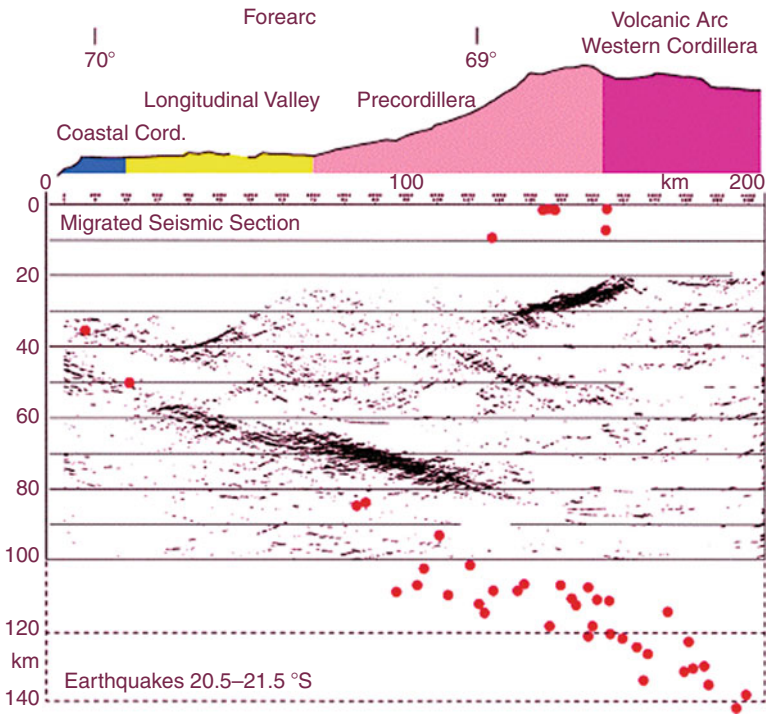
Shallow exposed structures, that are steep due to the earth's surface being stress-free, are expected to become sub-horizontal with increasing depth, and thus seismically imageable. Shear-zones, decollements, imbrications, laminae, metamorphism (facies changes, mylonitization), sill-like intrusions, and fluids (water, brine, melt, magma) have all been proposed as potential causes for crustal reflectivity, including that of Moho. Both the Kola (Smythe et al., 1994) and the KTB (Emmermann and

Lauterjung, 1997) super-deep drill holes have identified such conditions/structures at depth. Extrapolations from the drill-sites would point to the presence of crustal fluids in quantities more than that earlier expected, both in bound and free form; they undoubtedly play a major role in the tectono-thermal processes at depth (see Meissner et al. (2006) and the articles "*Earth's Structure, Continental Crust*" and "*Crustal Reflectivity (Oceanic) and Magma Chamber*" for further details).

Bright spots – zones of very strong reflectivity – present a particular challenge in this regard. Figure 7 is an interesting example of two reflective zones, probably with different, but related, origins. Nowadays, better data quality and amplitude-preserving processing often allow ruling out some causes, but may not always lead to positive conclusions. Pratt et al. (1993) investigated the Surrency bright spot (Georgia, USA), first assumed to be fluid related due to its flat nature, and concluded an (ultra)mafic body instead. Analyzing the Tibetan bright spots, Makovsky and Klemperer (1999) inferred 10% (volume) of free aqueous fluids as the cause.

Origin and nature of (*continental*) *Moho* is also not well understood, probably because this boundary – geophysically defined as a first-order transition of P-wave velocity from ≈ 6.8 to ≈ 8.2 km/s – does not always have the same evolutionary history. However, DSRRP has replaced the earlier model for (seismological) Moho, consisting of long refraction segments – more-or-less flat, by a (seismic) boundary with a complex and variable structure. At places, it seems to be the equilibrium surface established under the deviatoric stress-regime after a tectono-thermal event involving the lower crust and the upper mantle. At others, it seems to be underplated, or, overlain by sill-like intrusions, from a later magmatic episode. Elsewhere, Moho is only identifiable as the terminal zone of the diffused lower-crustal reflectivity. At many places, it exhibits enough strength though, to retain its deformation through later tectonic events; examples include *frozen subduction* (Figure 3) and *offset Moho* (Figure 4).

Role of (*multiple*) *scattering in lower crust* has attracted attention lately – both to explain observations and to relate to geological evidence. Surface outcrops of Moho (read: lower crustal and upper mantle) rocks are extremely rare in the continental setting, e.g., *Kapuskasing Uplift* (eastern Canada), Musgrave Range (central Australia), and *Ivrea-Verbano Zone, IVZ* (Northern Italy). While the first two have been studied by DSRRP (LITHOPROBE transect KSZ and Central Australian Basin transect respectively), IVZ – although sans seismics – has



Deep Seismic Reflection and Refraction Profiling, Figure 7 Migrated seismic section from the ANCORP line 2 in the Central Andean subduction zone (reproduced from http://www-app1.gfz-potsdam.de/www/pb3/dekorp/an-fig/An_line2.gif, with permission from the author). The strongly reflective zone on upper right is the *Quebrada Blanca* bright spot (QBBS), and the reflective zone dipping to the right is the Nazca reflector. Superimposed red dots show seismicity in this area of active subduction. For further details, see ANCORP working group (2003).

been extensively studied in the past by geologists, e.g., Zingg (1990) and geophysicists, e.g., Berckhemer (1969).

Recently, statistical analysis of detailed geological maps of several exposed (lower) crustal rocks have yielded the tantalizing possibility of a self-similar (fractal) description of their shapes in terms of typical *horizontal and vertical scale-lengths.* For the IVZ, a *2-D von Karman* distribution of the structure with bimodal petrology has been derived (Holliger and Levander, 1992). This raises important questions about the nature of the wave-field recorded in DSRRP experiments, and the techniques used to analyze/interpret the data. Holliger et al. (1993) computed synthetics for such a simulation of IVZ, and were able to qualitatively reproduce lower-crustal reflectivity observed in NMRS. At large distances (WARR configuration), the synthetics from the random-medium contained laterally correlatable events, which could be erroneously used for velocity analysis, migration, etc., a possibility already suggested earlier (e.g., Levander and Gibson, 1991; Emmerich et al., 1993), see also the *wide-angle squeeze* in Figure 1 between the refracted and reflected arrivals.

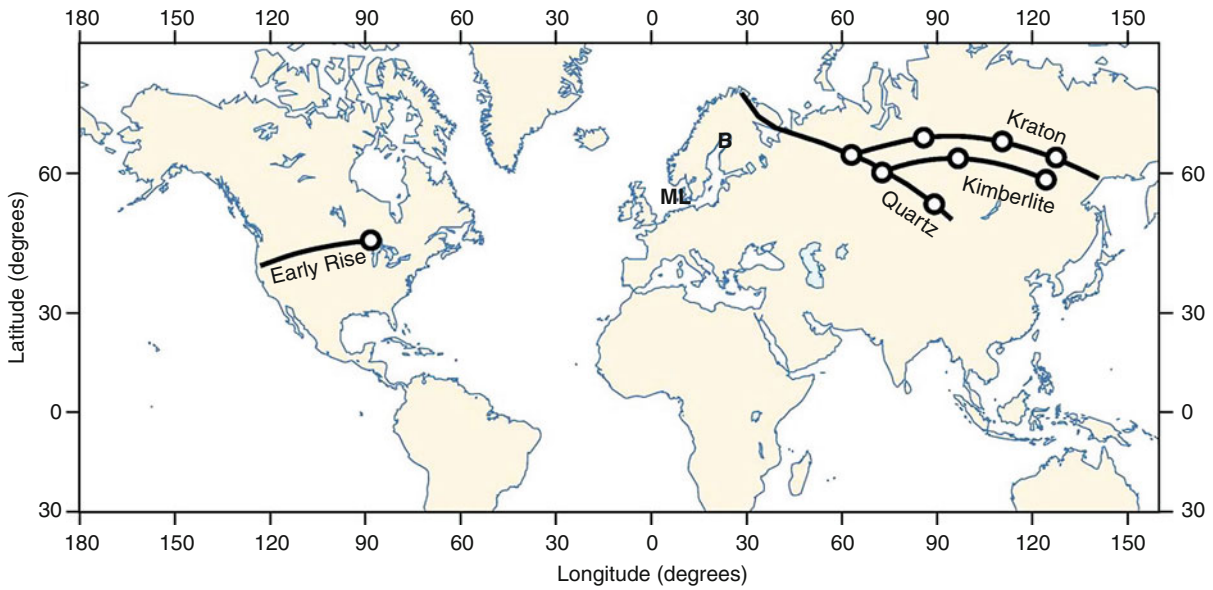
Another explanation offered for increased reflectivity in the lower crust, including transition zones, is lamellar structures with associated wave-propagation effects (amplitude, anisotropy); see Meissner et al. (2006) and "*Seismic Anisotropy*".

Both these possibilities, random heterogeneity and lamination, bring up the role of multiple scattering in the DSRRP wave-field and question the propriety of using conventional tools from basin exploration for processing such data. Douma and Roy Chowdhury (2001) used synthetics to show that multiple scattering has a limited effect for a 1-D bimodal model of IVZ, but also mentioned the need for 2-D full-wave numerical simulations.

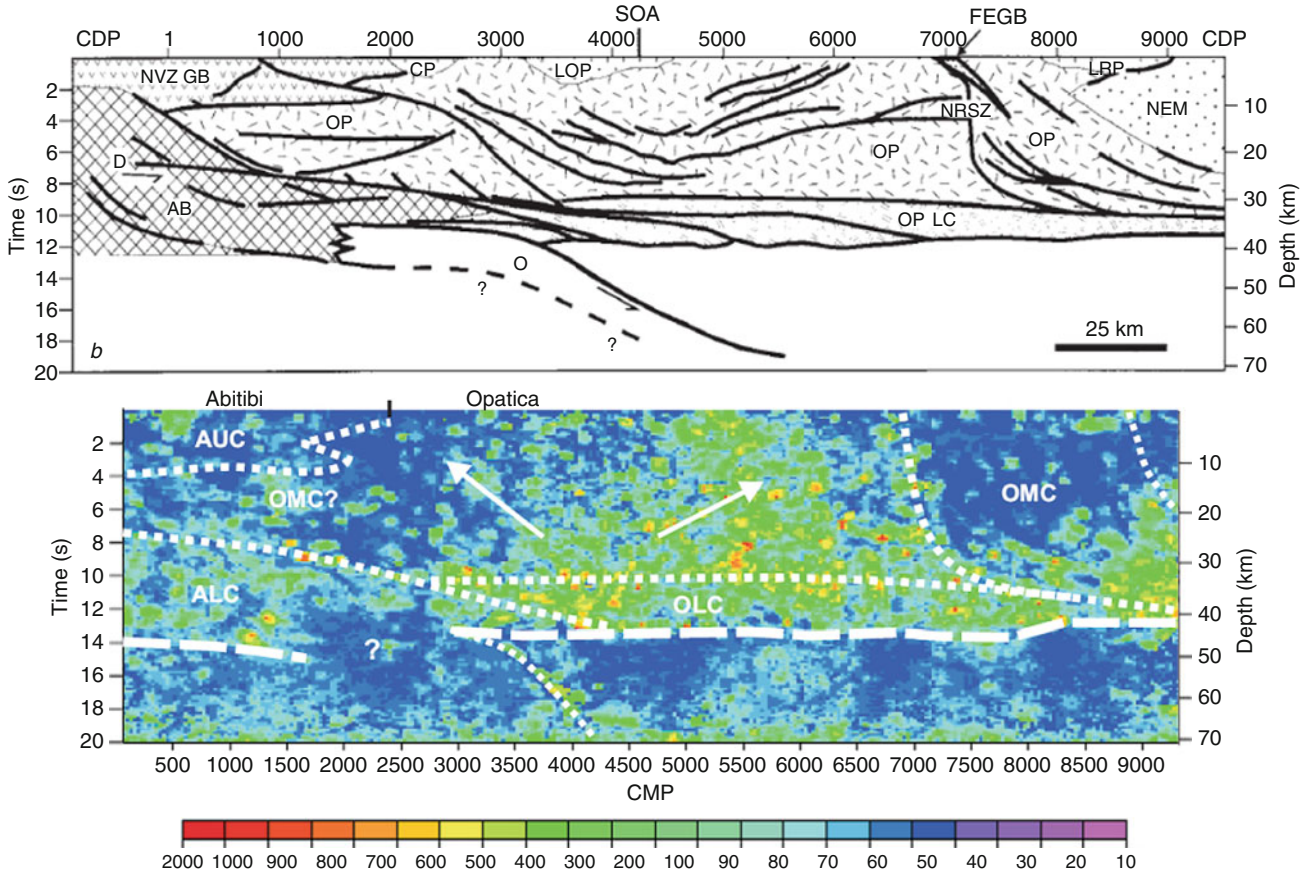
In the upper mantle too, multiple scattering seems to play an important role. Menke and Chen (1984) had invoked this to explain long-range propagation of P_n-phase from earthquakes. More recently, DSRRP data from PNE profiles, e.g., QUARTZ, has shown surprisingly strong propagation of high-frequency (5 Hz.) P_n-phase to distances of 3,000 km!

Estimating descriptive parameters of a possibly random medium in lower crust and upper mantle remains an active field of research, with two different approaches – modeling and direct estimation. To explain the long-distance P_n-phase in the QUARTZ data, Ryberg et al. (1995) modeled an upper-mantle zone of *horizontally stretched, randomly distributed velocity anisotropy.* Nielsen and Thybo (2006) modeled a larger dataset (Figure 8), and inferred random heterogeneity for both lower crust and upper mantle.

Following some earlier work (Hurich, 1996; Pullammanappallil et al., 1997; Hurich, 2003; Carpentier and Roy Chowdhury, 2007), Carpentier et al. (2010) have recently analyzed the data from Line 48 of the LITHOPROBE transect AG statistically. Assuming a 2-D *von Karman* medium, they estimated the horizontal scale-length of the medium directly from the seismic wave-field (Figure 9, below), which also shows their interpretation. The comparison with an earlier line-drawing



Deep Seismic Reflection and Refraction Profiling, Figure 8 Long-range seismic lines used for modeling by Nielsen and Thybo (2006). *Early Rise* used chemical explosives, the other lines used *PNEs.* B and ML mark the locations of NMRS lines *BABEL* and *MONA LISA* respectively (Courtesy *Tectonophysics*).



Deep Seismic Reflection and Refraction Profiling, Figure 9 *Above*: tectonic interpretation of LITHOPROBE line AG-48 (Abitibi-Grenville, Canada), taken from Calvert et al. (1995); figure courtesy *Nature*. *Below*: interpretation by Carpentier et al. (2010) overlain on their estimation of the horizontal scale-length (a_x) of the medium from the seismic data (Courtesy *Tectonophysics*).

based interpretation (above) illustrates the similarities and differences between the two approaches.

Summary

DSRRP has, over half-a-century, produced quality images of the continental crust and its margins, revealing their complex structure. These – including some unexpected results, e.g., *frozen subduction* – have contributed significantly to our ideas about the processes, current and ancient, involved in their evolution. As the deep structures, mostly inaccessible, play an important role in the development of the shallower geology, understanding these (*deep-faults*) also helps in the optimal exploration of economic resources, e.g., hydrocarbons, ore-deposits, etc., and in the study of natural hazards associated with volcanism and earthquakes.

Bibliography

Aki, K., and Richards, P. G., 2002. *Quantitative Seismology*, 2nd edn. Sausalito: University Science Books.

ANCORP working group, 2003. Seismic imaging of a convergent continental margin and plateau in the central Andes (Andean Continental Research Project 1996 (ANCORP'96)). *Journal of Geophysical Research*, **108**(B7), 2328.

Baan, M., 2000. Recognition and reconstruction of coherent energy with application to deep seismic reflection data. *Geophysics*, **65**(2), 656–667.

BABEL working group, 1990. Evidence for early Proterozoic plate tectonics from seismic reflection profiles in the Baltic shield. *Nature*, **348**, 34–38.

BABEL working group, 1993. Deep seismic reflection/refraction interpretation of critical structure along BABEL profiles A and B in the southern Baltic Sea. *Geophysical Journal International*, **112**, 325–343.

Barazangi, M., and Brown, L. (eds.), 1986a. *Reflection Seismology: A Global Perspective*. Washington, DC: American Geophysical Union. Geodynamic Series, Vol. 13.

Barazangi, M., and Brown, L. (eds.), 1986b. *Reflection Seismology: The Continental Crust*. Washington, DC: American Geophysical Union. Geodynamic Series, Vol. 14.

Benz, H., Huger, J., Leith, W., Mooney, W., Solodilov, L., Egorkin, A., and Ryaboy, V., 1992. Deep seismic sounding in northern Eurasia. *EOS, Transactions American Geophysical Union*, **73**, 297–300.

Berckhemer, H., 1969. Direct evidence for the composition of the lower crust and the Moho. *Tectonophysics*, **8**, 97–105.

Brewer, J., Matthews, D., Warner, M., Hall, J., Smythe, D., and Whittington, R., 1983. BIRPS deep seismic reflection studies of the British Caledonides. *Nature*, **305**, 206–210.

Brewer, J., and Smythe, D., 1984. MOIST and the continuity of crustal reflector geometry along the Caledonian-Appalachian orogen. *Journal of the Geological Society (London)*, **141**, 105–120.

Calvert, A., Sawyer, E., Davis, W., and Ludden, J., 1995. Archaean subduction inferred from seismic images of a mantle suture in the Superior Province. *Nature*, **375**, 670–674.

Carbonell, R., Gallart, J., and Tome, M. (eds.), 2000. *Deep Seismic Profiling of the Continents and Their Margins*. Amsterdam: Elsevier. Tectonophysics, Vol. 329.

Carpentier, S., and Roy Chowdhury, K., 2007. Underestimation of scale lengths in stochastic fields and their seismic response: a quantification exercise. *Geophysical Journal International*, **169**, 547–562.

Carpentier, S., Roy Chowdhury, K., and Hurich, C., 2010. Mapping correlation lengths of lower crustal heterogeneities together with their maximum-likelihood uncertainties. *Tectonophysics*, doi:10.1016/j.tecto.2010.07.008.

Chadwick, R., and Pharaoh, T., 1998. The seismic reflection Moho beneath the United Kingdom and adjacent areas. *Tectonophysics*, **299**, 255–279.

Clowes, R., Ellis, R., Hajnal, Z., and Jones, I., 1983. Seismic reflections from subducting lithosphere? *Nature*, **03**, 668–670.

Clowes, R., and Green, A. (eds.), 1994. *Seismic Reflection Probing of the Continents and Their Margins*. Amsterdam: Elsevier. Tectonophysics, Vol. 232.

Cook, F., and Vasudevan, K., 2006. Reprocessing and enhanced interpretation of the initial COCORP southern Appalachian traverse. *Tectonophysics*, **420**, 161–174.

Davey, F., and Jones, L. (eds.), 2004. *Special Issue – Continental Lithosphere*. Amsterdam: Elsevier. Tectonophysics, Vol. 388.

Diaconescu, C., Knapp, J., Brown, L., Steer, D., and Stiller, M., 1998. Precambrian Moho offset and tectonic stability of the East European platform from the URSEIS deep seismic profile. *Geology*, **26**, 211–214.

Dohr, G., 1957. Ein Beitrag der Reflexionsseismiek zur Erforschung des tieferen Untergrundes. *Geologischer Rundschau*, **46**, 17–26.

Dohr, G., and Fuchs, K., 1967. Statistical evaluation of deep crustal reflections in Germany. *Geophysics*, **32**(6), 951–967.

Douma, H., and Roy Chowdhury, K., 2001. Amplitude effects due to multi-scale impedance contrasts and multiple scattering: implications for Ivrea-type continental lower crust. *Geophysical Journal International*, **147**(2), 435–448.

Emmerich, H., Zwielich, J., and Müller, G., 1993. Migration of synthetic seismograms for crustal structures with random heterogeneities. *Geophysical Journal International*, **113**, 225–238.

Emmermann, R., and Lauterjung, J., 1997. The German Continental Deep Drilling Program KTB: overview and major results. *Journal of Geophysical Research*, **102**(B8), 18,179–18,201.

Finlayson, D. M., 2010a. BIRPS Startup: deep seismic profiling along the MOIST line on the continental shelf around the British Isles, 1981. Available from World Wide Web: http://www.earthscrust.org/earthscrust/science/startups/birps-su.html

Finlayson, D. M., 2010b. COCORP startup: the first deep seismic profiling of the continental crust in USA, Hardeman County, Texas, 1975. Available from World Wide Web: http://www.earthscrust.org/earthscrust/science/startups/cocorp-su.html

Flack, C., and Warner, M., 1990. Three-dimensional mapping of seismic reflections from the crust and upper mantle, northwest of Scotland. *Tectonophysics*, **173**, 469–481.

Gamburtsev, G., 1952. Deep seismic soundings of the Earth's crust. *Doklady Akademii Nauk SSSR*, **87**, 943–945.

Gebrande, H., Castellarin, A., Luschen, E., Neubauer, F., and Nicolich, R. (eds.), 2006. *TRANSALP – A Transect Through a Young Collisional Orogen*. Amsterdam: Elsevier. Tectonophysics, Vol. 414.

Gibbs, A., 1986. Seismic reflection profiles of precambrian crust: a qualitative assessment. In Barazangi, M., and Brown, L. (eds.), *Reflection Seismology: The Continental Crust*. Washington, DC: American Geophysical Union. Geodynamic Series, Vol. 14, pp. 95–106.

György, S. (ed.), 1972. *The Crustal Structure of Central and Southeastern Europe Based on the Results of Explosion Seismology*. Budapest: Hungarian Geophysical Institute Roland Eiotvois. Geophysical Transactions.

Hansen, K., Roy Chowdhury, K., and Phinney, R., 1988. The sign filter for seismic event detection. *Geophysics*, **53**(8), 1024–1033.

Hobbs, R., 1990. Effective Q determination using frequency methods on BIRPS data. *Tectonophysics*, **173**, 25–30.

Hole, J., 1992. Nonlinear high-resolution three-dimensional seismic travel time tomography. *Journal of Geophysical Research*, **97** (B5), 6553–6562.

Holliger, K., and Levander, A., 1992. A stochastic view of lower crustal fabric based on evidence from the Ivrea zone. *Geophysical Research Letters*, **19**(11), 1153–1156.

Holliger, K., Levander, A., and Goff, J., 1993. Stochastic modeling of the reflective lower crust: petrophysical and geological evidence from the Ivrea zone (northern Italy). *Journal of Geophysical Research*, **98**, 11,967–11,980.

Hurich, C., 1996. Statistical description of seismic reflection wavefields: a step towards quantitative interpretation of deep seismic reflection profiles. *Geophysical Journal International*, **125**, 719–728.

Hurich, C., 2003. The nature of crustal seismic heterogeneity: a case study from the Grenville Province. In *Heterogeneity in the Crust and Upper Mantle: Nature, Scaling, and Seismic Properties*. New York: Academic, pp. 299–320.

Hurich, C., and Kocurko, A., 2000. Statistical approaches to interpretation of seismic reflection data. *Tectonophysics*, **329**, 251–267.

IGCP-559, 2010a. International symposium on deep structure of the continents and their margins series. Available from World Wide Web: http://www.earthscrust.org/earthscrust/deep-structure-of-the-continents.htm

IGCP-559, 2010b. Seismic imaging programs. Available from World Wide Web: http://www.earthscrust.org/earthscrust/links.html

Kaila, K., Krishna, V., Roy Chowdhury, K., and Narain, H., 1978. Structure of the Kashmir Himalaya from deep seismic soundings. *Journal of the Geological Society of India*, **19**, 1–20.

Kaila, K., Roy Chowdhury, K., Reddy, P., Krishna, V., Narain, H., Subbotin, S., Sollogub, V., Chekunov, A., Kharetchko, G., Lazarenko, M., and Ilchenko, T., 1979. Crustal structure along Kavali-Udipi profile in the Indian peninsular shield from deep seismic sounding. *Journal of the Geological Society of India*, **20**, 307–333.

Kanasewich, E., Clowes, R., and McLoughan, C., 1969. A buried precambrian rift in western Canada. *Tectonophysics*, **8**, 513–527.

Khan, M., Mechie, J., Birt, C., Byrne, G., Gaciri, S., Jacob, B., Keller, G., Maguire, P., Novak, O., Nyambok, I., Pate, J., Prodehl, C., Riaroh, D., Simiyu, S., and Thybo, H., 1999. In MacNiocaill, C., and Ryan, P. (eds.), *Continental Tectonics; Special Publication*. London: Geological Society, Vol. 164, pp. 257–269.

Klemperer, S., and Mooney, W. (eds.), 1998a. *Deep Seismic Profiling of the Continents I: General Results and New Methods*. Amsterdam: Elsevier. Tectonophysics, Vol. 286.

Klemperer, S., and Mooney, W. (eds.), 1998b. *Deep Seismic Profiling of the Continents II: A Global Survey*. Amsterdam: Elsevier. Tectonophysics, Vol. 288.

Kong, S., Phinney, R., and Chowdhury, K. R., 1985. A nonlinear signal detector for enhancement of noisy seismic record sections. *Geophysics*, **50**(4), 539–550.
Levander, A., and Gibson, B., 1991. Wide-angle seismic reflections from two-dimensional random target zones. *Journal of Geophysical Research*, **96**(B6), 10,251–10,260.
Leven, J., Finlaysson, D., Wright, C., Dooley, J., and Kennet, B. (eds.), 1990. *Seismic Probing of Continents and Their Margins*. Amsterdam: Elsevier. Tectonophysics, Vol. 173.
Liner, C., 2004. *Elements of 3D Seismology*. Tulsa: PennWell.
Maguire, P., Keller, G., Klemperer, S., Mackenzie, G., Keranen, K., Harder, S., O'reilly, B., Thybo, H., Asfaw, L., Khan, M., and Amha, M., 2006. Crustal structure of the northern main Ethiopian Rift from the EAGLE controlled-source survey; a snapshot of incipient lithospheric break-up; special publication. In Yirgu, G., Ebinger, C., and Maguire, P. (eds.), *The Afar Volcanic Province within the East African Rift System*. London: Geological Society, Vol. 259, p. 269.
Makovsky, Y., and Klemperer, S., 1999. Measuring the seismic properties of Tibetan bright spots: evidence for free aqueous fluids in the Tibetan middle crust. *Journal of Geophysical Research*, **104**(B5), 10,795–10,825.
Matthews, D., and Smith, C. (eds.), 1987. *Deep Seismic Reflection Profiling of the Continental Lithosphere*. Oxford: Royal Astronomical Society. Geophysical Journal of the Royal Astronomical Society, Vol. 89.
Meissner, R., Brown, L., Dürbaum, H.-J., Franke, W., Fuchs, K., and Siefert, F. (eds.), 1991. *Continental Lithosphere: Deep Seismic Reflections*. Washington, DC: American Geophysical Union. Geodynamic Series, Vol. 22.
Meissner, R., Rabbel, W., and Kern, H., 2006. Seismic lamination and anisotropy of the lower continental crust. *Tectonophysics*, **416**, 81–99.
Menke, W., and Chen, R., 1984. Numerical studies of the coda falloff rate of multiply scattered waves in randomly layered media. *Bulletin of the Seismological Society of America*, **74**(5), 1605–1621.
Mereu, R., and Kovach, R., 1970. A portable inexpensive seismic system for crustal studies. *Bulletin of the Seismological Society of America*, **60**(5), 1607–1613.
Mooney, W., and Brocher, T., 1987. Coincident seismic reflection/refraction studies of the continental lithosphere: a global review. In *Reviews of Geophysics*. Washington, DC: American Geophysical Union, Vol. 25, pp. 723–742.
Morgan, J., Hadwin, M., Warner, M., Barton, P., and Morgan, R., 1994. The polarity of deep seismic reflections from the lithospheric mantle: evidence for a relict subduction zone. *Tectonophysics*, **232**, 319–328.
Morozov, I., Morozova, E., Smithson, S., and Solodilov, L. Long range profile Quartz.
Morozov, I., Morozova, E., Smithson, S., and Solodilov, L., 1998. 2-D image of seismic attenuation beneath the deep seismic sounding profile "Quartz," Russia. *Pure and Applied Geophysics*, **153**, 311–343.
Nielsen, L., and Thybo, H., 2006. Identification of crustal and upper mantle heterogeneity by modelling of controlled-source seismic data. *Tectonophysics*, **416**, 209–228.
Oliver, J., Dobrin, M., Kaufman, S., Meyer, R., and Phinney, R., 1976. Continuous seismic reflection profiling of the deep basement, Hardeman County, Texas. *Geological Society of America Bulletin*, **87**, 1537–1546.
PASSCAL. Program of Array Seismic Studies of Continental Lithosphere. Available from World Wide Web: http://iris.edu/hq/files/programs/passcal
Pavlenkova, G., and Pavelenkova, N., 2006. Upper mantle structure of northern Eurasia for peaceful nuclear explosion data. *Tectonophysics*, **416**, 33–52.
Peddy, C., Pinet, B., Masson, D., Scrutton, R., Sibuet, J., Warner, M., Lefort, J., Shroeder, I., BIRPS, and ECORS, 1989. Crustal structure of the Goban Spur continental margin, Northeast Atlantic, from deep seismic reflection profiling. *Journal Geological Society*, **146**, 427–437.
Percival, J., Green, A., Milkerei, B., Cook, F., Geis, W., and West, G., 1989. Seismic reflection profiles across deep continental crust exposed in the Kapuskasing uplift structure. *Nature*, **342**, 416–420.
Phinney, R., 1986. A seismic cross-section of the New England Appalachians: the orogen exposed. In Barazangi, M., and Brown, L. (eds.), *Reflection Seismology: The Continental Crust*. Washington, DC: American Geophysical Union. Geodynamic Series, Vol. 14, pp. 151–172.
Phinney, R., and Roy Chowdhury, K., 1989. Reflection seismic studies of crustal structure in the eastern United States. In Pakiser, L., and Mooney, W. (eds.), *Geophysical Framework of the Continental Unites States: GSA Memoir*. Geol. Soc. Am, Vol. 172, pp. 613–653.
Pratt, T., Mondary, J., and Brown, L., 1993. Crustal structure and deep reflector properties: wide angle shear and compressional wave studies of the midcrustal surrency bright spot beneath southeastern Georgia. *Journal of Geophysical Research*, **98**(B10), 17,723–17,735.
Pullammanappallil, S., Levander, A., and Larkin, S., 1997. Estimation of crustal stochastic parameters from seismic exploration data. *Journal of Geophysical Research*, **102**(B7), 15,269–15,286.
Reddy, P., Venkateswaralu, N., Koteswar Rao, P., and Prasad, A., 1999. Crustal structure of peninsular shield, India from DSS studies. *Current Science*, **77**, 1606–1611.
Roy Chowdhury, K., and Hargraves, R., 1981. Deep seismic soundings in India and the origin of continental crust. *Nature*, **291**(5817), 648–650.
Ryberg, T., Fuchs, K., Egorkin, A., and Solodilov, L., 1995. Observation of high-frequency teleseismic P_n on the long-range Quartz profile across northern Russia. *Journal of Geophysical Research*, **100**(B9), 18,151–18,163.
Sheriff, R. E., and Geldart, L. P., 1995. *Exploration Seismology*, 2nd edn. Cambridge: Cambridge University Press.
Smithson, S., and Brown, S., 1977. A model for lower continental crust. *Earth and Planetary Science Letters*, **35**, 134–144.
Smythe, D., Dobinson, A., McQuillin, R., Brewer, J., Matthews, D., Blundell, D., and Kelk, B., 1982. Deep structure of the Scottish Caledonides revealed by the MOIST reflection profile. *Nature*, **299**, 338–340.
Smythe, D., Smithson, S., Gillen, C., Humphreys, C., Kristoffersen, Y., Karev, N., Garipov, V., Pavlenkova, N., and The Kola-92 Working Group, 1994. Project images crust, collects seismic data in world's largest borehole. *EOS, Transactions American Geophysical Union*, **75**, 473–476.
Snyder, D., Eaton, D., and Hurich, C. (eds.), 2006. *Seismic Probing of Continents and Their Margins*. Amsterdam: Elsevier. Tectonophysics, Vol. 420.
Steinhart, J., and Meyer, R., 1961. *Explosion Studies of Continental Structure*. Washington, DC: Carnegie Institution of Washington, p. 622.
Texan. Miniature Seismic Recorder. Available from World Wide Web: http://iris.edu/hq/files/programs/passcal/docs/125A-01.pdf
The DOBREfraction'99 Working Group, 2003. "DOBREfraction'99" – velocity model of the crust and upper mantle beneath the Donbas Foldbelt (east Ukraine). *Tectonophysics*, **371**, 81–110.
Thybo, H. (ed.), 2002. *Deep Seismic Profiling of the Continents and Their Margins*. Amsterdam: Elsevier. Tectonophysics, Vol. 355.
Thybo, H. (ed.), 2010. *Deep Seismic Profiling of the Continents and Their Margins*. Amsterdam: Elsevier. Tectonophysics. Volume under preparation.

USARRAY. A continental-scale seismic observatory. Available from World Wide Web: http://www.usarray.org
Vasudevan, K., and Cook, F., 1998. Skeletons and fractals – a statistical approach to deep crustal seismic data processing and interpretation. *Tectonophysics*, **286**, 93–109.
Vermeer, G. J., 2002. *3-D Seismic Survey Design*. Tulsa: Society of Exploration Geophysicists.
Warner, M., 1990. Absolute reflection coefficients from deep seismic reflections. *Tectonophysics*, **173**, 15–23.
White, D., Ansorge, J., Bodoky, T., and Hajnal, Z. (eds.), 1996. *Seismic Reflection Probing of the Continents and Their Margins*. Amsterdam: Elsevier. Tectonophysics, Vol. 264.
Yilmaz, O., 2001. Seismic data analysis, processing, inversion and interpretation of seismic data. In *Investigations in Geophysics*, 2nd edn. Tulsa: Society of Exploration Geophysicists, Vol. I.
Zelt, C., 1995. Modelling strategies and model assessment for wide-angle seismic traveltime data. *Geophysical Journal International*, **139**, 183–204.
Zelt, C., and Smith, R., 1992. Seismic traveltime inversion for 2-D crustal velocity structure. *Geophysical Journal International*, **108**, 16–34.
Zhao, W., Nelson, K., and team, P. I., 1993. Deep seismic reflection evidence for continental underthrusting beneath southern Tibet. *Nature*, **366**, 557–559.
Zingg, A., 1990. The Ivrea crustal cross-section (northern Italy and southern Switzerland). In Salisbury, M. H., and Fountain, D. M. (eds.), *Exposed Cross Sections of the Continental Crust*. Dordrecht: Kluwer, pp. 1–19.

Cross-references

Continental Rifts
Crustal Reflectivity (Oceanic) and Magma Chamber
Earth's Structure, Continental Crust
Earth's Structure, Upper Mantle
Lithosphere, Continental
Plate Tectonics, Precambrian
Seismic Anisotropy
Seismic Data Acquisition and Processing
Seismic Imaging, Overview
Seismic Instrumentation
Seismic Monitoring of Nuclear Explosions
Seismic Waves, Scattering
Seismic, Ray Theory
Seismic, Viscoelastic Attenuation
Traveltime Tomography Using Controlled-Source Seismic Data

DIFFERENTIAL ROTATION OF THE EARTH'S INNER CORE

Xiaodong Song
Department of Geology, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Definition

Differential rotation of the inner core refers to the difference between the rotation of the Earth's solid inner core and the daily rotation of the Earth's crust and the mantle.

Background

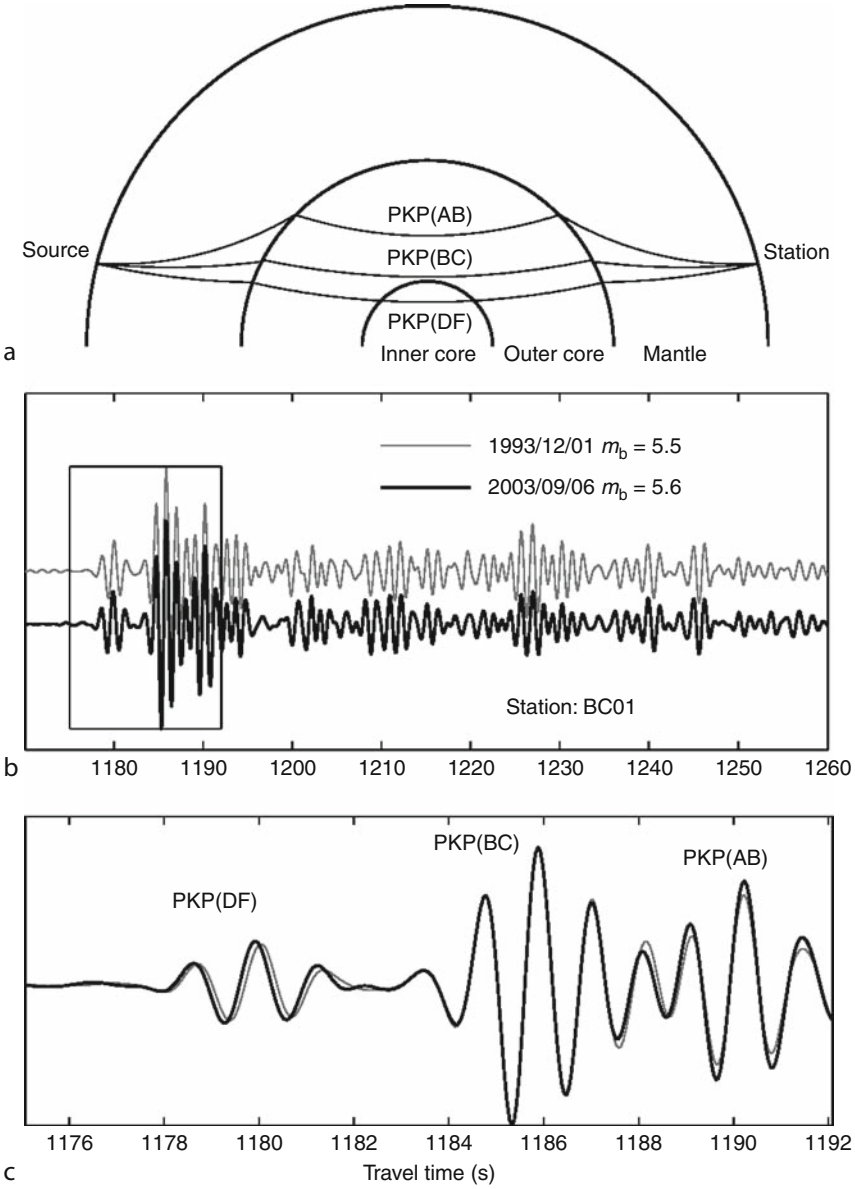
Driven by gravity, the Earth differentiated into iron core and rocky mantle at the early stage of its formation. The central core is liquid because of high temperature; but at the center of the Earth, an inner core is formed and gradually grows as the Earth cools, and the liquid iron freezes under tremendous pressure (3 million times the atmospheric pressure). The size of the inner core is slightly smaller than that of the moon. Separated from the solid mantle by the fluid outer core, the inner core is thus free to rotate independently under an external torque.

The early idea of inner core rotation came from studies of the geodynamo, which generated the Earth's magnetic field. Gubbins (1981) first suggested that electromagnetic forces between the electrically conducting inner core and the magnetic field generated in the fluid outer core would cause a differential rotation of the inner core. In a computer simulation of a three-dimensional, self-consistent geodynamo, Glatzmaier and Roberts (1995) predicted that the inner core is driven to rotate by electromagnetic coupling at a few degrees per year faster than the mantle and the crust.

Seismological observations

Song and Richards (1996) reported first evidence for differential rotation of inner core from seismic observations. The basic idea of their method is simple. They chose a fixed monitoring station and compare seismic waves that travel through the inner core from earthquakes at the same region but a few years apart. This is like the twinkling of a star, which is caused by disturbance of the light as it passes through the atmosphere (an analogy used by coauthor Paul Richards). The seismic waves used are short-period (about 1 s) longitudinal waves (P waves) that have propagated through the fluid core (generically called PKP waves) (Figure 1). At the distance range of about 146–154°, there are three possible paths, i.e., PKP(DF) waves that traverse the inner core, PKP(BC) waves that turn at the bottom of the outer core, and PKP(AB) waves that turn at the middle of the outer core. The basic data used in detecting the inner core rotation is the time difference between the PKP(BC) and PKP(DF) arrivals, abbreviated as the BC-DF time. Because these two waves travel very closely together in the crust and mantle and in most of the fluid core, the differential time BC-DF is relatively insensitive to uncertainty in source location and to three-dimensional heterogeneities of the crust and mantle along the ray paths. To remove influence of small difference in earthquake location and epicentral distance, the differential time residual is formed by subtracting the observed BC-DF time from the predicted difference for a standard earth model. In addition, residuals of the differential times between PKP(AB) and PKP(BC), which pass through the fluid core only, are used to examine possible systematic biases in source locations.

Song and Richards found that BC-DF residuals along certain paths have changed systematically with time over a few decades. In particular, the seismic waves that emanate from earthquakes in South Sandwich Islands in the South Atlantic travel through the inner core and reach



Differential Rotation of the Earth's Inner Core, Figure 1 Ray paths of PKP waves and example of waveform doublet used to detect temporal change of travel times through the inner core. (**a**) Ray paths of three branches of PKP waves turning in the solid inner core (DF), the bottom of the fluid outer core (BC), and the mid-outer core (AB). (**b**) Highly similar waveforms recorded at one station in Alaska from a waveform doublet from South Sandwich Islands. The two events are 10 years apart, one in 1993 and the other in 2003. (**c**) Superimposed and enlarged PKP waveforms from the box in (**b**). The waves through the outer core (BC and AB) are aligned, but the wave through the inner core (DF) shows a small time shift (about 0.1 s). (From Zhang et al., 2005).

College, Alaska seismic station. They have sped up systematically by about 0.3 s over 28 year-year time interval from 1967–1995. They interpreted the temporal change as evidence of an inner core rotation.

The estimate of the rotation rate depends on the underlying aspherical structure of the inner core that results the temporal change from a rotation. The inner core is known to be very anisotropic (i.e., seismic waves traveling through the inner core along different directions have different speeds) (Morelli et al., 1986; Woodhouse et al., 1986; see also review by Song, 1997; Tromp, 2001). The anisotropy is roughly axis-symmetric with the P-wave speed faster by some 3% along the Earth's rotation axis than along the equatorial plane. However, the anisotropy changes significantly laterally and with depth. The observed temporal change in BC-DF residuals was first interpreted as a change of the orientation of the fast axis of the inner core anisotropy, which yielded a rate of

1.1° per year (Song and Richards, 1996), but it was later and preferably interpreted as a shift of lateral velocity gradient in the inner core caused by the inner core rotation (Creager, 1997; Song, 2000), which yielded a rate ranging from 0.2–0.3 to 0.3–1.1° per year.

Subsequent studies provided further support and most estimates of the rotation rate are a few tenths of a degree per year faster than the rotation of the Earth (a super-rotation) (see reviews by Tromp, 2001; Song, 2003; and Souriau et al., 2003). The methods used include BC-DF times for additional paths (Ovchinnikov et al., 1998; Song and Li, 2000; Xu and Song, 2003), inner core scattering waves (Vidale et al., 2000), so-called earthquake waveform doublets (Li and Richards, 2003; Zhang et al., 2005; see below), and normal modes (Laske and Masters, 2003). However, whether the inner core rotation is real has also been hotly debated because of potential biases such as earthquake mislocation (e.g., Poupinet et al., 2000), failure to detect the motion (Souriau, 1998), and discrepancy in the inferred rotation rate (e.g., Creager, 1997; Vidale et al., 2000; Laske and Masters, 2003).

The most important source of errors is systematic event mislocation. Because global stations used to locate the earthquakes are not exactly the same in different years, the temporal variation could potentially be an artifact of the changes of the station distributions (Poupinet et al., 2000; Souriau et al., 2003). To resolve the debate, the two sides (Dr. Poupinet and the author) worked directly in 2004–2007. A thorough presentation of the collaborative work was published (Sun et al., 2006). The basic conclusion is that earthquake mislocation is too small to explain the observed time shifts in the data, and that inner core rotation is still the best explanation.

Perhaps the strongest support comes from studies of earthquake waveform doublets (Zhang et al., 2005). A waveform doublet is a pair of earthquakes occurring at different times but at essentially the same spatial position, as evidenced by their highly similar waveforms at each station recording both events (Poupinet et al., 1984). Zhang et al. (2005) observed that the waves that traveled through the area outside the inner core, in the crust, the mantle, and the outer core are all the same. Only when they travel through the inner core at different times are they different (Figure 1). The changes are two folds: first, and most prominently, the PKP(DF) waves (passing through the inner core) move faster in a systematic pattern over time, by about one tenth of a second per decade; second, the PKP(DF) waves themselves change in shape over time, an independent signature for motion of the inner core. The fact that the wave shapes are similar when they emanate from the source region allows them to measure precisely the small time shifts and to pin down precisely where the changes took place. The temporal change of BC-DF differential times for the South Sandwich Islands to College, Alaska path is constrained to be 0.0092 s/year with standard deviation of 0.0004 s/year. The best estimate of the rotation rate is 0.3–0.5° per year faster than the Earth's mantle. Subsequent studies of earthquake doublets suggest that the waves bouncing off the inner core boundary also shows temporal variation (Wen, 2006; Cao et al., 2007; Song and Dai, 2008).

Concluding remarks

The discovery of the inner core super-rotation has attracted attention of the academia and popular media. It has implications for understanding the geodynamo and angular momentum transfer in the interior of the Earth (Buffett and Glatzmaier, 2000). The origin of the Earth's magnetic field has been regarded as one of the great unsolved problems in modern physics. The observation of the inner core rotation provides a unique observational constraint on the geodynamo at the center of the earth.

The existence of the differential rotation of the inner core has now generally, although not universally, been accepted. However, major questions remain. What is the acceptable range of the rotation rate? Is the inner core rotation variable? Does the rotation change direction? Does the inner core oscillate within a certain range? The inner core possesses large hemispheric scale variations (Tanaka and Hamaguchi, 1997; Niu and Wen, 2001). This is difficult to reconcile with an inner core of a constant rotation, which may be expected to average out the lateral varying structure of the inner core over geological time as the inner core grows from the crystallization of liquid iron. If the rotation is variable or oscillating, what is the time scale? What is the driving force? How does it relate to the geodynamo processes? Recently, Lindner et al. (2010) proposed a nonparametric modeling method, which is able to separate mantle structure, inner core structure, and inner core motion. The study suggests an average rate of inner-core rotation of about 0.39° per year and that the rotation has accelerated from about 0.24° to about 0.56° per year within the last 55 years. The speed of the differential rotation, averaging about 10 km per year at the inner core equator, is by 50,000 times the fastest motion of the tectonic plates (20 cm per year) at the Earth's surface. The minimum torque acting on the inner core is estimated 1.19×10^{16} N m, which could easily result from the imbalance of much larger electromagnetic and gravitational torques. With the continuing accumulation of high quality seismic data, I believe we'll reach a new understanding on the inner core motion within next 50 years.

Bibliography

Buffett, B. A., and Glatzmaier, G. A., 2000. Gravitational braking of inner-core rotation in geodynamo simulations. *Geophysical Research Letters*, **27**, 3125–3128.

Cao, A. M., Masson, Y., and Romanowicz, B., 2007. Short wavelength topography on the inner-core boundary. *Proceedings of the National Academy of Sciences USA*, **104**(1), 31–35.

Creager, K. C., 1997. Inner core rotation rate from small-scale heterogeneity and time-varying travel times. *Science*, **278**, 1284–1288.

Glatzmaier, G. A., and Roberts, P. H., 1995. A three-dimensional convective dynamo solution with rotating and finitely conducting inner core and mantle. *Physics of the Earth and Planetary Interiors*, **91**, 63–75.

Gubbins, D., 1981. Rotation of the inner core. *Journal of Geophysical Research*, **86**, 11695–11699.
Laske, G., and Masters, G., 2003. The Earth's free oscillations and the differential rotation of the inner core. In Dehant, V., et al. (eds.), *Earth's Core: Dynamics, Structure, Rotation*. Washington, D. C: American Geophysical Union. Geodynamics Series, **31**, pp. 5–21.
Li, A., and Richards, P. G., 2003. Using earthquake doublets to study inner core rotation and seismicity catalog precision. *Geochemistry, Geophysics, Geosystems*, **4**(9), 1072, doi:10.1029/2002GC000379.
Lindner, D., Song, X., Ma, P., and Christensen, D. H., 2010. Inner core rotation and its variability from nonparametric modeling. *Journal of Geophysical Research*, **115**, B04307, doi:10.1029/2009JB006294.
Morelli, A., Dziewonski, A. M., and Woodhouse, J. H., 1986. Anisotropy of the inner core inferred from PKIKP travel times. *Geophysical Research Letters*, **13**, 1545–1548.
Niu, F. L., and Wen, L. X., 2001. Hemispherical variations in seismic velocity at the top of the Earth's inner core. *Nature*, **410**, 1081–1084.
Ovchinnikov, V. M., Adushkin, V. V., and An, V. A., 1998. On the velocity of differential rotation of the Earth's inner core. *Doklady Akademii Nauk*, **362**, 683–686.
Poupinet, G., Ellsworth, W. L., and Frechet, J., 1984. Monitoring velocity variations in the crust using earthquake doublets: an application to the Calaveras Fault, California. *Journal of Geophysical Research*, **89**, 5719–5731.
Poupinet, G., Souriau, A., and Coutant, O., 2000. The existence of an inner core super-rotation questioned by teleseismic doublets. *Physics of the Earth and Planetary Interiors*, **118**, 77–88.
Song, X. D., 1997. Anisotropy of the earth's inner core. *Reviews of Geophysics*, **35**, 297–313.
Song, X. D., 2000. Joint inversion for inner core rotation, inner core anisotropy, and mantle heterogeneity. *Journal of Geophysical Research*, **105**, 7931–7943.
Song, X. D., 2003. Three-dimensional structure and differential rotation of the inner core. In Dehant, V. M., Creager, K. C., Zatman, S., and Karato, S. (eds.), *Earth Core: Dynamics, Structure, Rotation*. Washington: American Geophys. Union, Geodynamics Series, Vol. 31, pp. 45–63.
Song, X. D., and Dai, W., 2008. Topography of Earth's inner core boundary from high-quality waveform doublets. *Geophysical Journal International*, **175**, 386–399, doi:10.1111/j.1365-246X.2008.03909.x.
Song, X. D., and Li, A. Y., 2000. Support for differential inner core superrotation from earthquakes in Alaska recorded at South Pole station. *Journal of Geophysical Research*, **105**, 623–630.
Song, X. D., and Richards, P. G., 1996. Seismological evidence for differential rotation of the Earth's inner core. *Nature*, **382**, 221–224.
Souriau, A., 1998. New seismological constraints on differential rotation of the inner core from Novaya Zemlya events recorded at DRV, Antarctica. *Geophysical Journal International*, **134**, F1–F5.
Souriau, A., Garcia, R., and Poupinet, G., 2003. The seismological picture of the inner core: structure and rotation. *Comptes Rendus Geoscience*, **335**, 51–63.
Sun, X. L., Poupinet, G., and Song, X. D., 2006. Examination of systematic mislocation of South Sandwich Islands earthquakes using station pairs: Implications for inner core rotation. *Journal of Geophysical Research*, **111**, B11305, doi:10.1029/2005JB004175.
Tanaka, S., and Hamaguchi, H., 1997. Degree one heterogeneity and hemispherical variation of anisotropy in the inner core from PKP(BC)-PKP(DF) times. *Journal of Geophysical Research*, **102**, 2925–2938.
Tromp, J., 2001. Inner-core anisotropy and rotation. *Annual Review of Earth and Planetary Sciences*, **29**, 47–69.
Vidale, J. E., Dodge, D. A., and Earle, P. S., 2000. Slow differential rotation of the Earth's inner core indicated by temporal changes in scattering. *Nature*, **405**, 445–448.
Wen, L., 2006. Localized temporal change of the Earth's inner core boundary. *Science*, **314**, 967–970.
Woodhouse, J. H., Giardini, D., and Li, X.-D., 1986. Evidence for inner core anisotropy from free oscillations. *Geophysical Research Letters*, **13**, 1549–1552.
Xu, X. X., and Song, X. D., 2003. Evidence for inner core super-rotation from time-dependent differential PKP travel times observed at Beijing Seismic Network. *Geophysical Journal International*, **152**, 509–514.
Zhang, J., Song, X. D., Li, Y. C., Richards, P. G., Sun, X. L., and Waldhauser, F., 2005. Inner core differential motion confirmed by earthquake waveform doublets. *Science*, **309**, 1357–1360.

Cross-references

Body Waves
Core Dynamo
Core-Mantle Coupling
Earth Rotation
Earth, Density Distribution
Earth's Structure, Core
Earth's Structure, Global
Earthquake, Location Techniques
Geodynamics
Geomagnetic Field, Global Pattern
Geomagnetic Field, Theory
Mantle Convection
Propagation of Elastic Waves: Fundamentals
Radioactivity in Earth's Core
Seismic Anisotropy
Seismic Properties of Rocks

E

EARTH ROTATION

Harald Schuh, Sigrid Böhm
Research Group Advanced Geodesy, Institute of Geodesy and Geophysics, Vienna University of Technology, Vienna, Austria

Definition

Earth. Solid Earth including oceans and atmosphere.
Earth rotation. Temporal variation of the orientation and the rotation speed of the Earth.

Introduction

The rotation of the Earth or Earth rotation, respectively, specifies the spatiotemporal change of the Earth rotation vector. The direction of the Earth rotation vector corresponds to the instantaneous rotation axis of the Earth and its absolute value equals the rotation speed. The Earth's rotation is not uniform and is given in terms of Earth orientation parameters (EOP): *precession* and *nutation* are long-term and periodic changes of the direction of the Earth rotation vector with respect to a space-fixed reference system. *Polar motion* is the variation of the direction of the Earth rotation vector with respect to an Earth-fixed reference system (qv *Geodesy, Networks and Reference Systems*). Changes in the Earth rotation speed are expressed as deviations of *Universal Time 1* (UT1) from the uniform atomic time (*Universal Time Coordinated*, UTC) *dUT1* = UT1−UTC or as variations in the *length of day* (LOD). The subgroup of *polar motion* and *dUT1* or LOD is called Earth rotation parameters (ERP). Fundamental information on Earth rotation theory and observation and about the relations of Earth rotation variations with geophysical processes is given in the seminal works of Munk and MacDonald (1960), Lambeck (1980), and Moritz and Mueller (1987).

While the existence of *precession* was already known to the Greek astronomer Hipparch in the second century before Christ, *nutation* was not discovered before the eighteenth century by James Bradley. Observations of *polar motion* were taken for the first time by the Bonn astronomer Friedrich Küstner at the end of the nineteenth century by measuring latitude variations. From the seventies of the twentieth century onward, space geodetic techniques like *Very Long Baseline Interferometry* (VLBI) (qv *Very Long Baseline Interferometry (VLBI)*), *Satellite Laser Ranging* (SLR) (qv *Satellite Laser Ranging*), *Lunar Laser Ranging* (LLR), *Doppler Orbitography and Radiopositioning Integrated by Satellite* (DORIS), and the *Global Navigation Satellite Systems* (GNSS) like the *Global Positioning System* (GPS) (qv *GPS, Data Acquisition and Analysis*) have been employed in addition to astronomical methods. Approximately since the year 2000, the latter are no longer routinely applied. Nowadays, the achievable accuracies of the measurements are better than 2×10^{-4} arcsec or less than 0.6 cm, if projected to the Earth surface. EOP can be determined from space geodetic observations within the respective parameter estimation process. The transformation between Earth-fixed and space-fixed system and thus the EOP are thereby components of the technique's observation equation and can be solved for as unknowns. The most precise technique to observe polar motion is GNSS, whereas *nutation* and *dUT1* can be measured directly only by VLBI. Due to the correlation of *nutation* and *dUT1* with the orbital elements of the satellites, satellite techniques are only sensitive to the time derivation of those parameters i.e., *nutation rates* and LOD. Promising new devices for the observation of high frequency variations of the instantaneous Earth rotation vector are large ring laser gyroscopes. These instruments can access absolute rotation by measuring the beat frequency of two laser beams rotating in opposite direction, the *Sagnac frequency* (Schreiber

Harsh K. Gupta (ed.), *Encyclopedia of Solid Earth Geophysics*, DOI 10.1007/978-90-481-8702-7,

et al., 2004). Information about the long-term behavior of the EOP is obtained from historical records of lunar and solar eclipses and from the analysis of sedimentary deposition. Evidence for a secular increase in LOD can be found for instance in paleontological studies of coral growth rate.

The observed EOP show a wide spectrum of variations. Its interpretation allows for drawing valuable conclusions about the structure of the Earth and the dynamical features of the Earth's interior, the atmosphere, hydrosphere, and cryosphere. Even anthropogenic influences on Earth rotation such as the impact of mass transfer due to an increased CO_2 emission are subject of scientific investigations (de Viron et al., 2002).

Medium and long-period as well as long-term mass displacements affect Earth rotation (cf. e.g., Gross (2007) or Seitz and Schuh (2010) for a comprehensive compilation and further references). According to the conservation of total angular momentum of the Earth in short terms, the rotation of the solid Earth undergoes variations, which are mirror image to changes in atmosphere and oceans. Any mass variation in one or more components of the system Earth which changes the Earth inertia tensor leads to a corresponding variation in the Earth rotation. One can also think of a mass movement which does not change the inertia tensor, because once a mass element moved away it is replaced by another, like in a circular ocean current. Such a mass transport causes a motion relative to the considered reference frame which affects Earth rotation as well. In which way changes of the inertia tensor and relative motions can be related to variations in Earth rotation is shown in the successive section.

Mathematical formulation of Earth rotation

The dynamical equation of motion of a rotating body with respect to a space-fixed system is given by

$$\boldsymbol{L} = \frac{\mathrm{d}\boldsymbol{H}}{\mathrm{d}t} \tag{1}$$

relating the torque L acting on the body to the temporal change of its angular momentum H. This is the basic relation for the development of (Newtonian) *precession/nutation* theories, since it describes the motion of the rotating body in a space-fixed reference system. To characterize *polar motion* and changes of the rotation rate of a rotating body, Equation 1 has to be referred to body-fixed axes, rotating with the angular velocity ω:

$$\boldsymbol{L} = \frac{\mathrm{d}\boldsymbol{H}}{\mathrm{d}t} + \boldsymbol{\omega} \times \boldsymbol{H} \tag{2}$$

This is one form of *Euler's dynamical equations* for rigid body rotation referred to a body-fixed coordinate system. The angular momentum H can be expressed as the product of the tensor of inertia I and the angular velocity vector ω. The inertia tensor is a symmetric matrix containing the moments of inertia and the products of inertia of a rotating body and thus characterizing the mass distribution in the body. Since single particles do not move with respect to the body-fixed system, this tensor is invariant in the case of a rigid body. For a nonrigid (deformable) body, the inertia tensor becomes time variable and the particles can move with respect to the body-frame, thus allowing for relative motions which introduce relative angular momentum. The angular momentum H of a rotating deformable body is then written as

$$\boldsymbol{H} = \boldsymbol{I}\boldsymbol{\omega} + \boldsymbol{h} \tag{3}$$

with h denoting relative angular momentum. The first summand is often referred to as *mass* or *matter term* and the second is called *motion term*. As the Earth is a nonrigid body, the equation of rotational motion 2 has to be extended considering the above stated differences to the motion of a rigid body, leading to

$$\boldsymbol{L} = \frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{I}\boldsymbol{\omega} + \boldsymbol{h}) + \boldsymbol{\omega} \times (\boldsymbol{I}\boldsymbol{\omega} + \boldsymbol{h}) \tag{4}$$

These are the *Euler–Liouville equations* or simply *Liouville equations*. The deviations from uniform, rigid rotation are formulated as follows for the rotation vector $\boldsymbol{\omega}$:

$$\omega_1 = \Omega m_1,\ \omega_2 = \Omega m_2,\ \omega_3 = \Omega(1 + m_3) \tag{5}$$

where Ω is the mean angular velocity of the Earth. The m_i are small dimensionless quantities describing the excursions of the rotation vector from its uniform rotation due to time-variable changes in the mass distribution of the Earth system and relative particle motion. Assuming the body-fixed axes to be principal axes of inertia and the body to be rotationally symmetric, these mass changes can be taken into consideration by composing the tensor of inertia from constant and time-variable parts:

$$\begin{aligned} &I_{11} = A + \Delta I_{11} \quad I_{22} = A + \Delta I_{22} \quad I_{33} = C + \Delta I_{33}, \\ &I_{ij} = \Delta I_{ij}, \qquad\qquad i \neq j \end{aligned} \tag{6}$$

where the constant parts are the polar moment of inertia C and the equatorial moment of inertia A of the undeformed Earth and the time-variable components are the quantities ΔI_{ij}. If second-order terms are neglected and the relations 5 and 6 are introduced, the linearized equations of motion can be rewritten as

$$\begin{aligned} m_1 - \frac{1}{\sigma_E}\dot{m}_2 &= \psi_1, \\ m_2 + \frac{1}{\sigma_E}\dot{m}_1 &= \psi_2, \\ m_3 &= \psi_3 + const. \end{aligned} \tag{7}$$

with $\sigma_E = \left(\frac{C-A}{A}\right)\Omega$ being the *Euler frequency*, which would be the frequency of the resonance of a rigid Earth, corresponding to a circular motion of the rotation axis with a period of approximately 305 days. If the Earth is considered to be an elastic or at least deformable body the *Euler frequency* is replaced by the complex *Chandler frequency*

$\sigma_C = \frac{2\pi}{T_C}\left(1+\frac{i}{2Q}\right)$, with $T_C \sim 433$ days denoting the observed period of the *Chandler wobble* and $Q = 30\ldots200$ being a dimensionless dissipation factor (see next section for a brief discussion of the *Chandler wobble*). The ψ_i are called *excitation functions*. They contain the changes of the tensor of inertia and the relative angular momenta:

$$\begin{aligned}
\psi_1 &= \frac{1}{\Omega^2(C-A)}\left[\Omega^2\Delta I_{13}+\Omega\Delta\dot{I}_{23}+\Omega h_1+\dot{h}_2-L_2\right]\\
\psi_2 &= \frac{1}{\Omega^2(C-A)}\left[\Omega^2\Delta I_{23}-\Omega\Delta\dot{I}_{13}+\Omega h_2-\dot{h}_1+L_1\right]\\
\psi_3 &= \frac{-1}{\Omega^2\mathbf{C}}\left[\Omega^2\Delta I_{33}+\Omega h_3-\Omega\int_0^t L_3\,\mathrm{d}t\right]
\end{aligned} \tag{8}$$

The first two equations of 7 for m_1 and m_2 express *polar motion*, while the third equation for m_3 describes the magnitude of the rotation speed and hence LOD variations. Provided that all the time-variable parts of the *excitation functions* (changes of the inertia tensor ΔI_{ij}, relative angular momenta h_i, and external torques L_i) are introduced as known quantities from models or observations, the *Euler–Liouville equations* can be solved for ω. Earth rotation variations can be thus calculated or predicted, respectively, from observed or modeled changes in the components of the system Earth, such as atmosphere or oceans. In general, relative motions (relative angular momenta) contribute more to LOD variations while the major effect on *polar motion* comes from alterations of the inertia tensor. The *excitation functions* are in practice mostly replaced by so-called *angular momentum functions*, introduced by Barnes et al. (1983). Elaborate explications of the basic Earth rotation formalism can be found in the fundamental books mentioned in the introduction.

Definition and observation of Earth orientation parameters

The elements which are used to perform the transformation between a space-fixed and an Earth-fixed reference system are commonly referred to as EOP. The realizations of such reference systems by assigning coordinates to selected celestial objects or terrestrial sites, respectively, are called reference frames. The definition and maintenance of the reference frames is one major task of the *International Earth Rotation and Reference Systems Service* (IERS). Conventional reference frames are the space-fixed *International Celestial Reference Frame* (ICRF) and the Earth-fixed *International Terrestrial Reference Frame* (ITRF). The current version of the ICRF is its second realization, the ICRF2 (Fey et al., 2009). The ITRF2008 has been adopted as the official terrestrial reference frame, following ITRF2005 (Altamimi et al., 2007). The definition of the EOP actually depends on the kind of the applied transformation method. The IERS recommends the transformation according to the IAU (*International Astronomical Union*) 2000 Resolutions (McCarthy and Petit, 2004). The transformation procedure is defined from the *Geocentric Celestial Reference System* (GCRS) to the *International Terrestrial Reference System* (ITRS) or vice versa. The ICRS is not a geocentric system – its origin is located in the barycenter of the solar system. To transfer from a barycentric system to a geocentric system, effects, which are not directly related to Earth rotation, like aberration and parallaxes have to be taken into account. The orientation of the ICRS however corresponds to the orientation of the GCRS. The transition from the GCRS to the ITRS is described as a sequence of time-dependent rotation matrices:

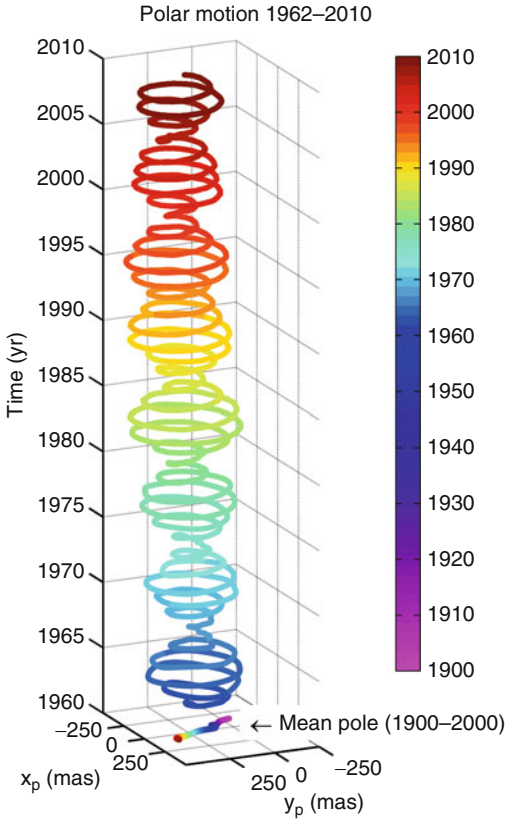
$$[\mathbf{ITRS}] = \boldsymbol{W}(t)\cdot\boldsymbol{R}(t)\cdot\boldsymbol{Q}(t)\cdot[\mathbf{GCRS}] \tag{9}$$

$\mathbf{W}(t)$ (with "W" for wobble) designates the *polar motion* matrix. It contains the coordinates x_p and y_p of the reference pole CIP (*Celestial Intermediate Pole*) in the Earth-fixed system and the angle s', which provides the position of the TIO (*Terrestrial Intermediate Origin*) on the equator of the CIP. The TIO and the *Celestial Intermediate Origin* (CIO) realize an instantaneous prime meridian in the respective system. These terms are part of the transformation concept using the *Non-Rotating Origin* (NRO), which replaced the older transformation scheme with ecliptic (plane of the Earth orbit) and equator (McCarthy and Petit, 2004). The rotation between TIO and CIO is performed with $\boldsymbol{R}(t)$ using a quantity named *Earth Rotation Angle* (*ERA*). The *ERA* is directly related to UT1, which is relevant for Earth rotation research. *Precession* and *nutation* are represented by the matrix $\boldsymbol{Q}(t)$. It comprises rotations around the angles X and Y, the coordinates of the CIP in the celestial system and around the angle s, which locates the CIO on the equator of the CIP. In case of using the transformation according to IAU 2000 resolutions, the five quantities $\{x_p, y_p, dUT1, X, Y\}$ therefore represent the EOP. If the older transformation concept based on ecliptic and equator is applied, X and Y are replaced by $\Delta\varepsilon$ and $\Delta\psi$, *nutation* in obliquity and longitude. The before cited *Celestial Intermediate Pole* is the reference pole measurements of space geodetic techniques are related to. The CIP thus defines the observed axis. This is a pure convention, which is realized by an accordingly adapted *precession-nutation* theory. The direction towards the CIP does not correspond to any physically defined axis, like the rotation axis, the figure axis, or the angular momentum axis; nevertheless, it is possible to mathematically connect it to each of those axes. Casually, it is often said that Earth rotation observations represent the motion of the Earth rotation axis. Regarding the measurements of space geodetic techniques, this is actually not entirely correct, since they are not sensitive to the instantaneous Earth rotation vector but to the complete rotation matrix only. The CIP defines an intermediate pole, separating the motion of the pole of the ITRS in the GCRS into a celestial part and a terrestrial part. The celestial part

(*precession* and *nutation*, $\{X, Y\}$) comprises all motions with periods greater than 2 days, as seen from space. This is equivalent to frequencies between −0.5 and +0.5 cycles per sidereal day (cpsd). With the minus sign labeling retrograde motions (opposite to the sense of Earth rotation) and the plus sign labeling prograde motions (in the sense of Earth rotation), all motions outside of the retrograde diurnal band in the Earth-fixed system, i.e., frequencies below −1.5 and above −0.5 cpsd, are allocated to the terrestrial part (*polar motion*, $\{x_p, y_p\}$).

The celestial motions, *precession* and *nutation,* are long-term and periodic changes of the direction of the Earth rotation axis or actually of the CIP axis, with respect to a space-fixed reference system, which is realized, for example, by the positions of extragalactic radio sources observed by VLBI. Due to *precession*, the Earth axis moves with respect to the space-fixed system on a cone with an aperture of 23.5°, which corresponds to the angle between the Earth equator plane and the ecliptic. The revolution period amounts to approximately 25,800 years. This time interval is also called Platonic year. The tidal forces of moon and sun are responsible for this steady motion. Since the Earth is not a sphere, but can be characterized as an oblate spheroid (qv *Geodesy, Figure of the Earth*) and its rotation axis is inclined by 23.5° with respect to the ecliptic normal, the gravitational torques force the equatorial plane into the ecliptic. Because of its rotation, the Earth acts like a gyroscope and swerves by moving on the above-mentioned cone, whose surface can be calculated very precisely.

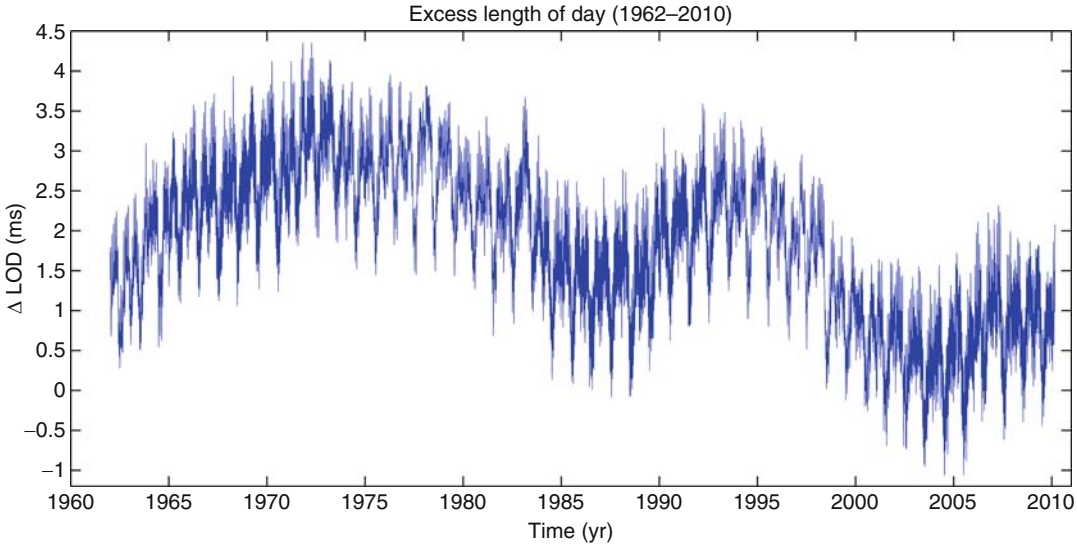
The smaller periodic change in the direction of the Earth rotation axis that is superimposed to *precession* is called *nutation*. This comprises motions of the Earth rotation axis with respect to the space-fixed system with periods from a few days to 18.6 years, caused by gravitational influences of sun, moon, and the planets of our solar system. *Precession* and *nutation* can be modeled and predicted precisely using time-dependent harmonic series expansions. The arguments of the individual harmonic terms are thereby calculated from combinations of five fundamental astronomical arguments. The currently most precise *precession-nutation* model adopted by IAU resolutions 2000 and 2006 (IAU Resolutions, 2000; IAU Resolutions, 2006) is IAU 2006/2000A. In this model effects of planets, ocean tides, mantle anelasticity, and electromagnetic coupling mechanisms between core and mantle as well as between inner and outer core are considered (Mathews et al., 2002). As for the planets, their direct effect on Earth as well as the indirect effect of the sun acting on a planet, which also causes subtle *nutation* terms, are taken into account. Remaining parts of the axis motion with respect to the space-fixed system that are not covered by the *precession–nutation* model can be measured by means of VLBI and are provided by the IERS as so-called *Celestial Pole Offsets*. These residuals originate from still deficiently modeled and unpredictable effects, like the *Free Core Nutation* (FCN). The FCN is a proper mode of the Earth, caused by a misalignment of the rotation axes



Earth Rotation, Figure 1 *Polar motion* from the combined EOP series C04 05 of the IERS (1962–2010, daily resolution) and position of the mean pole from 1900 to 2010.

of mantle and core. Whereas first theoretical estimations indicated a period of about 460 days for this mode, VLBI measurements show that the FCN period is more likely to be around 430 days with highly variable amplitude.

The terrestrial part of the change of the direction of the Earth axis or rather the CIP axis is designated *polar motion* (Figure 1). *Polar motion* has an order of magnitude of several meters and is expressed in a two-dimensional coordinate system by the *pole coordinates* x_p and y_p. According to the definition of the IERS and its precedent organizations, the *x*-axis is oriented in the direction of the Greenwich Meridian and the *y*-axis is oriented positively towards 90° west longitude. Already in 1765, the Swiss mathematician Leonhard Euler calculated a circular motion of the pole (of an Earth at that time assumed to be rigid) with a period of around 304 days. Today, we know that *polar motion* is mainly composed from an annual variation and the *Chandler oscillation* or *Chandler wobble*, named after Seth Carlo Chandler who first detected a period of approximately 14 months when analyzing *polar motion* observations at the end of the nineteenth century. The *Chandler wobble* is another proper motion of the Earth with strongly varying amplitude. More precisely, this nearly circular motion is a damped oscillation, which would have vanished due to



Earth Rotation, Figure 2 Excess length of day from the combined EOP series C04 05 of the IERS (1962–2010, daily resolution).

friction in the Earth interior after a few decades if there was not a constantly revolving excitation. Although the *Chandler oscillation* is known for more than 100 years, its underlying excitation mechanism is still under investigation. Today, there is broad consensus that the necessary excitation energy emerges from irregular processes in the atmosphere-ocean-system (Gross, 2000). A radical change in the phase of the *Chandler wobble* around 1925 could also not yet fully be explained. Variations of the Earth magnetic field are nowadays quoted as potential causes for the phase variations. The interference of the *Chandler wobble* and the annual motion leads to a beat-like rising and ebbing of the *polar motion* amplitude to a maximum of 9 m with a beat period of about 6.3 years. The *polar motion* spectrum below 1 year is dominated by irregular variations appearing in 3- and 5-months intervals. The most important short-period variations are due to the ocean tides caused by sun and moon with essentially diurnal and semidiurnal periods, albeit the total effect is only about 1/100 of the *Chandler wobble*. Analysis of long-term *polar motion* shows a distinct variation of approximately 11 years. It is usually associated with processes in the Earth interior but could also be connected to a corresponding cycle of the solar activity. In addition there are other, decadal variations with periods around 30 years and between 70 and 80 years. These are assumed to be caused by geodynamic couplings between Earth core and mantle. Secular *polar motion*, i.e., the long-term linear trend in the position of the pole, amounts to about 3.3 mas/year to the direction 76–78° west longitude (e.g., Schuh et al., 2001). This effect is supposed to be predominantly due to melting of the polar ice masses and postglacial rebound.

LOD is used to express the speed of Earth rotation alternatively to the difference *dUT1* between UT1, which is directly connected to the Earth rotation speed, and the regular atomic time UTC. The parameter, which is usually quoted to characterize changes in LOD, is actually *ΔLOD*, the so-called excess LOD which represents the deviation of the effective LOD from the nominal value of 86,400 s (Figure 2). Variations of LOD can be assigned to different period ranges. In the short-period range with periods from around 1 day and half a day, the strongest influence emerges from the ocean tides caused by sun and moon. Hydrodynamic models or altimetry (qv *Geoid Determination, Theory and Principles*) provide variations of ocean heights and velocities of the water masses, from which global changes of angular momentum of the world oceans are calculated. These are subsequently used to estimate the oceanic influence on the Earth rotation parameters. In this way high-frequency ERP variations can be predicted from tide and circulation models and they can already be observed by the high-precision space geodetic techniques mentioned above. In the range of a few weeks to months, periods of ∼14 and ∼28 days, due to solid Earth tides (qv *Earth Tides*), are dominant. In addition, there are other strong variations between 40 and 90 days basically excited by zonal winds. Seasonal variations due to changes in the angular momentum of the atmosphere show semiannual and annual periods. The variation of LOD with a period of approximately 1 year is predominantly due to annually changing wind patterns. A definite amplification of the annual variation every 4–6 years is associated with large-scale climate signals related to the El Niño phenomenon by which an ocean circulation with a very characteristic pattern in the Southern Pacific is connected to variations of the meteorological parameters. Decadal fluctuations of LOD are attributed to angular momentum changes in the fluid core, which are transferred to the mantle via mechanic or

electromagnetic coupling. Because of tidal friction and long-term mass variations, a secular prolongation of the day by about 1.8 ms in 100 years is observed as well (Morrison and Stephenson, 2001). This is, of course, an average long-term trend deduced from eclipse data over the last 2,500 years. Besides the secular trend, there is also evidence of fluctuations of LOD on a timescale of centuries with an amplitude of around 4 ms.

IERS International Earth Rotation and Reference Systems Service

The IERS is an international service for the determination of EOP and their dissemination to interested users. The definition and realization of reference systems for geodesy and astronomy can be regarded as its superior task. According to the terms of reference (Dick and Richter, 2009), the primary objectives of the IERS are to serve the astronomical, geodetic, and geophysical communities by providing the following:

- The International Celestial Reference System (ICRS) and its realization, the International Celestial Reference Frame (ICRF).
- The International Terrestrial Reference System (ITRS) and its realization, the International Terrestrial Reference Frame (ITRF).
- EOP required to study Earth orientation variations and to transform between the ICRF and the ITRF.
- Geophysical data to interpret time/space variations in the ICRF, ITRF, or EOP, and to model such variations.
- Standards, constants, and models (i.e., conventions) encouraging international adherence.

In addition, the IERS collects, archives, and distributes several products such as reference frames, monthly Earth orientation data, daily rapid service estimates of near real-time Earth orientation data and their predictions. The IERS also announces the differences between astronomical and civil time for time distribution by radio stations and leap seconds which – if necessary to keep the differences *dUT1* smaller than 0.9 s – are added at midnight of July 31st or December 31st. Further products are related to global geophysical fluids such as mass and angular momentum distribution, annual reports, and technical notes on conventions and other topics, and long-term Earth orientation information. The IERS began operation on 1st of January 1988 as common institution of the IAU and the *International Union of Geodesy and Geophysics* (IUGG) and replaced thereby the former *International Polar Motion Service* (IPMS) and the Earth rotation section of the *Bureau International de l'Heure* (BIH). The service consists among other parts of a Central Bureau, technique centers of the single-space geodetic techniques and several product centers, for example, for the collection of data about geophysical influences on Earth rotation (*Global Geophysical Fluids Centre*, GGFC). These are atmospheric and oceanic variations, hydrodynamical effects like groundwater variations and processes in the Earth interior which lead to changes in the rotational behavior of the Earth. The Central Bureau is located at the BKG (*Bundesamt für Kartographie und Geodäsie*) in Frankfurt am Main, Germany. Apart from regular publication of EOP, the IERS also issues guidelines (IERS Conventions, McCarthy and Petit (2004)), which contain models and standards recommended for the data processing of space geodetic techniques.

Summary

Earth rotation is conventionally described by the EOP, which represent the link between a space-fixed (celestial) and an Earth-fixed (terrestrial) coordinate system. Those EOP expressing the temporal variations of the orientation of the Earth correspond to *nutation/precession* (changes of the direction of Earth rotation axis with respect to a space-fixed reference frame) and *polar motion* (wobbling of the Earth with respect to its axis). Small variations of the speed of Earth rotation are expressed by UT1 minus the uniform atomic time (UTC) or as variations in the LOD. A precise knowledge of the Earth's attitude is needed for all positioning and navigation tasks on Earth and in space. It also gives fundamental information about the interactions between the various geophysical components of system Earth and allows deriving conclusions about phenomena of celestial mechanics. The EOP are nowadays monitored by space geodetic techniques and assembled and published by the IERS.

Bibliography

Altamimi, Z., Collilieux, X., Legrand, J., Garayt, B., and Boucher, C., 2007. ITRF2005: a new release of the International Terrestrial Reference Frame based on time series of station positions and Earth Orientation Parameters. *Journal of Geophysical Research*, **112**, B09401, doi:10.1029/2007JB004949.

Barnes, R. T. H., Hide, R., White, A. A., and Wilson, C. A., 1983. Atmospheric angular momentum functions, length-of-day changes and polar motion. *Proceedings. Royal Society of London*, **A387**, 31–73.

de Viron, O., Dehant, V., Goosse, H., Crucifix, M., and Participating CMIP Modeling Groups, 2002. Effect of global warming on the length-of-day. *Geophysical Research Letters*, **29**(7), 1146, doi:10.1029/2001GL013672.

Dick, W. R., and Richter, B. (eds.), 2009. *IERS Annual Report 2007. International earth rotation and reference systems service, central bureau*. Frankfurt am Main: Verlag des Bundesamts für Kartographie und Geodäsie.

Fey, A., Gordon, D., and Jacobs, C. S. (eds.), 2009. *The Second Realization of the International Celestial Reference Frame by Very Long Baseline Interferometry, Presented on behalf of the IERS/IVS Working Group. IERS Technical Note 35*. Frankfurt am Main: Verlag des Bundesamts für Kartographie und Geodäsie.

Gross, R. S., 2000. The excitation of the Chandler wobble. *Geophysical Research Letters*, **27**(15), 2329–2332.

Gross, R. S., 2007. Earth rotation variations – long period. In Herring, T. A. (ed.), *Physical Geodesy*. Amsterdam: Elsevier. Treatise on Geophysics, Vol. 3.

Lambeck, K., 1980. *The Earth's Variable Rotation, Geophysical Causes and Consequences*. Cambridge: Cambridge University Press.

Mathews, P. M., Herring, T. A., and Buffett, B. A., 2002. Modeling of nutation and precession: new nutation series for nonrigid Earth, and insights into the Earth's Interior. *Journal of Geophysical Research*, **107**(B4), doi:10.1029/2001JB000390.
McCarthy, D. D., and Petit, G. (eds.), 2004. *IERS Conventions (2003). IERS Technical Note 32*. Frankfurt am Main: Verlag des Bundesamts für Kartographie und Geodäsie.
Moritz, H., and Mueller, I. I., 1987. *Earth Rotation: Theory and Observation*. New York: Ungar.
Morrison, L. V., and Stephenson, F. R., 2001. Historical eclipses and the variability of the Earth's rotation. *Journal of Geodynamics*, **32**(1–2), 247–265.
Munk, W. H., and MacDonald, G. J. F., 1960. *The Rotation of the Earth: A Geophysical Discussion*. Cambridge: Cambridge University Press.
IAU Resolutions, 2000. http://www.iau.org/static/resolutions/IAU2000_French.pdf
IAU Resolutions, 2006. http://www.iau.org/static/resolutions/IAU2006_French.pdf
Schreiber, U., Velikoseltsev, A., Rothacher, M., Klügel T., Stedman, G., and Wiltshire, D., 2004. Direct measurement of diurnal polar motion by ring laser gyroscopes. *Journal of Geophysical Research*, **109**(B6), doi:10.1029/2003JB002803.
Schuh, H., Nagel, S., and Seitz, T., 2001. Linear drift and periodic variations observed in long time series of polar motion. *Journal of Geodesy*, **74**, 701–710.
Seitz, F., and Schuh, H., 2010. Earth Rotation. In Xu, G. (ed.), *Sciences of Geodesy*. Berlin: Springer.

Cross-references

Earth Tides
Geodesy, Figure of the Earth
Geodesy, Networks and Reference Systems
Geoid Determination, Theory and Principles
GPS, Data Acquisition and Analysis
Satellite Laser Ranging
Very Long Baseline Interferometry

EARTH TIDES

John M. Wahr
Department of Physics and Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA

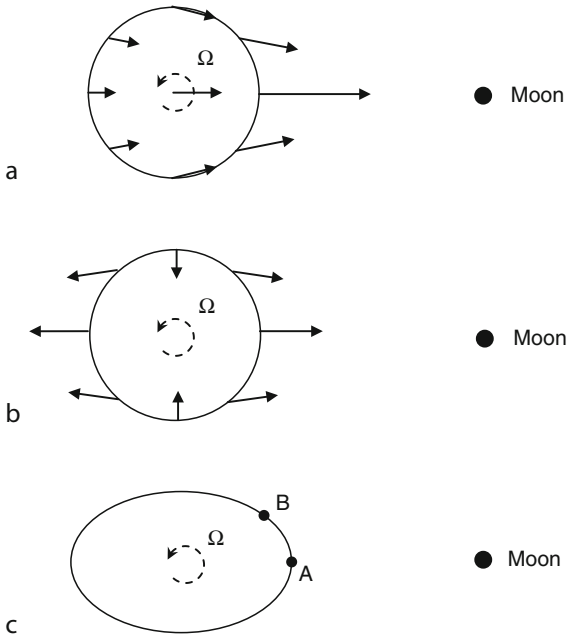
Definition

The earth tide is the deformation of the solid earth caused by the gravitational attraction of the Sun and moon. The most striking gravitational effects of the Sun and moon are the orbital motion of the earth's center of mass. The moon causes the earth to orbit the earth–moon center of mass, and the Sun causes that center of mass to orbit the Sun.

The Tidal Force

But the Sun and moon also cause points within the earth to be displaced relative to each other: they cause the earth to deform. For example, the side of the earth nearest the moon is attracted toward the moon more than is the center of the earth. And the side farthest from the moon is attracted less than the center of the earth. Consequently, both the far and near sides of the earth are pulled radially outward away from the center. The regions of the earth that are at right angles to the earth–moon vector are pulled radially inward.

This deformation pattern is illustrated in Figure 1. The total gravitational acceleration vectors due to the moon are shown in Figure 1a. The vectors all point toward the moon, but are of unequal length and direction, although the differences are greatly exaggerated in the figure. The orbital acceleration is, to a high degree of accuracy, equal to the acceleration vector at the earth's center of mass. Subtracting that vector from the other vectors in Figure 1a results in the pattern of vectors shown in Figure 1b. These residual vectors represent that portion of the lunar gravitational force tending to deform the earth. The residual acceleration field multiplied by the local material density is defined as the lunar tidal force, and the deformation it induces is called the lunar tide.



Earth Tides, Figure 1 All three panels show the earth as seen from above the north pole. The earth's diurnal rotation is represented by W. The *arrows* in (**a**) illustrate the magnitude and direction of the gravitational acceleration toward the moon. The differences in the lengths and directions of the *arrows* are greatly exaggerated. The lunar tidal acceleration is defined by subtracting the acceleration vector at the earth's center of mass from the other acceleration vectors. The resulting vectors are shown in (**b**). The tidal force (the local density times the acceleration) deforms the earth into the elliptical shape shown greatly exaggerated in (**c**). Points A and B are used in the text to illustrate tidal strain and tilt.

Note from Figure 1b that because of the earth's diurnal rotation the tidal force at a fixed point in the earth varies through two complete cycles in 1 day; a fixed point rotates through one outward bulge, then through an inward bulge, and then into the other outward bulge, all within a half a rotation period (i.e., within 12 h). This semidiurnal time dependence is split into many periodic terms with frequencies closely spaced about 2 cycles/day, due to the time variability of the orbital motion of the moon. Furthermore, because the moon is not always in the plane of the earth's equator, there is also significant variability at frequencies closely spaced about 1 cycle/day (imagine, e.g., that the earth–moon vector is inclined at 45° to the earth's rotation axis) and about 0 cycles/day (imagine that the moon is directly above the north pole). The solar tidal force and the solar tide are defined in a similar manner, and can also be decomposed into semidiurnal, diurnal, and long period terms.

For a quantitative description, it is useful to work with the tidal potential, denoted here as V_T and defined so that its gradient is the tidal force per unit mass. Define a nonrotating coordinate system with origin at the center of the earth. The total lunar gravitational potential at the point **r** inside the earth is (assuming the moon to be a point mass)

$$V = \frac{GM}{|\mathbf{r} - \mathbf{R}|} \tag{1}$$

where M is the mass of the moon, G is Newton's gravitational constant, and **R** is the position vector of the moon. Let r, θ, ϕ be the radial distance, colatitude, and eastward longitude of the earth-fixed point, **r**, so that $\phi + \Omega t$ (where $\Omega = 1$ cycle/day is the earth's angular velocity of rotation) is the azimuthal angle of **r** in nonrotating, inertial space. The factor $1/|\mathbf{r} - \mathbf{R}|$ can be expanded as a sum of complex spherical harmonics, $Y_n^m(\theta, \phi)$, so that

$$V = \mathrm{Re}\left[\frac{GM}{R}\sum_{n=0}^{\infty}\sum_{m=0}^{n}\left[\frac{r}{R}\right]^n d_n^m Y_n^m(\theta,\phi)e^{im\Omega t}\right] \tag{2}$$

where *Re* denotes the real part (included in Equation 2 so that the sum over m includes only $m \geq 0$), R is the radial coordinate of the moon, and d_n^m are coefficients that depend on the angular position of the moon relative to nonrotating inertial space. The coordinate R and the moon's angular position depend on time, due to the moon's orbital motion.

The $n = 0$ term in Equation 2 is a spatial constant and so is physically meaningless. It can be ignored.

The $n = 1$ term represents a spatially constant force, equal to the lunar force at the earth's center of mass. Consequently, the n=1 term is removed from Equation 2 when defining the tidal potential V_T.

Since $r \leq a$ where a is the earth's radius, and since $a/R \approx 1/60$ for the moon (and is much smaller for the Sun), the factor $(r/R)^n$ in Equation 2 causes the contributions to Equation 2 to fall off rapidly with increasing n. For most purposes it is sufficient to keep only the $n = 2$ terms in V_T, so that V_T has the approximate form

$$V_T = \mathrm{Re}\left[\left[\frac{r}{a}\right]\sum_{m=0}^{2} c_2^m Y_2^m(\theta,\phi)e^{im\Omega t}\right] \tag{3}$$

where $c_2^m (= GM/R(a/r)^2 d_2^m)$ are complex, time-varying coefficients that depend on the orbital coordinates of the moon.

The dominant time dependence in Equation 3 comes from the $e^{im\Omega t}$ term. This term results in semidiurnal, diurnal, or long period tides, depending on whether m is 2, 1, or 0, respectively. The time-dependent c_2^m in Equation 3 split the $e^{im\Omega i}$ time dependence into terms with frequencies closely spaced about mΩ.

The Earth's Response

The tidal force causes ocean tides as well as earth tides. The observed ocean tide is actually the difference between the response of the ocean and the response of the solid earth. Ocean tides are complicated and difficult to model, for a number of reasons that need not concern us here. Open ocean tides are typically less than 1 m in height, although ocean tides at some coastal locations can be as large as several meters peak-to-peak.

Earth tides are much easier to understand and model. The tidal force shown in Figure 1b tends to deform the solid earth into the elliptical shape shown greatly exaggerated in Figure 1c. Tidal displacements of the solid earth are typically several tens of centimeters. But, unlike ocean tides, earth tides cannot be observed without sensitive instruments. The reason is that earth tides cause both the ground and the observer to be displaced by the same amount.

One way to detect earth tides is to use a gravimeter. There are three contributions to the observed tidal variation in the gravitational acceleration: (1) the direct attraction of the Sun and moon, (2) the change in the earth's gravity field due to tidal deformation within the earth, and (3) the change in the gravitational acceleration at the gravimeter caused by the radial tidal displacement of the earth's surface beneath the gravimeter (if the surface moves outward, gravity decreases, and vice versa). All three contributions are roughly of the same order.

The total tidal variation of the observed gravitational acceleration is on the order of 100 μgals or more (1 gal = 1 cm/s^2). The unperturbed gravitational acceleration at the earth's surface is about 10^3 gals. So, tides affect gravity at about the 10^{-7} level. This is roughly the relative importance of tides on any physical observable, and can be traced to the fact that tidal displacements (a few tens of centimeters) are about 10^{-7} of the earth's radius (about 6×10^8 cm).

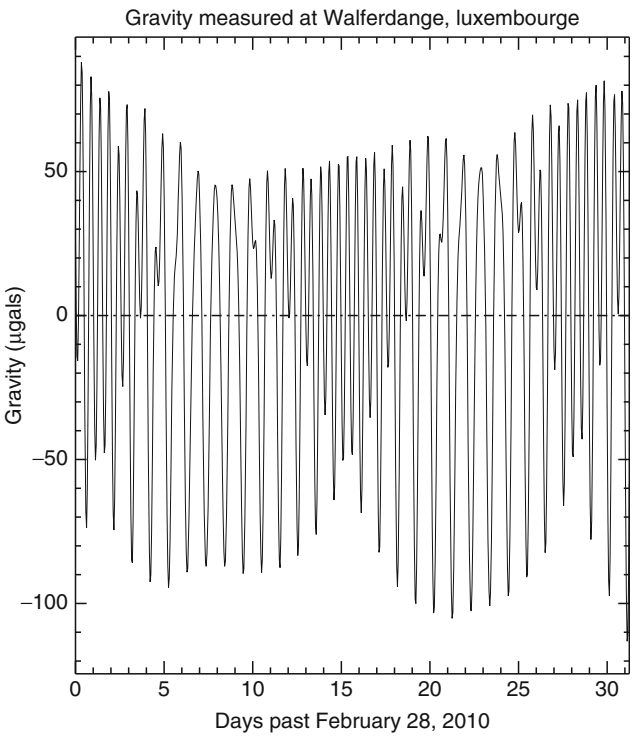
A simple, idealized gravimeter is a mass hanging from a spring. The amount the spring is stretched at equilibrium is proportional to the local gravitational acceleration. So

by continually monitoring the length of the spring, tidal variations in gravity can be determined. Although real gravimeters are more complicated than this, the basic idea is the same. Most instruments, in fact, do use springs.

To illustrate the effects of tides on gravity, Figure 2 shows 1 month of hourly gravitational acceleration data as observed with a spring-type instrument at Walferdange, Luxembourg (data courtesy of Olivier Francis). The tides are the dominant signal in the data. The semidiurnal and diurnal tides are especially evident. The longer period (2 weeks and 1 month) amplitude modulations evident in the figure are caused by beating between frequencies separated by 1 cycle/13.7 days and between frequencies separated by 1 cycle/27.6 days. Long period gravity tides have small amplitudes at the latitude of Walferdange, and are hard to distinguish in Figure 2.

One problem common to all relative gravimeters is calibration. For spring-type instruments, the proportionality constant between the length of the spring and gravity depends on the spring constant, which can never be precisely known. This, for example, could cause errors in the vertical scale in Figure 2.

Another method of detecting earth tides is to use a strainmeter. This is an instrument that continually monitors the distance between two points on the earth's surface, separated by anywhere from a few centimeters to a kilometer. The fact that earth tides perturb the distance between points can be seen by considering points A and B in Figure 1c. The distance between those points has been changed by the tidal deformation, because A has moved radially outward more than B. Also, not evident in the figure, A and B have been displaced horizontally by different amounts. Typical tidal variations in the distance between two points are on the order of 10^{-7} times the unperturbed distance.



Earth Tides, Figure 2 One month of hourly gravity data, taken with a spring-type gravimeter at Walferdange, Luxembourg (data courtesy of Olivier Francis). The tides are the dominant signal and are described further in the text.

A third way to detect tidal deformation is with tiltmeters. These instruments measure the tilt of the earth's surface with respect to the direction of gravity. Tidal tilt of the surface can be understood from Figure 1c by noting that the tidal bulge has affected the slope of the line between points A and B. The direction of gravity also tips during the deformation, but by a different amount. It is this difference that is observed. Typically, observed tidal tilts are a few tenths of a microradian.

Tidal deformation can also be observed using space geodetic techniques, including satellite laser ranging (SLR), very-long-baseline-interferometry (VLBI), and GPS positioning. These techniques are sensitive to displacements of the ground beneath the observing stations. In fact, one of the primary geophysical objectives of these techniques is to detect tectonic motion of the stations. The effects of earth tides are evident in the data, but they cannot generally be determined with as much relative accuracy as can the tidal effects on surface gravity, strain, or tilt.

Satellite orbit tracking can also be used to detect tidal variations in the earth's gravity field. Those variations affect the orbit of a satellite, and so can show up in the satellite tracking data. In fact, for lower-flying satellites the ranging data are usually more sensitive to these tidal gravity variations than they are to tidal displacements of the ground stations.

These various observable tidal effects can be conveniently described, mathematically, by using dimensionless parameters known as Love numbers, defined as follows. The deformation of a spherically symmetric, nonrotating earth in response to a $Y_n^m(\theta, \lambda)$ applied external potential can be described with that same $Y_n^m(\theta, \lambda)$ angular dependence. Mathematically, this is because the Y_n^m separate spherically symmetric differential equations. The change in the earth's gravitational potential at $r = a$ (the unperturbed surface of the earth) and the radial displacement at $r = a$ are both proportional to Y_n^m, and the horizontal displacement vector at $r = a$ is proportional to ∇Y_n^m. The proportionality constants depend on n but are independent of m. Thus, the same constants are pertinent for all three Y_2^m terms in Equation 3.

Let $\Phi(\theta, \lambda)$, $U_r(\theta, \lambda)$, $U_\theta(\theta, \lambda)$, and $U_\lambda(\theta, \lambda)$ denote tidal effects at the earth's surface ($r = a$) on, respectively, the earth's gravitational potential, and the radial, southward, and eastward displacements of the point (θ, λ). Then, Φ, U_r, U_θ, and U_λ for our assumed spherical earth have the form

$$\begin{aligned}\Phi(\theta,\lambda) &= kV_T^a(\theta,\lambda)\\ U_r(\theta,\lambda) &= \frac{h}{g}V_T^a(\theta,\lambda)\\ U_\theta(\theta,\lambda) &= \frac{l}{g}\partial_\theta V_T^a(\theta,\lambda)\\ U_\lambda(\theta,\lambda) &= \frac{l}{g\sin\theta}\partial_\lambda V_T^a(\theta,\lambda)\end{aligned} \tag{4}$$

where k, h, and l are the dimensionless Love numbers, g is the unperturbed gravitational acceleration at $r = a$, and

$$V_T^a(\theta,\lambda) = \mathrm{Re}\left[\sum_{m=0}^{\infty} c_2^m Y_2^m(\theta,\lambda)e^{im\Omega t}\right] \tag{5}$$

is the tidal potential (Equation 3) evaluated at $r = a$.

Tidal variations in gravity, strain, tilt, surface displacements, and satellite orbital motion can all be parameterized in terms of k, h, and l. As one example, consider the gravitational acceleration at a fixed point on the earth's surface. There are three contributions to this acceleration, as discussed above. First, the direct gravitational acceleration from the moon (or Sun) in the inward radial direction is $-\partial_r V_T(r=a) = -\frac{2}{a}V_T^a$ (using the r^2 radial dependence of V_t shown in Equation 3). Second, Φ (the tidal change in the earth's gravitational potential at $r = a$) has an angular dependence Y_2^m and so, outside of the earth, has radial dependence r^{-3}. Thus, the effect of Φ on the radially inward gravitational acceleration is $-\partial_r\Phi(r=a) = \frac{3}{a}\Phi = \frac{3}{a}kV_T^a$ Third, in the absence of tides, the surface gravitational acceleration, g, varies with radius as r^{-2}. So, if the surface point is displaced radially by U_r, the resulting perturbation in the gravitational acceleration at the displaced surface point is $U_r\partial_r g = -\frac{2}{a}gU_r = -\frac{2}{a}hV_T^a$.

Adding these three contributions together gives the total tidal effect on the observed acceleration as

$$\Delta g(\theta,\lambda) = -\delta\frac{2}{a}V_T^a(\theta,\lambda) \tag{6}$$

where $\delta = 1 - \frac{3}{2}k + h$ is called the gravimetric factor. Similar exercises show that tidal tilt is described by the diminishing factor $\gamma = 1 + k - h$, and that tidal strain, although more complicated than either tilt or gravity, depends on the Love numbers h and l.

The numerical values of the Love numbers depend on the earth's internal properties. It is possible to learn about some of those properties by comparing tidal observations with predictions based on theoretical results for the Love numbers. For example, tidal observations have been used to place constraints on the earth's anelastic properties. The effects of anelasticity are frequency- and stress-dependent, but are not well understood. Because tides sample the earth in unique frequency and stress regimes (at lower frequencies and larger deep-earth stresses than seismology, e.g.), tidal studies have proven useful as complements to other types of anelastic observations.

The use of tides to study the earth's internal structure can be difficult, for several reasons. One problem is instrument calibration, as described above for gravimeters. Any error in the calibration maps directly into a frequency independent, multiplicative error in the estimated Love numbers.

But there are also a number of other complicating effects, most related to the fact that the earth is not really spherically symmetric and nonrotating. Some of those effects are potentially useful. For example, the earth does rotate, and because of that rotation the earth's internal properties are closer to being elliptically, rather than spherically, symmetric. In this case, the results of Equation 4 are still approximately valid, but the Love numbers k, h, and l are notably dependent on frequency near 1 cycle/day. This diurnal frequency dependence is particularly sensitive to the shape of the core-mantle boundary, and so diurnal tidal results can be used to help constrain that shape.

Another omission in the theory described above is the ocean. Tides in the ocean cause time-varying pressure loads on the surface of the solid earth with the same frequencies as the earth tides. These loads cause the earth to deform, and this "load tide" (the tidal deformation that would occur in the absence of oceans is called the "body tide") affects all earth tide observations to some extent. Tidal gravity observations, for example, can be perturbed by up to 10% near coasts, and typically a few percent in the interior of continents. Tidal tilt and strain can be perturbed by several hundred percent near coasts.

It is often difficult to model and remove the load tide well enough to use the remaining body tide to learn about the earth's deep interior. There are uncertainties in ocean tide models. Furthermore, the earth's response to the ocean tides, particularly the contributions to tilt and strain near coasts, can be sensitive to the local material properties that may or may not be adequately known. On the other hand, because of these possible uncertainties, people have sometimes been able to use tidal observations to help constrain the nearby ocean tide or the underlying material properties.

In fact, for tilt and strain local effects can be important on both the load tide and the body tide. By comparing observed tidal amplitudes from instruments in an array of tiltmeters or strainmeters, geophysicists can learn about the local geology and underlying structure. There have even been attempts to look for time-dependent variations in tidal amplitudes near active earthquake faults that might be caused by sudden changes in local material properties preceding an earthquake.

A related problem is that tilt and strain amplitudes are also affected by the local topography and by the shape and size of any cavity the instrument is placed in. (Tiltmeters and strainmeters are often placed in boreholes or tunnels to minimize the effects of the surface environment.) These effects are rarely interesting, and they cannot be observationally separated from the effects of local geology. Instead, they must be modeled separately and removed from the data.

Summary

The effects of earth tides can be detected in several types of geophysical measurements. These tidal observations can be inverted to learn about large-scale material properties of the earth's deep interior and about local geological structure. In practice, a useful inversion is often difficult due to the large number of possible geophysical interpretations, and due to uncertainties associated with instrument calibration, topographic and cavity effects, and ocean tidal loading of the solid earth.

Bibliography

Agnew, D. C., 2007. Earth tides. In Schubert, G., and Herring, T. (eds.), *Treatise on Geophysics*. Oxford: Elsevier, Vol. 3, pp. 163–195.

Baker, T. F., 1984. Tidal deformations of the earth. *Science Progress (Oxford)*, **69**, 197–233.

Harrison, J. C., 1984. *Earth Tides (Benchmark Papers in Geology)*. Pennsylvania: Hutchinson Ross.

Melchior, P., 1983. *The Tides of the Planet Earth*. Oxford: Pergamon.

Wilhelm, H., Zurn, W., and Wenzel, H. G., 1997. *Tidal Phenomena*. Berlin: Springer.

Cross-references

Earth Rotation
Earth, Density Distribution
Geodesy, Figure of the Earth
Geodesy, Networks and Reference Systems
Geodesy, Physical
Geodynamics
GPS, Tectonic Geodesy
Gravity Method, Satellite

EARTH, DENSITY DISTRIBUTION

Frank D. Stacey[1], Paul M. Davis[2]
[1]Division of Exploration and Mining, CSIRO, Kenmore, Australia
[2]Earth and Space Sciences, UCLA, Los Angeles, CA, USA

Definition

Except for iron meteorites, the Earth is the densest body in the solar system. As has been recognized for more than two centuries, it is, on average, about twice as dense as common surface rocks. But many of the early speculations on its internal structure were bizarre, ignoring the most basic physics and the simplest density information, and prompting a remark in one of the early papers on the subject with a sound scientific base (Oldham, 1906): "Many theories of the earth have been propounded . . . till geologists have turned in despair from the subject, and become inclined to confine their attention to the outermost crust of the earth, leaving the centre as a playground for mathematicians." Seismological studies, for which Oldham was one of the pioneers, have changed that by giving us a clear picture of the interior. But debates on many details and improvements in the methods of investigating them continue.

The mean density

The first well-informed estimate of the Earth's mean density was by Isaac Newton, who appealed to four pieces of evidence: the Earth's surface gravity, the period of the Moon's orbit, the size of Jupiter, and the orbital periods of its satellites. It was evident to him that the density of Jupiter is a quarter of the density of the Earth. Arguing that the density of Jupiter would not be lower than that of water (it is now known to be 1,327 kg m^{-3}), and recognizing that the densest Earth materials would have sunk to the deep interior, he inferred that the mean Earth density is between 5,000 and 6,000 kg m^{-3}, a remarkable coincidence with the now established value, $\bar{\rho}$ = 5513.4(6) kg m^{-3}. More precise estimates awaited the determination of the gravitational constant, G. Any astronomical or geophysical measurement of gravity, g, or, equivalently the acceleration of a body in a gravitational field, depends on the product of G and a mass. In the approximation of a spherical, nonrotating Earth, mass M, and radius R, as Newton showed, the surface gravity is $g = GM/R^2$ (the departure from this simple equation caused by rotation and ellipticity is about 0.5%). Since g and R have been reasonably well known since Newton's time, a determination of G is equivalent to a determination of M, or $\bar{\rho}$, and was referred to as such by early investigators. Satellite measurements have given the most precise value of the product $GM = 3.986004415(8) \times 10^{14}$ m^3 s^{-2}, where M includes the mass of the atmosphere, 8.84×10^{-7} of the total.

The first attempts to measure G were by M. Bouguer, who, in the 1730s, was conducting an astrogeodetic survey in the extreme topography of the Andes and noted a plumb line deflection by Mount Chimborazo, in Ecuador. He also compared the gravity on a plateau, at Quito, with observations on the nearby flood plain of the Esmeralda River, the first use of what we now know as the Bouguer method of gravity interpretation. In both cases Bouguer recognized that he was thwarted by heterogeneity of the geological structures and that his estimates of $\bar{\rho}$ were unsatisfactory, but in 1774 his plumb line deflection method was repeated more successfully by N. Maskelyne on the slopes of Mt. Schiehallion in Scotland. Maskelyne's estimate was revised several times by others, using different assessments of the density of the mountain. This introduced a measure of uncertainty, so that geophysical measurements were generally disregarded for about 50 years after 1798, when H. Cavendish made a series of measurements of the gravitational force between laboratory masses, using a torsion balance developed for the purpose by J. Michell. However, Cavendish's result was also a subject of doubt. Although the measurements were excellent, the final result was clouded by an arithmetic error, recognized in 1842 by F. Baily, leading to a revival of

interest in geophysical methods. The most significant of these was by G. B. Airy, who, in 1856, measured the variation of gravity with depth in a coal mine near Newcastle, in northeast England. Although limited by inadequate measurements of rock density, Airy's experiment was well conceived. It did not require an absolute measurement of gravity but only the differences in its value at different levels in the mine, as indicated by the fractional differences between the periods of pendulums operated by observers who exchanged electrical signals to indicate their timing. The linked variations in gravity and density, recognized by Airy, are central to the development of earth models. The Airy method was used by others for the following 50 years, but with mixed results, and by 1900 the superiority of torsion balance measurements was clear. Now, it appears that conventional balance experiments, pioneered in the late 1800s by P. von Jolly and J. H. Poynting, are better still and the current best estimate of G derives primarily from them: $G = 6.6743(7) \times 10^{-11}$ m^3 kg^{-1}s^{-2}. The uncertainty (1 standard deviation), 1 part in 10^4, is far larger than the uncertainties in the other fundamental physical constants and translates to a similar uncertainty in M, and therefore $\bar{\rho}$, through the very precise value of (GM).

The moment of inertia constraint

The understanding of the Earth's density structure made little progress, beyond Newton's inference that denser materials would have sunk to the centre, until the late 1800s. By then astronomical measurements of precession, combined with astrogeodetic measurements of ellipticity, had given an estimate of the moment of inertia of the Earth, C. This is expressed as a coefficient, f, relating C to the mass, M, and equatorial radius, a

$$C = fMa^2 \tag{1}$$

For a uniform sphere $f = 0.4$ and a lower value is a measure of the concentration of mass toward the centre. Using recent data the value for the Earth is $f = 0.330968(2)$, with the uncertainty in parenthesis. This number represents a quite strong central concentration of mass and prompted speculation on the Earth's deep composition. The cosmic abundance of iron, apparent from meteorites and the solar spectrum, invited the supposition that the Earth has an iron core. This idea was pursued by E. Wiechert who, in 1897, showed that both M and f could be matched by a model with an iron core and rocky mantle, with both materials approximating their familiar low pressure densities, if the core radius was about $0.78a$. But in 1906 R. D. Oldham reported seismological evidence for a smaller core, with a radius that he estimated to be about $0.4a$ (it is now known to be $0.546a$), requiring higher densities for both the core and the mantle. Two developments were needed: recognition of the increase in density by the compression of deep burial and the introduction of a transition zone of increasing intrinsic density within the mantle. Both required more detailed seismological data than were available to Wiechert and Oldham.

Self-compression

Williamson and Adams (1923) recognized that information about self-compression in a homogeneous layer was available from seismic wave velocities. Using the velocities of both compressional and shear waves

$$V_\mathrm{P} = \{[K_S + (4/3)\mu]/\rho\}^{1/2}; V_\mathrm{S} = \{\mu/\rho\}^{1/2} \tag{2}$$

they separated the adiabatic bulk modulus, K_S, from the rigidity modulus, μ, by writing

$$\Phi = V_\mathrm{P}{}^2 - (4/3)V_\mathrm{S}{}^2 = K_S/\rho = (\partial P/\partial\rho)_S \tag{3}$$

That is, the seismological parameter Φ gives directly the variation of pressure, P, with density, ρ, in a homogeneous region with an adiabatic temperature gradient. Since $\mathrm{d}P/\mathrm{d}z = \rho g$, g being gravity at depth z,

$$\mathrm{d}\rho/\mathrm{d}z = \rho g/\Phi \tag{4}$$

This is the Williamson–Adams equation for the variation of density with depth in a homogeneous region. An additional term, $-\alpha\rho\tau$, was added by F. Birch to account for a temperature gradient differing from the adiabat

$$\tau = \mathrm{d}T/\mathrm{d}z - (\mathrm{d}T/\mathrm{d}z)_\mathrm{Adiabatic} \tag{5}$$

where α is the volume expansion coefficient. However, this term is very small except at the limited depths of steep temperature gradients, the lithosphere, and the base of the mantle (D″). In the outer core it is certainly negligible.

Another reason for imperfection of the Williamson–Adams equation is that, for composite materials, such as mantle rocks that are comprised of several minerals with different properties, there is a frequency dependence of elasticity. The unrelaxed elastic moduli that apply to rapid cycles of low stress, as in seismic waves, are slightly higher than the relaxed moduli applying to strong static compression. Averaged over the lower mantle the difference is about 0.5% (Stacey and Davis, 2008, Table 18.1). This means that $\mathrm{d}\rho/\mathrm{d}z$ is 0.5% greater than by the Williamson–Adams equation. Interpreted as an error in the inferred temperature gradient, it amounts to ~100 K over the depth of the lower mantle, but earth models are not so accurate that this is regarded as serious.

When self-compression, as represented by the Williamson–Adams equation, is allowed for, a simple two-layer model with the observed core size and a rocky mantle cannot match the moment of inertia. It is necessary to bring some mass up higher into the mantle. This was the vital next step to realistic earth models, introduced by K. E. Bullen and B. Gutenberg in the 1930s: the mantle transition zone. By then, seismological information was sufficiently detailed to indicate the depth range of the transition zone and the earth models that followed were quite close to our present understanding. As we now know, the transition zone is a combination of transitions at several levels in the depth range of 220–660 km, where mantle minerals are converted to denser phases. The transition at 660 km is conventionally regarded as the boundary

between the upper and lower mantles. The lower mantle, which has two-thirds of the mantle volume and three-fourths of its mass, appears to be more or less homogeneous over almost the whole depth range to the core, at 2,900 km depth, and conforms well to the Williamson–Adams equation with a density 18% higher than for the low pressure forms of the minerals. The upper mantle is more complicated. Bullen (1975) reviewed this stage in the development of our understanding of the Earth's density.

Free oscillation data

In the 1960s observations of the Earth's modes of free oscillation introduced a new constraint on the density structure. Whereas body wave seismology gives the ratio K_S/ρ throughout the Earth but provides no formal separation of K_S and ρ, the spheroidal modes involve oscillations in the radial distribution of mass with consequent gravitational as well as elastic restoring forces. The equations of motion and the mode frequencies involve not just the ratios K_S/ρ and μ/ρ but also $G\rho$. With G a known constant, the $G\rho$ term allows an assessment of absolute density independent of elasticity. This is illustrated by considering the frequencies of two modes for the simple case of a uniform solid sphere. The spheroidal mode ${}_0S_2$ is an alternation between prolate and oblate deformations and, since there is radial motion, the $G\rho$ term has an important control on the frequency, f, which is given by

$$f^2({}_0S_2) = (2/3\pi)G\rho + (5/3\pi^2)\mu/\rho R^2 \quad (6)$$

whereas the lowest toroidal mode, ${}_0T_2$, is an alternating twist between two hemispheres, with no radial motion, and a frequency given by

$$f^2({}_0T_2) = (7/4\pi^2)\mu/\rho R^2 \quad (7)$$

so that

$$G\rho = (3\pi/2)f^2({}_0S_2) - (10\pi/7)f^2({}_0T_2) \quad (8)$$

The frequencies of more than 500 modes have been identified (Masters and Widmer, 1995), providing a direct measure of global average properties, including the radial density profile. Fine details, such as the sharpness of boundaries, still require analysis of body wave data, but the broad scale structure is derived from the mode frequencies. The Preliminary Reference Earth Model (PREM) of Dziewonski and Anderson (1981) was produced in this way. It signalled a turning point in seismological studies, with less attention to global average properties and increasing emphasis on lateral heterogeneities, boundary structure, and anisotropy.

A powerful new use of free mode data to clarify the radial density structure was presented by Masters and Gubbins (2003), who selected a combination of modes sensitive to a restricted depth range, rather than inverting the whole data set. By concentrating on a region close to the inner core boundary, they obtained an estimate of the density difference between the inner and outer cores, 820 kg m^{-3}, compared with the PREM difference, 597 kg m^{-3}. The revision amounts to 1.8% of the density at that level, compared with the estimated 0.5% precision of the new technique, and may be interpreted as a measure of the revisions to PREM that may be expected in an updated model. From the splitting of mode frequencies, free oscillation data allow identification of lateral density variations independently of elasticity. This is referred to in a later paragraph on lateral heterogeneity.

A density gradient anomaly

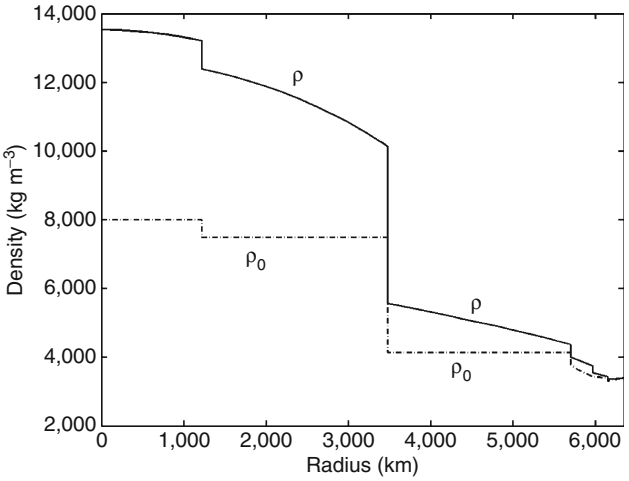
An interesting feature of PREM is the decrease in density with depth in the lithosphere and asthenosphere, over the depth range 24.4–220 km. The obvious interpretation is a steep temperature gradient, but this presents difficulty. The inverted density gradient is slight, little more than 0.1 kg m^{-3} per kilometer of depth, but the difficulty would arise even with zero gradient. For a uniform layer, density is a function only of pressure and temperature, so we can write

$$\begin{aligned} d\rho/dz &= (\partial\rho/\partial P)_T dP/dz + (\partial\rho\partial T)_P dT/dz \\ &= -0.109 \text{ kgm}^{-3}/\text{km} \end{aligned}$$

and substitute the physical properties of mantle material in this depth range to obtain $dT/dz = 8.6$ K/km. (For $d\rho/dz = 0$ this would be 7.9 K/km). Over the depth range of 195.6 km the anomalous gradient gives a temperature range of 1,680 K and since the temperature at 24.4 km (the averaged crust-mantle depth of PREM) cannot be taken as less than about 700 K, the implied temperature at 220 km is close to 2,400 K. This is about 600 K too high; the temperature at 660 km depth required to meet the (P,T) conditions of the phase transition there is about 1900 K (Boehler, 2000) and the 220 km temperature must be lower than that. The discrepancy can be understood if PREM underestimates the effect of anisotropy in the asthenosphere, causing the inverted density gradient as a modelling artifact.

Extrapolation to zero pressure and inferences about composition

The density profile of PREM, with minor equation of state adjustments, is plotted in Figure 1, together with equation of state extrapolations to zero pressure and a temperature of 300 K. The extrapolations use equations of state fitted to each of the ranges, assuming that the materials can be decompressed and cooled without phase changes. Thus, the broken line in the figure represents the densities that the high pressure phases would have under ambient conditions. For each of the inner and outer cores and lower mantle the extrapolated densities are uniform, within the uncertainties of the calculation, and so indicate homogeneity. However, this is not true for the upper mantle, where the model indicates heterogeneity additional to the boundaries that mark recognized phase changes.



Earth, Density Distribution, Figure 1 Density profile of the Earth based on PREM (*solid line*) and the profile of $\rho_{0,0}$, the density obtained by extrapolating to zero pressure and a temperature of 300 K (*broken line*).

The density estimates provide the basic data needed for a discussion of composition. Two long-standing debates that hang on this are the composition of the core, that is, the light elements that must be dissolved in iron-nickel alloy to reduce its density to that of the core, and the possibility of a significant compositional difference between the upper and lower mantles. As Birch (1952) recognized, there is no plausible alternative to iron as the major core constituent, but the outer core is about 10% less dense than pure iron and Poirier (1994) reviewed the arguments for rival candidate light solutes. At core pressures iron occurs as a hexagonal close-packed (ε) phase, for which the extrapolated density would be $\rho_{0,0} = 8{,}352 \pm 23$ kg m^{-3} (Stacey and Davis, 2004), compared with the core density, extrapolated to the same conditions and also in the ε phase, 7,488 kg m^{-3} for the outer core and 7,993 kg m^{-3} for the inner core. The effects on the iron density of each of the candidate light elements are known from metallurgical literature; although none of the observations were made on ε iron, they are taken as a sufficient guide to their potential contributions to the core density deficit. There are undoubtedly many elements dissolved in the core, but the favored important ones, with the mass percentages required in the outer core if each were the only light ingredient, are H (1.4%), C (10.6%), O (12.7%), Si (17.7%), and S (18.2%). The core densities can, in principle, be explained by any combination of these elements, but geochemical arguments restrict the range. The mass percentages of each, in the inner and outer core, respectively, favored by Stacey and Davis (2008, Table 2.5) are H (0.07%, 0.08%), C (0.45%, 0.50%), O (0.11%, 5.34%), Si (nil, nil), and S (8.02%, 8.44%).

There are two main arguments bearing on the similarity of the compositions of the upper and lower mantles, the seismological evidence for penetration of the 660-km boundary by subducting slabs and the densities of high pressure phases of mantle minerals. The perovskite phase of $(Mg,Fe)SiO_3$ is believed to be the dominant lower mantle mineral. It is produced by high pressure laboratory experiments and survives in a metastable state at $P = 0$ provided it is not heated. With no Fe the density at ambient conditions is 4,107 kg m^{-3}. The second most important lower mantle mineral is almost certainly ferro-periclase (magnesiowustite), (Mg, Fe)O, which may account for as much as 20% of the composition. It is less dense than the perovskite (3,585 kg m^{-3} for MgO) but more readily accepts substitution of Fe for Mg and would be 5,956 kg m^{-3} for FeO. The magnesium silicate perovskite can accept Al ions, but neither of these minerals readily accepts Ca, so a calcium perovskite is believed to occur also, but, unlike the magnesium perovskite and periclase, it does not survive decompression and is not as well studied. It is evident that, with a modest iron content, a mixture of these minerals can match the lower mantle density, $\rho_{0,0} = 4{,}144$ kg m^{-3}, and that, on the basis of density, there is no evidence for a compositional difference between the upper and lower mantles.

Lateral density variations

Convection is ubiquitous in the Earth and this means that lateral density variations, with hot, dilated materials rising, and replaced by denser, cool materials. Heterogeneity of the mantle inferred from body wave tomography is generally consistent with the convective pattern, acknowledging that relics of earlier convective patterns are still apparent, but elastic moduli are more sensitive to temperature than is density and a thermodynamic analysis is needed for interpretation (Stacey and Davis, 2008, Sect. 19.7). The splitting of mode frequencies presents evidence of lateral density variations that does not rely on such interpretation, but, in spite of several detailed studies, the difficulties are such that available results must be regarded as preliminary (Ishii and Tromp, 1999; Masters et al., 2000; Romanowicz, 2001). The general conclusion is that, in addition to temperature variations, compositional heterogeneity of the deep mantle must be invoked to explain the observations, in confirmation of a thermodynamic assessment (Forte and Mitrovica, 2001; Stacey and Davis, 2008, p. 288) of results of body wave tomography (Robertson and Woodhouse, 1996; Su and Dziewonski, 1997).

With respect to the outer core, observable lateral heterogeneity must be discounted as quite implausible, in view of the speed of convective motion apparent from the geomagnetic secular variation. Compositional layering, with a stable low density layer at the top of the core has been suggested (Braginsky, 1999), but is not generally supported.

Summary

The Earth's density structure is now known in sufficient detail to provide a sound basis for discussions of internal

composition. High pressure equations of state (Stacey and Davis, 2004), the modern version of finite strain theory, provide reliable extrapolations of density to zero pressure, as well as distinguishing internal heterogeneities caused by temperature variations from compositional and mineralogical differences. Although fine details in the deep interior are not all resolvable, it is evident that the bulk of the Earth does not have very strong lateral heterogeneities, such as are obvious in the crust. Nevertheless, the heterogeneities that are seen are indications of the dynamics of the Earth, making them prime targets for continuing investigation.

Bibliography

Birch, F., 1952. Elasticity and constitution of the Earth's interior. *Journal of Geophysical Research*, **57**, 227–286.

Boehler, R., 2000. High pressure experiments and the phase diagram of lower mantle and core materials. *Reviews of Geophysics*, **38**, 221–245.

Braginsky, S. I., 1999. Dynamics of the stably stratified ocean at the top of the core. *Physics of the Earth and Planetary Interiors*, **111**, 21–34.

Bullen, K. E., 1975. *The Earth's Density.* London: Chapman and Hall.

Dziewonski, A. M., and Anderson, D. L., 1981. Preliminary reference earth model. *Physics of the Earth and Planetary Interiors*, **25**, 297–356.

Forte, A. M., and Mitrovica, J. X., 2001. Deep-mantle high viscosity flow and thermochemical structure inferred from seismic and geodynamic data. *Nature*, **410**, 1049–1056.

Ishii, M., and Tromp, J., 1999. Normal-mode and free-air gravity constraints on lateral variations in velocity and density of Earth's mantle. *Science*, **285**, 1231–1236.

Masters, G., and Gubbins, D., 2003. On the resolution of density within the Earth. *Physics of the Earth and Planetary Interiors*, **140**, 159–167.

Masters, G., and Widmer, R., 1995. Free oscillations: frequencies and attenuation. In Ahrens, T. J. (ed.), *A Handbook of Physical Constants 1: Global Earth Physics*. Washington: American Geophysical Union, pp. 104–125.

Masters, G., Laske, G., and Gilbert, F., 2000. Matrix autoregressive analysis of free oscillation coupling and splitting. *Geophysical Journal International*, **143**, 478–489.

Oldham, R. D., 1906. The constitution of the interior of the Earth as revealed by earthquakes. *Quarterly Journal Geological Society of London*, **62**, 456–475.

Poirier, J.-P., 1994. Light elements in the Earth's core: a critical review. *Physics of the Earth and Planetary Interiors*, **85**, 319–337.

Robertson, G. S., and Woodhouse, J. H., 1996. Ratio of relative S to P heterogeneity in the lower mantle. *Journal of Geophysical Research*, **101**, 20041–20052.

Romanowicz, B., 2001. Can we resolve 3D density heterogeneity in the lower mantle? *Geophysical Research Letters* **28**(6), 1107–1110.

Stacey, F. D., and Davis, P. M., 2004. High pressure equations of state with applications to the lower mantle and core. *Physics of the Earth and Planetary Interiors*, **142**, 137–184.

Stacey, F. D., and Davis, P. M., 2008. *Physics of the Earth*, 4th edn. Cambridge: Cambridge University Press.

Su, W. J., and Dziewonski, A. M., 1997. Simultaneous inversion for 3D variations in shear and bulk velocity in the mantle. *Physics of the Earth and Planetary Interiors*, **100**, 135–156.

Williamson, E. D., and Adams, L. H., 1923. Density distribution in the Earth. *Journal of Washington Academy of Sciences*, **13**, 413–428.

Cross-references

Geodesy, Physical
Gravity Measurements, Absolute
Gravity, Global Models
Seismic Velocity-Density Relationships

EARTH'S STRUCTURE, CORE

Lianxing Wen
Department of Geosciences, State University of New York at Stony Brook, Stony Brook, NY, USA

Definition

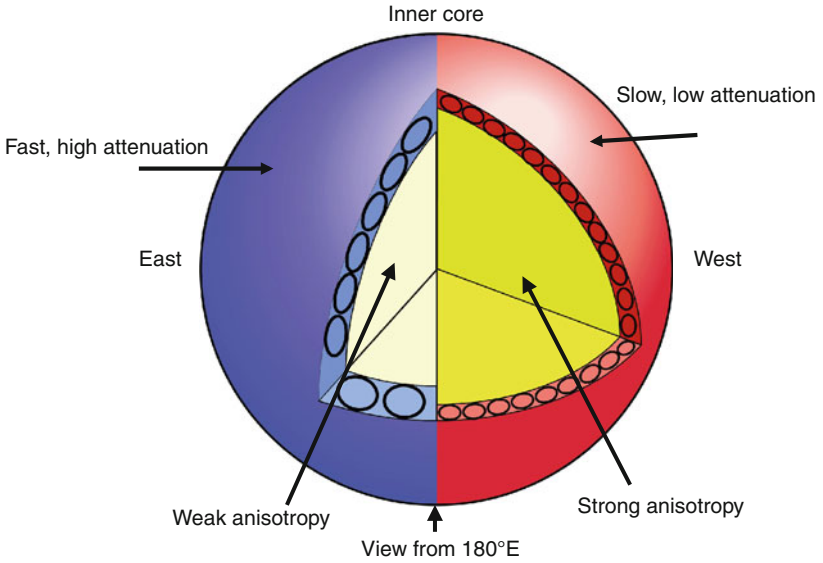
The Structure of Earth's core is referred to as the properties of velocity, density, attenuation, anisotropy, and composition of Earth's core, and how these properties change with depth, geographic location, and time.

Attenuation is a measure of material's ability to absorb energy as seismic waves pass through it.

Anisotropy is a material property that is directionally dependent.

Overall structure: The Earth's core occupies the center portion of Earth with a radius of 3,480 km. It consists of a liquid outer core and a solid inner core. Earth's core is composed of Fe/Ni and some minor light elements, such as S, C, O, and Si. Those light elements are preferentially enriched in the outer core. Seismic compressional velocity and density exhibit sharp jumps at the inner core boundary at the radius of 1,221 km. The inner core is anisotropic in both velocity and attenuation, with seismic waves exhibiting higher velocities and higher attenuation as they propagate along the polar paths than along the equatorial paths. The anisotropy in the inner core is probably caused by preferential alignment of anisotropic Fe crystals.

Radial and lateral variations: With possible exception at the bottom and the top, the outer core is homogeneous due to vigorous mixing of the liquid iron/nickel. The inner core exhibits a complex structure, with the most prominent feature being the hemispheric difference in seismic velocity, attenuation, and anisotropy, between the "eastern" hemisphere (defined between longitudes from 40°E to 180°E) and the "western hemisphere" (defined between longitudes from 180°W to 40°E). The eastern hemisphere exhibits a higher velocity, higher attenuation, weaker anisotropy, and a larger transitional depth from top isotropy to deep anisotropy. The top 235 km of the eastern hemisphere also exhibits a flat velocity gradient with depth (Figure 1). At depth, the innermost 300 km of the inner core may have different forms of anisotropy from the rest of the inner core.



Earth's Structure, Core, Figure 1 Cartoon illustration of hemispheric differences in seismic velocity, attenuation, anisotropy, and depth extent of anisotropy in the Earth's inner core.

Temporal changes: In some localized regions, the inner core surface changes its radius by several kilometers in a decadal scale. Physical mechanisms for such rapid change remain unknown.

Bibliography

Cao, A., Romanowicz, B., and Takeuchi, N., 2005. An observation of PKJKP: inferences on inner core shear properties. *Science*, **308**, 1453–1455.

Ishii, M., and Dziewonski, A. M., 2002. The innermost inner core of the Earth: Evidence for a change in anisotropic behavior at the radius of about 300 km. *Proceedings of the National Academy Sciences*, **99**, 14026–14030.

Morelli, A., Dziewonski, A. M., and Woodhouse, J. H., 1986. Anisotropy of the inner core inferred from PKIKP travel times. *Geophysical Research Letters*, **13**, 1545–1548.

Niu, F., and Wen, L., 2001. Hemispherical variations in seismic velocity at the top of the Earth's inner-core. *Nature*, **410**, 1081–1084.

Tanaka, S., and Hamaguchi, H., 1997. Degree one heterogeneity and hemispherical variation of anisotropy in the inner core from PKP(BC) – PKP(DF) times. *Journal of Geophysical Research*, **102**, 2925–2938.

Wen, L., 2006. Localized temporal change of the Earth's inner core boundary. *Science*, **314**, 967–970, doi:10.1126/science.1131692.

Yu, W., and Wen, L., 2006. Inner core attenuation anisotropy. *Earth and Planetary Science Letters*, **245**, 581–594.

Cross-references

Differential Rotation of the Earth's Inner Core
Earth, Density Distribution
Earth's Structure, Global
Seismic Anisotropy

EARTH'S STRUCTURE, CONTINENTAL CRUST

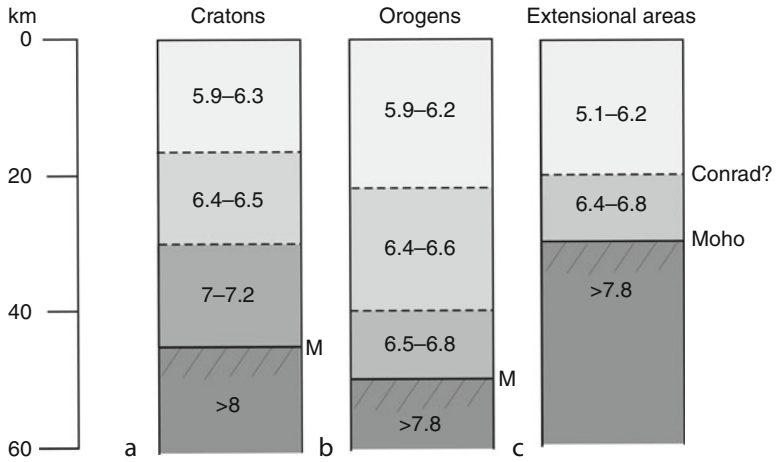
Rolf Meissner, Hartmut Kern
Institut für Geowissenschaften, Universität Kiel, Kiel, Germany

Introduction

In the last 20 years much progress has been made in the development of geophysical techniques that increased our knowledge about the variability of crustal structure substantially. A big misunderstanding, still often presented today, is the use of the phrase "THE" crust. This is a marked oversimplification. Continental crust has been found to be extremely different. There are at least three main types to be distinguished:

1. Thick, old cratons, shields and platforms
2. Orogens, often related to continental subduction
3. Extensional areas, including shelves and rifts

Thick Precambrian cratons, disrupted by early plate tectonics, are presently found in all continents; they contribute more than 70% to the continental lithosphere. They are characterized by a thick crust (45–50 km) with high velocities (>7.2 km/s) in the lower part. The two other crustal types were generated by modern plate tectonics: the *orogens* by convergence and plate collisions, most of them showing deep crustal roots, and the *extensional areas* by divergence and wrench tectonics, often accompanied by thermal and magmatic processes. They generally show a shallow crustal thickness without a high-velocity lower crust. Special forms are rifts, basins, and igneous



Earth's Structure, Continental Crust, Figure 1 Type sections of the three main crustal structures (**a**–**c**) and ranges of P-wave velocities in the upper, middle, and lower crust. Numbers are velocities in km/s.

provinces. Type sections of the three main crustal structures are presented in Figure 1.

In general, the continental crust has lower seismic velocities and consists of lighter material than the underlying mantle; V_P is smaller than 7.8 km/s, V_S smaller than 4.3 km/s, and ρ below 3.1 g/cm^3. Crustal thickness is between 20 km (shelves) and more than 70 km (some orogens) (Mooney et al., 2005). Mohorovičić was the first to determine the base of the crust, today called the "Moho" for short. It is observed worldwide and is clearly recognized by seismic methods. It is worthy to note that the position of the seismic Moho is not always identical with the petrological Moho (defined as the boundary between non-peridotitic crustal rocks and olivine-dominated rocks). In case mafic rocks are transformed into eclogites at the base of orogenically thickened crust, P-wave velocities of eclogites largely overlap with those of peridotites so that seismic field studies may not detect eclogites as crustal rocks (e.g., Mengel and Kern, 1992). An intracrustal boundary – the "Conrad" – seems to separate upper sialic from more mafic lower crust in some areas. Transport of crustal material into the mantle and delamination of the lithosphere and the lowermost part of a thickened crust may give rise to a reduction of crustal thicknesses.

Methods

Reliable studies of the continental crust were started in the late 1940s by seismic methods, applying first refraction studies with man-made explosions and lines of simple geophones. Wide-angle reflection seismics followed, and steep-angle reflection seismics became more and more important. In the last 30 years, digital technology, computer programs for ray-tracing and tomography, and large national and international research programs with an increased number of recording devices and energy sources were initiated. Large, national reflection programs, often with a substantial support by the industry or large funding agencies were developed in nearly all industrialized countries, spearheaded by COCORP in the USA. Both methods – refraction-wide-angle and steep-angle reflection – complement each other, the former supplying velocity-depth functions and relations to petrology, the latter providing reflectivity images of the subsurface that reveal past tectonic processes. Both methods form the backbone of crustal interpretation. They confirmed identical Mohos, either by a jump of velocities or by termination of (crustal) reflectivity because the mantle is less reflective (Meissner and Brown, 1991). Vibroseis or dynamite shooting are used for reflection programs, the latter one generally providing a higher energy output, which is sometimes used for piggy-back wide-angle work. In the last 20 years, a relatively cheap seismological method, the "seismic receiver function technique," supported interpretation of seismic data, making use of global seismicity and a wave change at seismic boundaries. This "seismic receiver function technique" is based on seismic phase conversions of teleseismic waves at the Moho and at boundaries in the crust and mantle.

Non-seismic geophysical techniques are complementary to seismic methods: gravity, for instance, for investigating isostatic problems or selecting between competing seismic interpretations or – together with magnetics – looking for inhomogeneities or controlling structure. Electromagnetic and magnetotelluric methods help to find and to interpret conductivity anomalies. Measurements of heat flow help to observe heat sources and decipher the thermal or magmatic history of a specific area. Depending on the heat flow, temperature-depth functions and viscosities that control rheology can be estimated. The brittle upper crust extends to a depth of 12–20 km and is a zone where generally earthquakes occur. It is separated from the more ductile, aseismic part of the middle and lower crust (some

exceptions). Various experimental studies of creep, stress, and strength support seismic interpretation.

Cratons

Cratons are relatively stable Precambrian continental units with a rather deep lithosphere (Figure 1a). They are called shields when a sediment cover is lacking and platforms when covered by sediments. Today, cratons are found in all continents, possibly created from a few early Supercontinents that were later disrupted by plate tectonics. No consensus exists about the number of cratons. They comprise up to ten in Africa and even more in North America. The latter were finally combined into the big "North American Craton" forming the whole interior of North America (without its margins in the west and east). Even "stable" cratons are influenced by Proterozoic and Phanerozoic processes like rifts or plumes. In North America there is a 2,500 km long Proterozoic Midcontinental rift with a thick cover of sediments, volcanic intrusions, and a reduced crustal thickness. There are also Paleozoic rifts in the southern part of the East European Craton, and there is an unusual 60 km thick crust in the cold Baltic Shield, well within the eclogite stability field in the lowermost part. However, high velocities (>7.8 km/s) corresponding to eclogitic rocks are not observed in the lowermost crust of the Baltic Shield (Korja et al., 2009).

The crystalline crust of cratons is similar worldwide. Cratons are characterized by low heat-flow values. Their thickness is in the range 35–55 km with a "classical" 3-velocity structure (Pawlenkowa, 1996), as shown by wide-angle plus refraction studies. In general, cratons have an upper crust with V_P around 6 km/s, a middle crust around 6.5 km/s, and a roughly 10-km thick lower crust with velocities around 7–7.2 km/s, sometimes 6.7–7.5 km/s (Figure 1a). A typical example is the EUROBRIDGE profile (Guterch et al., 1997) crossing the crust of the East European platform (Figure 2).

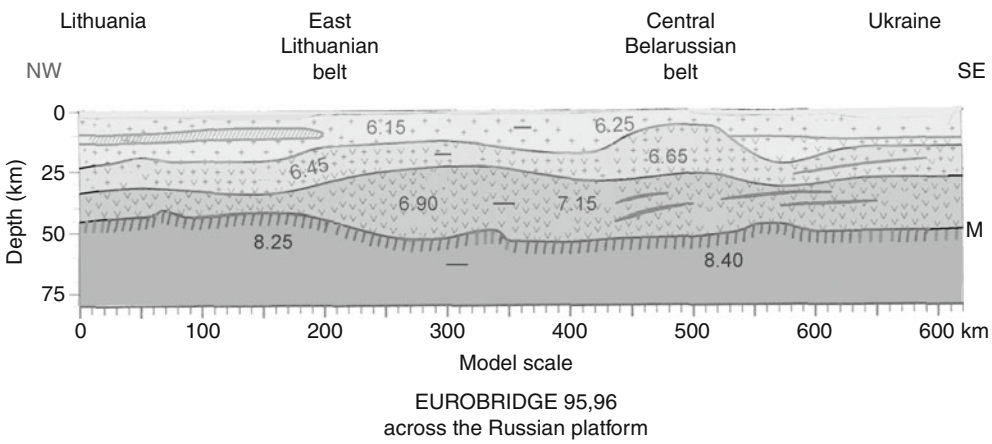
Reflection seismic experiments generally show no special enhancement of reflectivity in the (mafic?) lower crust, as shown by the large transects of LITHOPROBE in Canada, by the COCORP data in the USA or by the BABEL surveys in the northern Baltic Sea (some exceptions).

Orogens

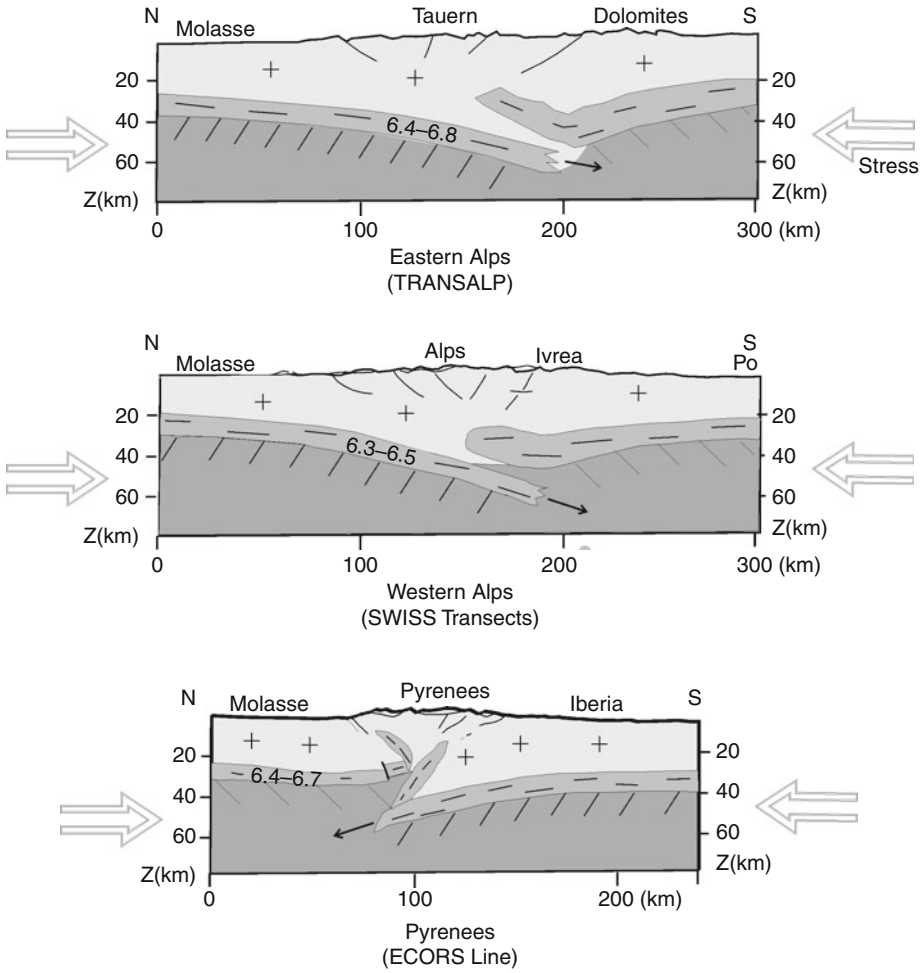
Collisions of plates or terranes create orogens by compressing material upward (mountain ranges) and downward (roots), and dynamic processes initiate oceanic or continental subduction. While the elevation of an orogen apparently depends on the time and speed of the colliding plates (example: Himalaya, Tibet), processes of crust/mantle interaction are quite complex. Magmatic processes are found behind subducting plates in many orogens (Alps, Sierra Nevada, Andes). In continental collision zones, rock types of the lower crust resemble those of the neighboring plates, guided into the orogens. There are high-velocity (cratonic) crustal rocks with mafic lower crusts in the Uralides (from the East European Platform) and sialic, extensional crusts with a reflecting lower crust in the Alps (mostly from the Variscides). Continental subduction often follows oceanic subduction. Remnants of these processes may be reflected by the crustal "roots." Figure 3 shows simplified cross sections through the Alps and Pyrenees. Orogens are more or less isostatically compensated; its Bouguer Anomaly is strongly negative (Ebbing, 2004).

Extensional areas

Extensional processes cover many dimensions: from rifts, basins, (the Basin and Range of North America), continental shelves up to large Phanerozoic areas like Europe, west of the Tesseyre-Tornquist line. This large extensional area was created in Post-Avalonian, Post-Variscan time at the end of the Paleozoic. A surprisingly uniform crust has been formed, showing three prominent differences



Earth's Structure, Continental Crust, Figure 2 Typical cratonic crust from the East European platform (after Guterch et al., 1997).



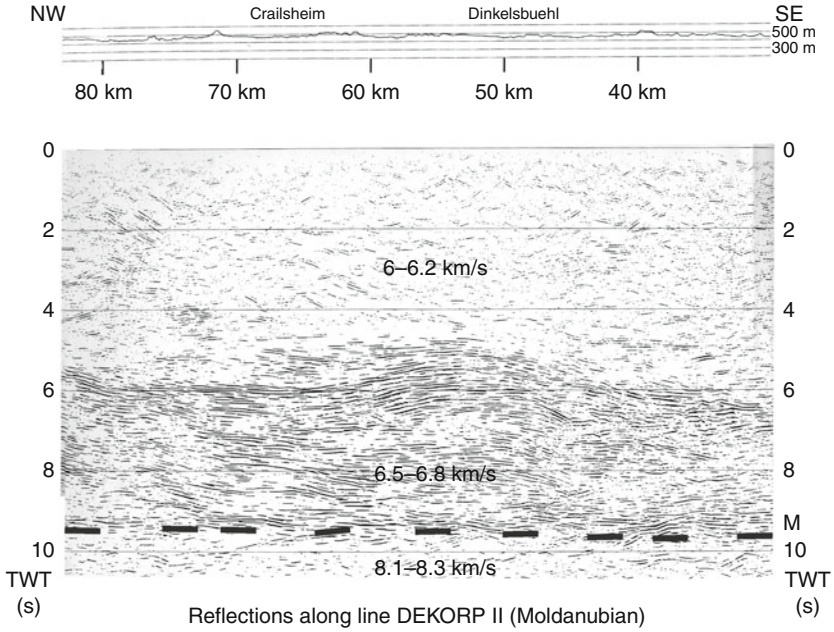
Earth's Structure, Continental Crust, Figure 3 Typical sections of orogenic crust (Alps and Pyrenees).

compared to the cratons: their thickness is only 30 km (Figure 1c), their velocities being lower than 6.8 km/s, and in many places they show prominent reflectivity (lamellae) in the lower crust.

Apparently, all three characteristic crustal patterns are genetically interrelated. While a former lower crust (in the Basin and Range province and in the Variscides) got lost in the previous orogeny (by delamination or subduction) subsequent moderate, but widespread, extension dominated. Mafic-ultramafic dykes intruded into the lower crust and formed sill-like intrusions, giving rise to high impedance contrasts after cooling. Importantly, sills are mostly rather thin and do not generally increase the lower crustal overall (sialic) seismic velocities. Figure 4 gives an example of strong lower crust reflectivity and low average velocities.

In contrast, the formation of rifts and basins is a result of local (extensional) stress and/or strong local heat anomalies. As a consequence, large volumes of mantle-derived basaltic magmas intrude into the crust, and after cooling, the mafic rocks (gabbros or metamorphic equivalents) cause the observed high velocities at the base of most rifts and basins. Examples are the Oslo Graben, the Danish Basin, and other adjacent rifts and basins north of the Variscan front, the great East African rift, the Donetz Basin, and the Mid-continental North American rift. In general, the massive intrusions created a transparent lower crust, but some basins with high lower crust velocities (above 7.2 km/s) show some thin lamellae (apparently ultramafic?) on top of the Moho. Also in some areas of continental/ocean transition high-velocity lamellae are observed in the lower crust (Roberts et al., 2009).

In general, lower crustal reflectivity seems to depend on the volume and thickness of the mafic intrusions compared to the host rock. Whereas sill-like intrusions are suggested to generate the observed lower crustal reflectivity, massive intrusions in the lower crust reduce reflectivity and enhance the average velocities to values of 6.8–7.2 km/s. The volcanic Messum ring complex in Namibia shows no lamellae in the high-velocity interior, but a good reflectivity in the surrounding (Bauer et al., 2003). In contrast, the Viking Graben in the North Sea



Earth's Structure, Continental Crust, Figure 4 Multiple reflections (laminations) in the Moldanubian lower crust along the DEKORP II line, along with ranges of V_P-velocities at different depths.

shows lower crust reflectivity and no reflection around it (McBride et al., 2004).

Most continental shelves show extension and a stretched continental crust with a thickness of only 20–30 km, often containing aborted rifts (Korja et al., 2009). Reflectivity in the lower crust is found in the West Atlantic margin (Faroe or Hatton) and along the Greenland coast. Often, seaward dipping reflections, caused by intrusive and extrusive magmatism, mark the onset of oceanic crust with high velocities. BIRPS's WAM–profile from the British Isles toward the Atlantic ocean shows a preferred lower crust reflectivity below the (continental) Mesozoic basins, while the oceanic crust is free from reflections.

Constraints on crustal petrology

The measured seismic velocity structures are very helpful in providing a rough characterization of the earth's crust in different tectonic environments, but they are nonunique so that inferences about composition cannot be drawn from wave velocities (at least P-wave velocities). The seismic properties at depth are determined by a number of lithologic and physical factors that control the in situ rock properties in a very complex manner. Velocities are controlled by the intrinsic properties of the rocks (mineralogical composition, chemical composition, metamorphic grade, crystallographic preferred orientation of constituent minerals, etc.) and by the physical environment of the crust (temperature, pressure, porosity, fluid content, etc.).

Direct information about the composition and structure of the deep crust can be obtained either from crustal terrains exposed at the surface (e.g., Fountain and Salisbury, 1981; Kern and Schenk, 1988) or from xenoliths brought to the surface by magmas (Downes et al., 1990). Deeply eroded Precambrian terrains (e.g., Sri Lanka), and upthrust tectonic slices in orogenic belts (e.g., Ivrea Zone, N. Italy; Serre Mountains, S. Calabria; Kapuskasing Zone, Canada) provide perhaps the best geologic guides to structural style and composition at depth. Such rocks are important in providing direct data although they contain a mixed message as they are no longer the in situ deep crust.

A unique ability to correlate the seismic data with the structure and composition of the in situ deep crust can be provided by coupling experimentally determined or calculated P- and S-wave velocities for relevant crustal rocks collected from surface outcrops or from xenoliths, simulating in situ conditions: (1) by laboratory seismic measurements at actual PT conditions (e.g., Christensen and Wepfer, 1989; Kern et al., 1999) and (2) by calculations from modal analyses and elastic properties of the rock-forming minerals and their pressure and temperature derivatives (e.g., Jackson et al., 1990).

High P-wave velocities (>6.9 km/s) generally defining the lowermost crust (below about 20–25 km depth) are typical for metamorphic rocks in the granulite facies. These are mafic granulites, mafic amphibolite facies rocks, anorthosites, and high-grade metapelites.

Intermediate P-wave velocities (6.5–6.9 km/s) represent the upper-lower to mid-crust and are characteristic for intermediate granulites and metapelites.

Low P-wave velocities (6.2–6.5 km/s) occurring at mid-crustal levels in Palaeozoic and more recent fold belts

and in the uppermost crust in shields and platforms are likely to be composed of felsic rocks in the amphibolite and granulite facies. It should be noted, however, that not all felsic (granitic) upper crustal rocks are intermediate- to high-metamorphic grade. The Superior Province (Canada) is an example of low metamorphic shield rocks.

Mineralogically, feldspar (K-feldspar, plagioclase) is the most abundant mineral, followed by quartz and hydrous minerals (such as mica and amphiboles).

The bulk chemical composition of the upper crust is felsic, and the middle crust is intermediate in composition. The average composition of the lower crust is in general mafic in cratons, but it tends to be intermediate or even felsic in some regions. The continental crust as the whole (average of all crustal types) is suggested to be intermediate in average bulk composition (Rudnick and Fountain, 1995).

Importantly, most of the rocks constituting the earth's crust exhibit marked anisotropy of elastic properties (e.g., Kern et al., 2001; Ji et al., 2002). Typical values of intrinsic P-wave anisotropy (quartz-mica-schists, felsic gneisses, granulite-facies metapelites, amphibolites) range from 5.4% to 10.7%. Anisotropy may be caused by crystallographic (CPO) and morphological (or shape) preferred orientation (SPO), by oriented microcracks, or by thin layers of otherwise isotropic materials with different properties.

Whereas oriented cracks may contribute a large fraction of observed seismic anisotropy in the uppermost crust (Crampin, 1987), CPO- and SPO-related seismic anisotropy is a major contribution in ductily deformed and foliated crustal structures (shear zones, gneiss sections) of the deeper crust. It is mainly due to the alignment of mica and hornblende minerals and their strong single crystal anisotropies (Meissner and Kern, 2008). In addition to mineral and chemical composition and rock fabric, pore fluids may have an important bearing on in situ seismic properties. Saturation of pore space ($P_{pore} \approx 0$) increases V_P, whereas V_S remains unaffected (Nur and Simmons, 1969). In case of pore pressure approaching lithostatic pressure ($P_{pore} \approx P_{lith}$), microcracks and grain boundaries are more or less kept open. As a consequence, both P- and S-wave velocities are smaller than in dry rocks, due to a reduction of the effective pressure ($P_{eff.} = P_{lith.} - n \times P_{pore}$; $n \approx 1$). Importantly, V_P-anisotropy is significantly higher in dry rocks than in water-saturated rocks at atmospheric pressure conditions (Popp and Kern, 1994). As pressure is increased, differences progressively decrease. The corresponding S-wave velocities and S-wave splitting data are only weakly affected by intergranular fluids.

Seismic anisotropy is becoming increasingly important also in understanding the evolution of the Earth's crust (and upper mantle). Since the measured seismic anisotropy has a structural (macro- and/or micro-) origin, seismic azimuthal anisotropy manifested by shear wave splitting, in particular, can provide important geophysical evidence of deformation because the orientation and magnitude of anisotropy is, in general, strongly related to the internal strain-induced rock fabric. Although variations of lithology is suggested to be most important for the generation of seismic reflections in the lower crust (laminated lower crust) (e.g., Meissner et al., 2006), seismic reflectivity may be enhanced by the fabric-related anisotropy of ductily deformed structures (Kern and Wenk, 1990). It should be noted, however, that in crustal rocks the variability due to variation in composition is generally higher than the variation caused by velocity anisotropy (except for schists). The reverse is true in the olivine-rich upper mantle rocks. This probably implies that, on a regional scale, fabric-related anisotropy is more important in the oceanic and continental upper mantle than in the continental crust.

Summary

Seismic refraction and reflection surveys have revealed three main types of continental crust: (1) the thick old cratons, constituting large parts of all continents; (2) the orogens, generated by plate collisions; and (3) the extensional areas with a relatively thin crust, exhibiting ample reflectivity (lamellae) in the lower part. Crustal thickness is between 20 km (shelves) and more than 70 km (some orogens). The Mohorovičić discontinuity (Moho), defined by a jump in velocity and density, separates the lowermost crust from the upper mantle. High P-wave velocities (>6.9 km/s) are typical for the lowermost crust of platforms and orogens; they stand for high-grade mafic rocks in the granulite facies. In the thin lower crust of extensional areas, high-velocity mafic rocks are almost lacking. The upper-lower to mid-crust is characterized by intermediate P-wave velocities (6.5–6.9 km/s) and is likely to be composed of intermediate granulites and metapelites. Low P-wave velocities (6.2–6.5 km/s) of mid-crustal levels in Paleozoic and more recent fold belts as well as in the uppermost crust in shields and platform point to felsic rocks in the amphibolite and granulite faces. Seismic anisotropy is an important property of most rocks constituting the earth crust. It is basically caused by crystallographic and morphological (or shape) preferred orientation, by oriented microcracks, or by thin layers of otherwise isotropic material with different properties.

Acknowledgments

We greatly appreciate the assistance of Gudrun Reim (Kiel) in preparing the figures, and we thank Walter Mooney (Menlo Park) for his review leading to many improvements of the manuscript.

Bibliography

Barruol, G., and Kern, H., 1996. Seismic anisotropy and shear-wave splitting in lower-crustal and upper-mantle rocks from the Ivrea Zone – experimental and calculated data. *Physics of the Earth and Planetary Interiors*, **95**, 175–194.

Bauer, K., Trumbull, R. B., and Victor, T., 2003. Geophysical images and a crustal model of the intrusive structures beneath the Messum ring complex, Namibia. *Earth and Planetary Science Letters*, **216**, 65–80.

Christensen, N. I., and Mooney, W. D., 1995. Seismic velocity structure and composition of the continental crust: a global view. *Journal of Geophysical Research*, **100**, 9761–9788.
Christensen, N. I., and Wepfer, W. W., 1989. Laboratory techniques for determining seismic velocities and attenuations, with applications to the continental lithosphere. In Pakiser, L. C., and Mooney, W. D. (eds.), *Geophysical Framework of the Continental United States*. Boulder: Geophysical Society of America Memoir, p. 172.
Crampin, S., 1987. The geological and industrial implications of extensive-dilatancy anisotropy. *Nature*, **328**, 491–496.
Downes, H., Dupuy, C., and Leyreloup, A., 1990. Crustal evolution of the Herzynian belt of Western Europe: evidence from lower crustal granulitic xenoliths. *Chemical Geology*, **68**, 291–303.
Ebbing, J., 2004. The crustal structure of the Eastern Alps from a combination of 3D gravity modeling and isostatic investigations. *Tectonophysics*, **350**, 89–104.
Fountain, D., and Salisbury, M. H., 1981. Exposed cross sections through the continental crust; implications for the crustal structure, petrology and evolution. *Earth and Planetary Science Letters*, **56**, 263–277.
Guterch, A., Grad, M., Thybo, H., Keller, G. R., and POLONAISE Working Group, 1997. POLONAISE '97 – an international seismic experiment between Precambrian and Variscan Europe in Poland. *Tectonophysics*, **314**, 101–121.
Jackson, I., Rudnick, R. L., O'Reilly, S. Y., and Bezant, C., 1990. Measured and calculated elastic wave velocities for xenoliths from the lower crust and upper mantle. *Tectonophysics*, **173**, 207–210.
Ji, S. C., Wang, Q., and Xia, B., 2002. *Handbook of Seismic Properties of Minerals, Rocks, and Ores*. Montreal: Polytechnic International Press, p. 630.
Kern, H., and Schenk, V., 1988. A model of velocity structure beneath Calabria, South Italy, based on laboratory data. *Earth and Planetary Science Letters*, **87**, 325–337.
Kern, H., and Wenk, H.-R., 1990. Fabric related velocity anisotropy and shear wave splitting in rocks from the Santa Rosa mylonite Zone, California. *Journal of Geophysical Research*, **95**, 11213–11223.
Kern, H., Gao, S., Jin, Z., Popp, T., and Jin, S., 1999. Petrophysical studies on rocks from the Dabie ultrahigh-pressure (UHP) metamorphic belt, Central China: implications for the composition and delamination of the lower crust. *Tectonophysics*, **301**, 191–215.
Kern, H., Popp, T., Gorbatsevich, F., Zharikov, A., Lobanov, K. V., and Smirnov, Yu. P., 2001. Pressure and temperature dependence of V_P and V_S in rocks from the superdeep well and from surface analogues at Kola and the nature of velocity anisotropy. *Tectonophysics*, **338**, 113–134.
Korja, A., Hyönen, T., Tira, T., and Heikkinen, P., 2009. Examining three-dimensional crustal heterogeneity in Finland. *EOS, Transactions American Geophysical Union*, **90**(15), 129–130.
McBride, J. H., White, R. S., Smallwood, J. R., and England, R. W., 2004. Must magmatic intrusion in the lower crust produce reflectivity? *Tectonophysics*, **388**, 271–297.
Meissner, R., and Brown, L., 1991. Seismic reflections from the Earth's crust: comparative studies of tectonic patterns. *Geophysical Journal International*, **105**, 1–2.
Meissner, R., and Kern, H., 2008. Eathquakes and strength in the laminated lower crust – can they be explained by the corset model? *Tectonophysics*, **448**, 49–59.
Meissner, R., Rabbel, W., and Kern, H., 2006. Seismic lamination and anisotropy of the lower continental crust. *Tectonophysics*, **416**, 81–99.
Mengel, K., and Kern, H., 1992. Evolution of the petrological and seismic Moho – implications for the continental crust-mantle boundary. *Terra Nova*, **4**, 109–116.
Mooney, W. D., Rao, V. V., Chulik, G. S., and Detweiler, S. T., 2005. Comparison of the deep crustal structure and seismicity of North Amarica with the Indian Subcontinent. *Current Science*, **88**, 1639–1651.
Nur, A., and Simmons, C., 1969. The effect of saturation on velocity in low-porosity rocks. *Earth and Planetary Science Letters*, **7**, 183.
Pawlenkowa, N., 1996. Crust and upper mantle structure in northern Eurasia from seismic data. In Dmowska, R., and Saltzmann, B. (eds.), *Advances in Geophysics*. New York: Academic, Vol. 37.
Popp, T., and Kern, H., 1994. The influence of dry and water-saturated cracks on seismic velocities of crustal rocks – a comparison of experimental data with theoretical model. *Surveys in Geophysics*, **15**, 443–465.
Roberts, A. W., White, R. S., and Christie, P. A. F., 2009. Imaging igneous rocks on the North Atlantic rifted continental margin. *Geophysical Journal International*, **179**, 1029–1038.
Rudnick, R. L., and Fountain, D. M., 1995. Nature and composition of the continental crust: a lower crustal perspective. *Reviews of Geophysics*, **33**, 267–309.

Cross-references

Deep Seismic Reflection and Refraction Profiling
Plate Tectonics, Precambrian
Sedimentary Basins
Seismic Anisotropy
Seismic, Receiver Function Technique
Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting

EARTH'S STRUCTURE, GLOBAL

Jean-Paul Montagner
Seismological Laboratory, Institut de Physique du Globe, University Paris-Diderot, UMR CNRS/7154, Paris, France

Definition

Structure of the Earth. The manner in which the Earth is constructed. It is characterized through physical and chemical parameters.
Crust. The outer layer of the solid Earth that lies above the Mohorovicic discontinuity.
Mantle. A thick layer of rock below the crust down to the core–mantle boundary CMB at 2,900 km depth.
Core. The central part of the Earth below CMB. It is composed of the fluid outer core (2,900 km down to 5,100 km) and the solid inner core below.
Anelasticity. Property of a substance in which there is no definite (linear) relation between stress and strain.
Anisotropy. Property of a substance being directionally dependent, as opposed to isotropy (identical properties in all directions).

Introduction

The Earth is the largest in size and mass of the Solar System's four terrestrial planets. Of these four planets, Earth also has the highest density, the highest surface gravity, the strongest magnetic field, and fastest rotation.

It is also the third planet from the Sun, and the densest and fifth largest of the eight planets in the Solar System. Earth is a terrestrial planet, meaning that it is a rocky body, rather than a gas giant like Jupiter. Earth is currently the only place where life is known to exist, and the only terrestrial planet with active plate tectonics.

Earth Sciences in general, and Geophysics in particular, are built on the foundations of physics, mathematics, geology, astronomy, and are closely connected to other disciplines such as chemistry and biology. Its roots therefore go far back in history, but the science has blossomed only in the last century with the explosive increase in our ability to measure the properties of the Earth and the processes in play inside the Earth, on and above its surface. Geophysics includes the study of the Earth's solid layers, fluid envelopes, and its near-space environment. However, in this contribution, the subject is narrowed to the solid Earth.

The interior structure of the Earth, similar to the outer and like that of the other terrestrial planets, is divided into layers by their physical (rheological) or chemical properties. Scientific understanding of Earth's internal structure is based on geological, geochemical, and geophysical observations at the surface of the Earth including topography and bathymetry, observations of rock in outcrops, samples brought to the surface from greater depths by volcanic activity. Geophysics can make use of measurements of different physical fields such as the gravity, magnetic fields, and the fields associated with the propagation of seismic waves passing through the Earth. This interdisciplinary observational approach is completed by laboratory experiments with minerals at pressures and temperatures characteristic of the Earth's deep interior.

The technological advances of the last century in laboratory and field instrumentation, computing, and satellite-based remote sensing are largely responsible for the explosive growth of geophysics. These advances enabled geophysicists to make more and more accurate measurements, to collect and to analyze enormous amounts of data, and to model more and more complex systems. This new view of how the Earth works enabled a fundamental understanding of structure of the Earth as well as earthquakes, volcanoes, mountain building, indeed all of geological processes and objects.

1D structure of the earth

The global structure of Earth can be defined in two different ways, either by considering its physical properties, or its chemical properties. Different physical fields are used by geophysicists. The Earth's gravity field can be used to calculate its mass, its moment of inertia. By adding some additional physical constraints, the density variations within the Earth can be calculated. The magnetic field of internal origin provides fundamental information on the deepest layer of the Earth, the Core (see *Magnetic Data Enhancements and Depth Estimation* and *Magnetic Methods, Principles*). Geophysics is the only Earth science discipline enabling the imaging of the invisible deep inner structure of our planet by a direct measurement of its physical properties. To first order, the structure of the earth is dominated by its radial structure (spherically symmetric model). During the first part of the twentieth century, geophysicists determined the onion-like fine structure of the Earth. Seismic waves enable the investigation of the elastic and to some extent anelastic properties of the whole Earth.

Characterizing the interior of the Earth with earthquake waves

Global seismology starts in 1889 when Ernst von Rebeur-Paschwitz in Potsdam (Germany) associated a disturbance recorded on a tiltmeter to the occurrence of a remote Japanese earthquake. Since then, seismologists modeled the propagation of seismic waves from teleseismic events by using the ray theory considering seismic waves as virtual particles following rays between earthquake focus and receivers and their travel times are sensitive to the structure of the Earth.

During the first decades of the twentieth century (see Dziewonski and Romanowicz (2007) for details), the layering of Earth was indirectly inferred using the travel times of refracted and reflected seismic waves generated by earthquakes. The changes in seismic velocity of P-waves (compression waves) and S-waves (shear waves) between different layers cause refraction owing to Snell's law. Reflections are caused by a large increase in seismic velocity and are similar to light reflecting from a mirror. Seismologists were able to find and characterize the main seismic discontinuities (depth and velocity jump) within the different layers of the Earth (see *Seismic Discontinuities in the Transition Zone*). First of all, the Mohorovicic discontinuity between crust and mantle in the depth range 10–60 km then the core–mantle boundary at 2,900 km depth by Oldham (1906). The core does not allow shear waves to pass through it, showing that its outer part is fluid. Finally, Inge Lehmann (1936) demonstrated the existence of a solid inner core.

Mantle itself can be divided into upper mantle and lower mantle by a discontinuity at 660 km depth. In the upper mantle, the finding of a discontinuity around 410 km depth defines the "transition zone" between 410 and 660 km depth. At the base of the mantle, about 200 km above the core–mantle boundary, a discontinuity was discovered, identifying the still mysterious D″-layer. There are some other candidates for entering the club of seismic discontinuities, the Hales discontinuity at 80 km depth, the Gutenberg discontinuity at the base of continental and oceanic lithosphere, the Lehmann discontinuity at 220 km, and also at 520 km and around 900 km depths, but their global nature is still debated. Most seismic discontinuities are interpreted as a mineralogical phase transition between minerals of mantle silicates (olivine–perovskite system) but some of them might be a boundary between different chemical compositions.

From its elastic properties, the different layers are presented in Table 1 and Figure 1. The Earth can be divided into the outer silicate solid crust, the upper mantle (including lithosphere, asthenosphere, and transition zone), the lower mantle (with the D″-layer at its base), the liquid outer core much less viscous than the mantle, and the solid inner core.

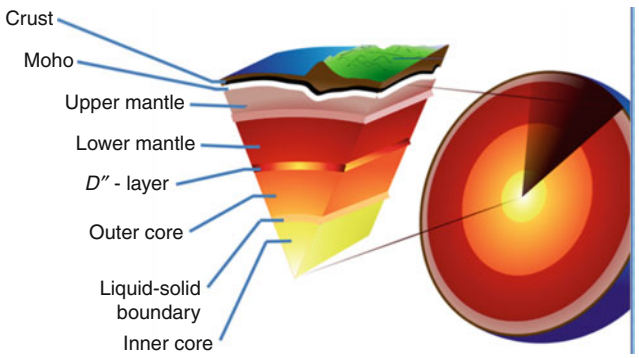
From 1D models to 3D tomographic models

Many reference models (spherically symmetric models) have been developed during the last century, but the most popular ones are the PREM (Preliminary Reference Earth Model by Dziewonski and Anderson, 1981) (Figure 2), IASP91 (Kennett and Engdahl, 1991) and later AK135 (Kennett et al., 1995). PREM introduced for the first time an anisotropic layer (with the specific case of radial anisotropy with a vertical symmetry axis) in the uppermost 220 km of the mantle. Montagner and Kennett (1996) by trying to reconcile body wave and normal mode data showed that a significant amount of radial anisotropy is also necessary in the mantle transition zone in a broad sense between 410 and 900 km depth.

However, important deviations between observed and theoretical travel times of seismic waves with respect to 1D reference models cannot be explained by these simple 1D models. These differences, though relatively small (<10%) are due to the existence of lateral heterogeneities of physical properties within the different layers of the Earth. In the 1960s, the acceptance of the paradigm of plate tectonics and mantle convection leads the seismological community to a better appraisal of the coupling between lithosphere and the underlying mantle convection. Plate tectonics demonstrated that the internal dynamics is very active and that tectonic plates are the surface evidence of deeper convective phenomena. According to geodynamic calculations, large lateral variations in temperature must be associated with lateral variations of seismic velocities. This new approach of global geodynamics has been facilitated by the revolution of broadband seismology in the 1980s (see *Seismological Networks*), which came up with the rapid expansion of seismic networks, the free accessibility of digital data, and more and more powerful computers. Time was ripe for the new era of seismic tomography (imaging of 3D seismic anomalies) and its application to global geodynamics.

Earth's Structure, Global, Table 1 The 1D structure of the Earth

Layer (Km)	
0–80	Lithosphere (locally varies between 5 and 200 km)
0–35	. . . Crust (locally varies between 5 and 70 km)
35–80	. . . Uppermost part of mantle
35–2,890	Mantle
80–220	. . . Asthenosphere
410–660	. . . Transition zone
35–660	. . . Upper mantle
660–2,890	. . . Lower mantle
2,740–2,890	. . . D″-layer
2,890–5,150	Outer core
5,150–6,360	Inner core

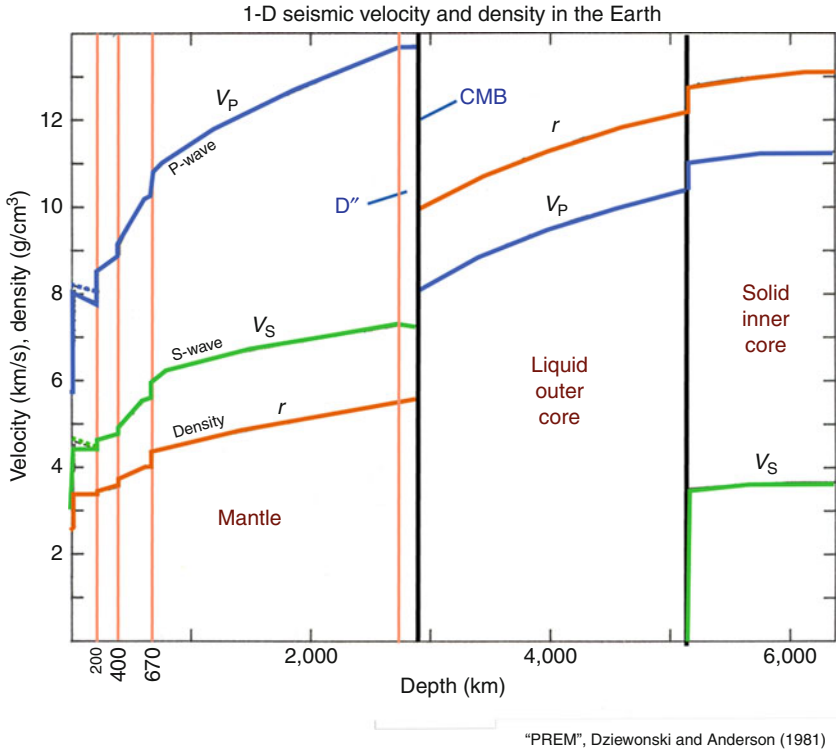


Earth's Structure, Global, Figure 1 The different layers of the Earth.

Mantle convection jigsaw

The investigation of mantle convection is the best example of the implementation of these new available tools in the framework of an interdisciplinary approach. Thermal convection in the Earth has been recognized for a long time (Holmes, 1928; Pekeris, 1935) as the cause of dynamic processes such as mountain building, volcanic and seismic activity, and, more recently, of seafloor spreading. With the advent of plate tectonics, it has become possible to describe quantitatively the motions at the surface of the Earth. The plates, considered as rigid bodies, are the surface manifestation of deep-seated movements related to mantle convection (e.g., Turcotte and Schubert, 2002). This evidence constitutes the first manifestation of pecularities of mantle convection, for it demonstrates large variations in the rheological properties of Earth material between rigid plates and underlying mantle, which cannot be explained by 1D models. Meanwhile, hotspots were identified (Wilson, 1963) and their fixity with respect to plate motions demonstrated (Morgan, 1971). There is a good consensus on the origin of plumes in a thermal boundary layer, but the depth of origin of hotspots is still an open question. The rheology of the lithosphere and the existence of hotspots are two simple examples that illustrate the fact that mantle convection differs in many respects from "classical" convection (such as Benard convection) observed in usual viscous fluids.

The understanding of mantle dynamics is similar to a police investigation: Each earth science field provides miscellaneous and often contradictory clues, which must



Earth's Structure, Global, Figure 2 Preliminary Reference Earth Model (PREM Dziewonski and Anderson, 1981).

be assembled like a jigsaw puzzle pieces. Different approaches have been used to characterize and understand mantle convection (see *Mantle Convection*). None of them could provide a general and consistent explanation of simple features such as the shape and geometry of plates, the origin of hotspots, and the number of layers in the mantle. Many questions arose: How is surface geology (continents, plate boundaries, mountain ranges, hot spots) reflected at depth? How are characterized the interactions and the coupling between the different layers of the Earth? How deep mantle flows affect the surface geology or conversely how surface phenomena affect deep dynamics?

Another issue was the characterization of the spatial scales of heterogeneities at the global scale. Geology makes evident a multiplicity of scales. Some tectonic plates are gigantic (larger than 10,000 km), some others smaller (1,000 km). Continents have as well very various sizes. The different plate boundaries (subduction zones, ridges, transform faults) have a very small lateral extent (a maximum of 100 km except continent–continent collision zones). Hotspots are even smaller and their role in global geodynamics cannot be explained by plate tectonics. This multiscale approach is also present for the temporal scales. Whereas times associated with plate tectonics count in tens of millions of years, catastrophic events such as hotspot birth, mantle avalanches, seismic and volcanic cycles last only for centuries. According to time scales, rheological properties change. Therefore, natural phenomena are by essence, multiscale and make it necessary a multidisciplinary approach. The role of geophysicists consists in quantifying, understanding processes, proposing simple laws in spite of the complexity of processes.

One of the most powerful approach to investigate mantle convection is the numerical approach. The basic equations of physics governing mantle convection are well known (Navier–Stokes equations and thermal equation of energy conservation). This system of equations is very complex and nonlinear, expressed in terms of adimensional numbers such as the Rayleigh number Ra. Actual estimates of Ra make it largely beyond the critical Rayleigh number, which means that mantle convection is thermally turbulent. In other words, mantle convection is chaotic at the geological time scales, meaning that the convective flow is not stationary in time and space. With the rapid expansion of computer facilities, it is now possible to numerically simulate the equation system. However, it turns out difficult to reproduce plate behavior with numerical experiments. Laboratory experiments complete the whole numerical battery, enabling the fine investigation of some specific phenomena. Material science, by studying physical properties of earth minerals (rheology, anelastic and elastic properties) at very high pressure

(by using diamond anvil cells and synchrotrons, etc) also provide important information on mineralogical composition of the mantle. Most seismic discontinuities can be interpreted as pressure- and temperature- induced phase changes in the olivine system (Birch, 1952), such as the 410- and 660-km discontinuities. However, some composition changes might also occur around these depths and deeper in the lower mantle where thermochemical convection might be predominant.

In that context, the determination of 3D models of seismic parameters of the mantle (also coined tomographic models) can be considered as a pragmatic approach that consists in imaging the structure of convection in order to provide clues to the nature of driving mechanism and to investigate mantle convective processes, the engine of plate tectonics. Thanks to the intensive use of computers that can handle very large data sets, the techniques of seismic tomography (see *Seismic Tomography*) can image the 3D deep structure of the Earth and provide unavoidable constraints of possible models of mantle convection and makes it possible to test different competing hypotheses. However, the main limitation of seismic tomography is that it only provides an instantaneous image of our planet Earth and it is necessary to confront seismic results with other Earth science fields providing information on the history of our planet.

Seismic 3D tomography: the Earth scanner

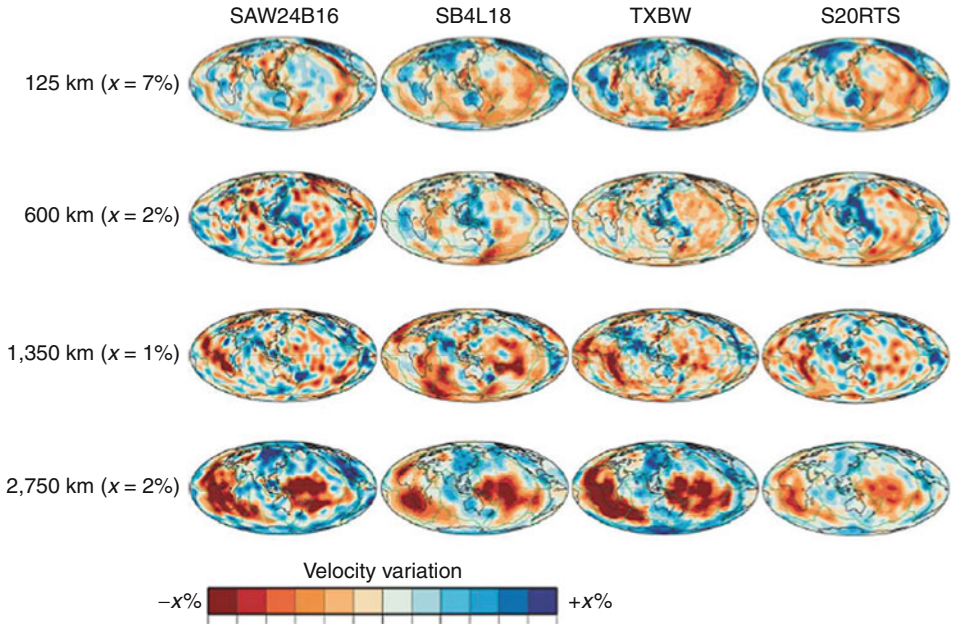
Seismic wave propagation is very sensitive to temperature, chemical, and mineralogical anomalies. These anomalies reflect the active dynamics of the Earth which should be a dead planet without them. It was observed for many decades that seismic velocity below old continents is much larger than the velocity beneath young oceans. This difference between tectonic provinces can be explained by a different history: Cratons older than 1 billion years had a lot of time to get colder, whereas below oceans, the younger and warmer lithosphere is permanently recycled within the mantle. An increase in temperature is usually translated into decrease in seismic velocity, whereas an increase in pressure is translated into velocity increase. At a given depth (approximately constant pressure), the lateral variations of seismic velocity can be primarily interpreted in terms of temperature variations. A large variety of tomographic techniques similar to ultrasonography or scanner has been developed for getting the 3D structure of the Earth, notwithstanding the geological nature of the surface.

The first 3D models of the Earth's mantle were proposed by the end of the 1970s (Dziewonski et al., 1977; Sengupta and Toksoz, 1976), but the true revolution of seismic tomography took place in the mid-1980s when a complete set of tomographic models was proposed from the surface down to the center of the Earth (see Woodhouse and Dziewonski (1984), Nataf et al. (1984), Tanimoto and Anderson (1985) for the upper mantle, Dziewonski (1984) for the lower mantle). The present lateral resolution at global scale for the whole Earth is approximately 1,000 km. Tomographic models (Figure 3) fundamentally renewed our vision of Earth dynamics, and shed new light on deep processes.

At 100 km depth, the distribution of seismic velocity anomalies reflects the geological surface structure. Almost all plate boundaries are slow since they are often associated with active volcanism. In oceanic zones, seismic velocity increases with the age of the sea floor confirming the cooling of plates with age. This remarkable agreement between observed seismic structure and structure deduced from geology was the first proof of the validity of tomographic approach. At larger depths, the correlation between surface geology and seismic structure progressively vanishes and the amplitude of seismic anomalies decreases. The continental roots can be found down to 250–300 km but mid-ocean ridges are no longer visible. In contrast, subducting slabs associated with high velocities can be traced down very deeply. The transition zone (410–660 km) is still one of the less well-resolved depth ranges. It is characterized by weak heterogeneities but the velocity distribution is dominated by a simple scheme, with high velocities around the Pacific Ocean and slow velocities below central Pacific and Africa, the so-called degree 2 pattern (Masters et al., 1982).

At the global scale, it is usual to expand the velocity anomalies in terms of spherical harmonics $Y_l^m(\theta,\phi)$, where θ,ϕ are the spherical geographical coordinates, l the angular degree, and m the azimuthal order. The spectral amplitude of individual degree l enables the assessment of the importance of different spatial wavelengths λ, since λ is inversely proportional to l. The "degree 2 pattern" predominant in the transition zone as well as in the D″-layer is associated with large-scale mantle flows, upwellings on both sides of Pacific, and upwellings below central Pacific and Africa. The spectrum of heterogeneities is "red" (see Dziewonski and Romanowicz, 2007 for a review), therefore dominated by very large wavelengths but the role of smaller scale objects such as slabs and mantle plumes is still discussed.

The behavior of slabs in regional studies does not seem to be uniform (Figure 4a). Some slabs seem to fall down to the core–mantle boundary (van der Hilst et al., 1997), whereas others, e.g., beneath Japan (Fukao et al., 2001) are stagnant in the transition zone. Mantle plumes and the associated hotspots might initiate continental breakup, inducing the opening of new ridges of oceans, and might dramatically affect biodiversity. After plate tectonics, will we see the development of plume tectonics? There is still a controversy on the origin of plumes. It is likely that they do not have a single origin, but can originate from the different boundary layers within the Earth, in the asthenosphere, transition zone, and D″-layer. There are still lively debates about the origin at depth of mantle plumes (Figure 4b shows an example of plume imaging), and two opposite extreme models (plate model and plume model) are summarized in Figure 5. At the base of the mantle, the D″-layer is still very mysterious since it is



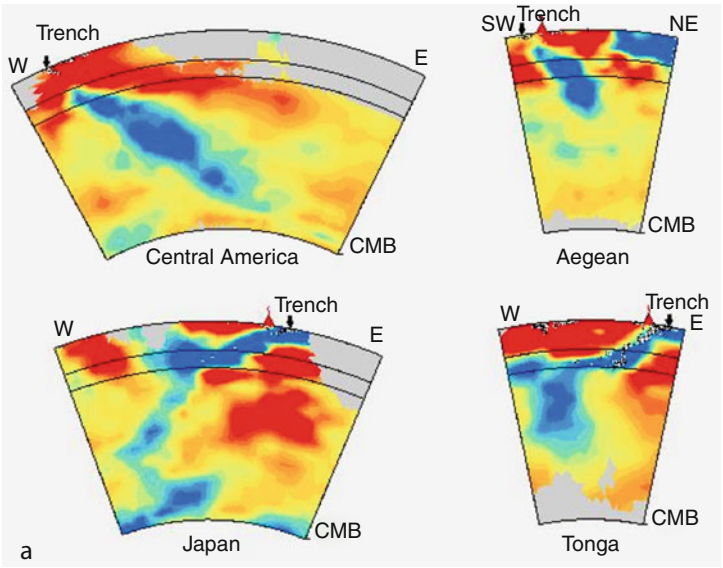
Earth's Structure, Global, Figure 3 Maps through S-velocity models SAW24B16, SB4L18, TXBW, and S20RTS and P-velocity models P-MIT and FF-PRI at depths of 125, 600, 1,350, and 2,750 km (From Thurber and Ritsema, 2007).

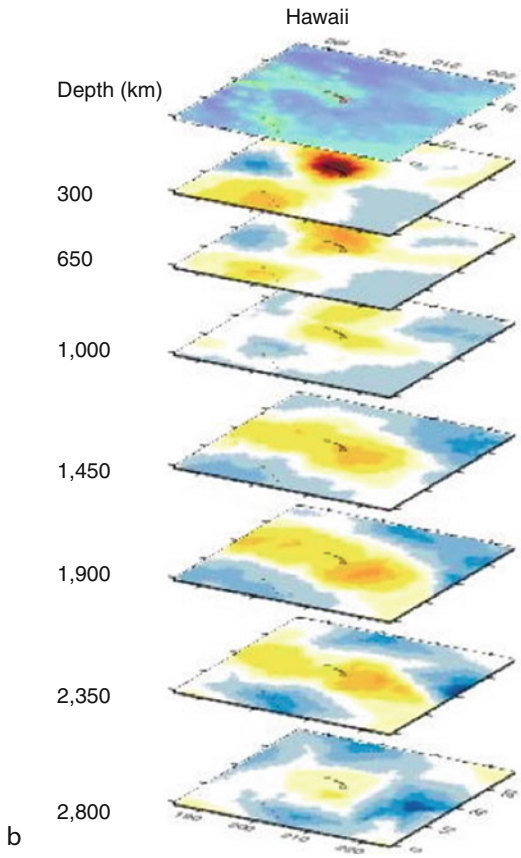
the probable graveyard of slabs and a good candidate for being the matrix of some mantle plumes giving birth to hotspots at the surface of the Earth. The recent and unexpected discovery of post-perovskite (Murakami et al., 2004), which might be present in the lowermost mantle has implications for the D″-layer that influences the convective mixing in the mantle responsible for plate tectonics.

An anisotropic and anelastic Earth

Most earth minerals cannot be limited to their isotropic and elastic properties. When they are deformed over geological timescales, mineral grains and aggregates tend to align by lattice-preferred orientation (LPO), which can lead to bulk anisotropy (see Becker, *Seismic Anisotropy*). Additional anisotropy can also come up when crack distribution (in the crust) or fluid inclusions (asthenosphere, inner core) are present within the material (shape-preferred orientation, SPO). Water content might also play a key role in alignment mechanisms (Karato et al., 2008). In the upper mantle, the different minerals present are strongly anisotropic. The difference of velocity between the fast axis and the slow axis is larger than 20% for olivine, the main constituent of the upper mantle. Other important constituents such as orthopyroxene or clinopyroxene are anisotropic as well (>10%) (see Babuska and Cara (1991), Anderson, (2007), Mainprice, (2007) for reviews). Consequently, the petrological models that are assemblages of different minerals are less anisotropic than pure olivine. The amount of anisotropy is largely dependent on the percentage of these different minerals and on the mechanisms aligning the crystallographic axes according to LPO. For example, the anisotropy of the pyrolitic model, mainly composed of olivine and orthopyroxene (Ringwood, 1975), will depend on the relative orientation of the crystallographic axes of different constituents (Christensen and Lundquist, 1982).

For sake of mathematical simplicity, it was usually assumed that the propagating elastic medium is isotropic, in spite of the large amplitude of anisotropy in several depth ranges. And only 1D reference models of attenuation are incorporated in seismic modeling. Global tomographic models improved over years not only by an increase in the number of data but also by more general parameterizations. They are now including anisotropy (radial anisotropy in Nataf et al. (1984); general slight anisotropy in Montagner and Tanimoto (1991), and anelasticity (Romanowicz, 1995). Whereas isotropic heterogeneities enable the mapping of hot and cold regions within the Earth, seismic anisotropy gives access to the convective flow. Gaboret et al. (2003) and Becker et al. (2003, 2008) demonstrated that mantle circulation inferred from geodynamic models is in good agreement with radial and azimuthal anisotropy distributions (Figure 6). Seismic anisotropy has many other applications such as the determination of the base of continental roots (Montagner, 1994; Gaherty and Jordan, 1995; Silver, 1996; Gung et al., 2003; Debayle et al., 2005), the

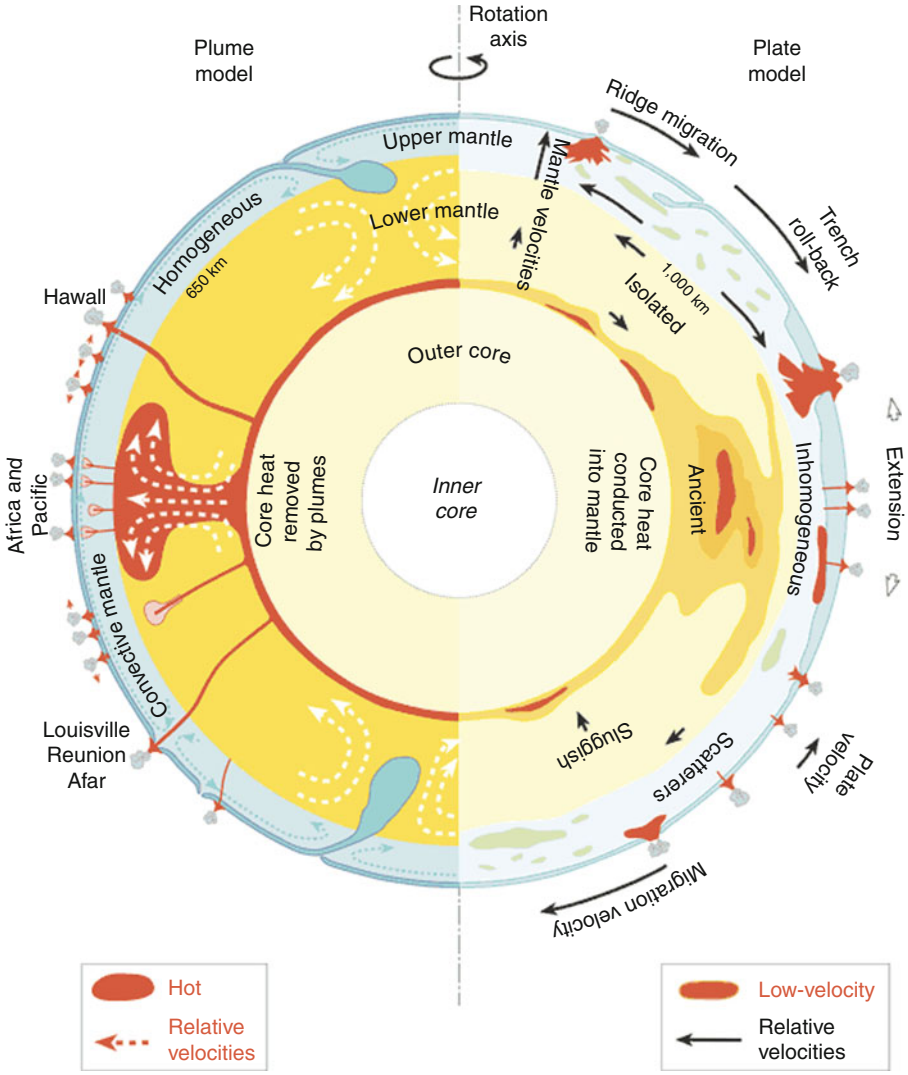




Earth's Structure, Global, Figure 4 Tomographic images of slabs (**a**, van der Hilst et al., 1997) and plume (**b**, Montelli et al., 2004).

investigation of small-scale convection at the base of the lithosphere, of convection boundary layers within the Earth, in the transition zone (Trampert and van Heijst, 2002), D″-layer (see Lay, and Garnero et al.,) and even within the inner core (see Souriau (2007), Wen in structure of the inner core).

Because of its strong dependence on temperature, partial melting, and water content, mapping anelastic



Earth's Structure, Global, Figure 5 Competing models for explaining the origin of hotspots (Left, Courtillot et al., 2003; Anderson, 2001) (From Foulger et al., 2005).

attenuation in the Earth has the potential to provide valuable information on Earth's three-dimensional (3D) structure and dynamics, in complement to what can be learned from mapping elastic isotropic velocities and anisotropic parameters. A significant challenge is to separate the effects of anelastic (or intrinsic) attenuation from those of elastic scattering and focusing due to propagation in elastic 3D structure.

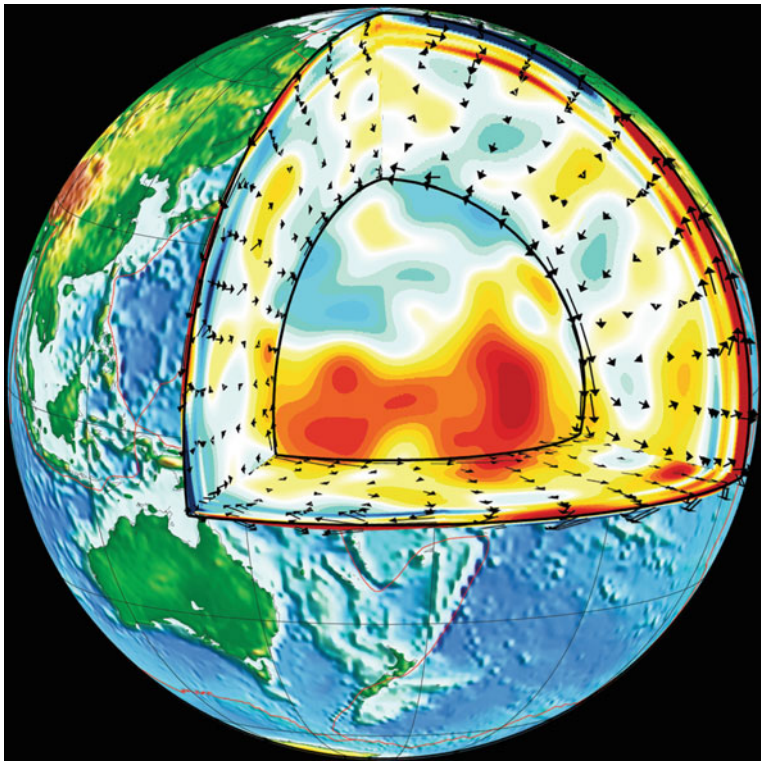
Seismic anisotropy and anelasticity provide new ways to investigate geodynamic processes within the different layers of the Earth.

Conclusions and new challenges

So far, global 3D tomographic models are well resolved for very large spatial scales (>1,000 km). And the global flow circulation in most of the mantle is dominated by the "degree 2 pattern." However, smaller wavelengths are not so well resolved even though many geological objects with a lateral extent of the order of 100 km, such as slabs or mantle plumes play a key role in global geodynamics.

New innovative numerical techniques for forward modeling (such as Spectral Element Method) and for inverse problems (adjoint tomography of Tarantola, 1986; Tromp et al., 2005) are now available and make it possible to use the full waveforms of seismograms. Figure 4b shows an image of the Hawaiian plume (Montelli et al., 2004) obtained by incorporating finite frequency effects. Therefore, we can expect that they should lead to a new revolution in global tomography.

Impressive progress through global tomography, geochemistry, and mineral physics has been made during the last 20 years in our understanding of global geodynamics,



Earth's Structure, Global, Figure 6 Numerical modeling of flow circulation superimposed to 3D tomographic model (Gaboret et al., 2003).

demonstrating how active and turbulent is our planet. All layers are heterogeneous, interact, and exchange matter. The geosciences community must improve the lateral resolution and the quality of the 3D images and incorporate on a routine basis, anisotropy, and anelasticity. There is a real need for a 3D seismic reference Earth model in agreement with geological, mineralogical data, and fluid dynamics modeling.

This progress brings about the expansion of new instruments such as very dense seismic networks on land (e.g., US-array, Hi-Net, Euro-Array, . . .), on the sea floor, by exploring other planets, and by incorporating new data such as seismic noise (Shapiro et al., 2005; Nishida et al., 2009) implementing more powerful numerical techniques. Geophysics only provides some pieces for our puzzling global geodynamics and a multidisciplinary effort is necessary to fully understand the spatiotemporal evolution of our planet. The exploration of other telluric planets will provide other examples of solutions chosen by nature.

Bibliography

Aki, A., and Richards, P. G., 2002. *Quantitative Seismology*, 2nd edn. Sausalito: University Science Books.

Anderson, D. L., 2007. *New Theory of the Earth*. New York: Cambridge University Press.

Babuska, V., and Cara, M., 1991. *Seismic Anisotropy in the Earth*. Dordrecht: Kluwer Academic.

Dahlen, F. A., and Tromp, J., 1998. *Theoretical Global Seismology*. Princeton: Princeton University Press.

Gutenberg, B., and Richter, C. F., 1954. *Seismicity of the Earth*. Princeton: Princeton University Press.

Lay, T., and Wallace, T. C., 1995. *Modern Global Seismology*. San Diego: Academic.

Nolet, G., 2008. *A Breviary of Seismic Tomography*. Cambridge: Cambridge University Press.

Poirier, J.-P., 1991. *Introduction to the Physics of the Earth Interior*. Cambridge: Cambridge University Press.

Richter, C. F., 1958. *Elementary Seismology*. San Francisco: Freeman.

Schubert, G. (Editor-in-Chief), 2007. *Treatise on Geophysics*. Amsterdam: Elsevier.

Turcotte, D. L., and Schubert, G., 2002. *Geodynamics*, 2nd edn. New York: Cambridge University Press.

Other references

Anderson, D. L., 2001. Top-down tectonics? *Science*, **293**, 2016–2018.

Becker, T. W., Kellogg, J. B., Ekström, G., and O'Connell, R. J., 2003. Comparison of azimuthal seismic anisotropy from surface waves and finite-strain from global mantle-circulation models. *Geophysical Journal International*, **155**, 696–714.

Bercovici, D., and Karato, S.-I., 2003. Whole-mantle convection and the transition-zone water filter. *Nature*, **425**, 39–44.

Birch, F., 1952. Elasticity and constitution of the Earth's interior. *Journal of Geophysical Research*, **57**, 227–286.

Christensen, N. I., and Lundquist, S., 1982. Pyroxene orientation within the upper mantle. *Bulletin of the Geological Society of America*, **93**, 279–288.

Courtillot, V., Davaille, A., Besse, J., and Stock, J., 2003. Three distinct types of hotspots in the Earth's mantle. *Earth and Planetary Science Letters*, **205**, 295–308.
Debayle, E., Kennett, B. L. N., and Priestley, K., 2005. Global azimuthal anisotropy and the unique plate-motion déformation of Australia. *Nature*, **433**, 509–512.
Dziewonski, A. M., 1984. Mapping the lower mantle: determination of lateral heterogeneity in P velocity up to degree and order 6. *Journal of Geophysical Research*, **89**, 5929–5952.
Dziewonski, A. M., and Anderson, D. L., 1981. Preliminary reference earth model. *Physics of the Earth and Planetary Interiors*, **25**, 297–356.
Dziewonski, A. M., and Romanowicz, B., 2007. Overview. In Romanowicz, B., and Dziewonski, A. (eds.), *Treatise on Geophysics*. Amsterdam: Elsevier, Vol. 1, pp. 1–29.
Dziewonski, A. M., Hager, B. H., and O'Connell, R., 1977. Large-scale heterogeneities in the lower mantle. *Journal of Geophysical Research*, **82**, 239–255.
Foulger, G. R., Natland, J. H., Presnall, D. C., and Anderson, D. L. (eds.), 2005. Plates, Plumes, and Paradigms. Geological Society of America Special, Vol. 388, 881 pp.
Fukao, Y., Widiyantoro, S., and Obayashi, M., 2001. Stagnant slabs in the upper and lower mantle transition region. *Reviews of Geophysics*, **39**, 291–323.
Gaboret, C., Forte, A. M., and Montagner, J.-P., 2003. The unique dynamics of the Pacific Hemisphere mantle and its signature on seismic anisotropy. *Earth and Planetary Science Letters*, **208**, 219–233.
Gaherty, J. B., and Jordan, T. H., 1995. Lehmann discontinuity as the base of an anisotropic layer beneath continents. *Science*, **268**, 1468–1471.
Gung, Y., Panning, M., and Romanowicz, B., 2003. Global anisotropy and the thickness of continents. *Nature*, **422**, 707–711.
Holmes, A., 1928. Radioactivity and Earth movements. *Transactions of the Geological Society of Glasgow*, **18**, 559–606.
Karato, S.-I., Jung, H., Katayama, I., and Skemer, P., 2008. Geodynamic significance of seismic anisotropy of the upper mantle: new insights from laboratory studies. *Annual Review of Earth and Planetary Sciences*, **36**, 59–95.
Kennett, B. L. N., and Engdahl, E. R., 1991. Traveltimes for global earthquake location and phase identification. *Geophysical Journal International*, **122**, 429–465.
Kennett, B. L. N., Engdahl, E. R., and Buland, R., 1995. Constraints on seismic velocities in the Earth from traveltimes. *Geophysical Journal International*, **122**, 108–124.
Lehmann, I., 1936. Publications du Bureau Central Séismologique International, série A. *Travaux Scientifiques*, **14**, 87–115.
Mainprice, D., 2007. Seismic anisotropy of the deep Earth from a mineral and rock physics perspective. In Schubert, G., and Bercovici, D. (eds.), *Treatise on Geophysics*. Amsterdam: Elsevier, Vol. 2, pp. 437–492.
Masters, G., Jordan, T. H., Silver, P. G., and Gilbert, F., 1982. Aspherical Earth structure from fundamental spheroidal mode data. *Nature*, **298**, 609–613.
Montagner, J.-P., 1994. What can seismology tell us about mantle convection? *Reviews of Geophysics*, **32**, 115–137.
Montagner, J.-P., and Kennett, B. L. N., 1996. How to reconcile body wave and normal-mode reference Earth models? *Geophysical Journal International*, **125**, 229–248.
Montagner, J.-P., and Tanimoto, T., 1991. Global upper mantle tomography of seismic velocities and anisotropies. *Journal of Geophysical Research*, **96**, 20337–20351.
Montelli, R., Nolet, G., Dahlen, F., Masters, G., Engdahl, E., and Hung, S., 2004. Finite-frequency tomography reveals a variety of plumes in the mantle. *Science*, **303**, 338–343.
Morgan, W. J., 1971. Convection plumes in the lower mantle. *Nature*, **230**, 42–43.
Murakami, M., Hirose, K., Kawamura, K., Sata, Na, and Ohishi, Y., 2004. Post-perovskite phase transition in $MgSiO_3$. *Science*, **304**, 5672.
Nataf, H.-C., Nakanishi, I., and Anderson, D. L., 1984. Anisotropy and shear velocity heterogeneity in the upper mantle. *Geophysical Research Letters*, **11**, 109–112.
Nishida, K., Montagner, J.-P., and Kawakatsu, H., 2009. Global surface wave tomography using seismic hum. *Science*, **326**(5949), 112, doi:10.1126/science1176389.
Oldham, R. D., 1906. The constitution of the Earth. *Quarterly Journal Geological Society of London*, **62**, 456–475.
Pekeris, C., 1935. Thermal convection in the interior of the Earth. *Geophysical Journal*, **3**, 343–367.
Ringwood, A. E., 1975. *Composition and Petrology of the Earth's Mantle*. New York: McGraw–Hill, 618 pp.
Romanowicz, B., 1995. A global tomographic model of shear attenuation in the upper mantle. *Journal of Geophysical Research*, **100**, 12375–12394.
Romanowicz, B., and Mitchell, B. J., 2007. Deep earth structure – Q of the Earth from crust to core. In Romanowicz, B., and Dziewonski, A. (eds.), *Treatise on Geophysics*. Amsterdam: Elsevier, Vol. 1, pp. 732–774.
Sengupta, M. K., and Toksoz, M. N., 1976. Three dimensional model of seismic velocity variation in the Earth's mantle. *Geophysical Research Letters*, **3**, 84–86.
Shapiro, N. M., Campillo, M., Stehly, L., and Ritzwoller, M. H., 2005. High-resolution surface-wave tomography from ambient seismic noise. *Science*, **307**, 1615–1618.
Silver, P. G., 1996. Seismic anisotropy beneath the continents: probing the depths of geology. *Annual Review of Earth and Planetary Sciences*, **24**, 385–432.
Souriau, A., 2007. The Earth's core. In Romanowicz, B., and Dziewonski, A. (eds.), *Treatise on Geophysics*. Amsterdam: Elsevier, Vol. 1, pp. 655–693.
Tanimoto, T., and Anderson, D. L., 1985. Lateral heterogeneity and azimuthal anisotropy of the upper mantle: Love and Rayleigh waves. 100–250 s. *Journal of Geophysical Research*, **90**, 1842–1858.
Tarantola, A., 1986. A strategy for nonlinear elastic inversion of seismic reflection data. *Geophysics*, **51**, 1893–1903.
Thurber, C., and Ritsema, J., 2007. Theory and observations – seismic tomography and inverse methods. In Romanowicz, B., and Dziewonski, A. (eds.), *Treatise on Geophysics*. Amsterdam: Elsevier, Vol. 1, pp. 323–360.
Trampert, J., and van Heijst, H. J., 2002. Global azimuthal anisotropy in the transition zone. *Science*, **296**, 1297–1299.
Tromp, J., Tape, C., and Liu, Q., 2005. Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels. *Geophysical Journal International*, **160**, 195–216.
van der Hilst, R. D., Widiyantoro, S., and Engdahl, R., 1997. Evidence of deep mantle circulation from global tomography. *Nature*, **386**, 578–584.
Vinnik, L., Kosarev, G. L., and Makeyeva, L. I., 1984. Anisotropy of the lithosphere from the observations of *SKS* and *SKKS* phases. *Proceedings of the Academy of Sciences of USSR*, **278**, 1335–1339.
Wilson, J. T., 1963. Evidence from Islands on the spreading of ocean. *Nature*, **197**, 536–538.
Woodhouse, J. H., and Dziewonski, A. M., 1984. Mapping the upper mantle; three-dimensional modeling of Earth structure by inversion of seismic waveforms. *Journal of Geophysical Research*, **89**, 5953–5986.

Cross-references

Body Waves
Core-Mantle Coupling

Differential Rotation of the Earth's Inner Core
Earth's Structure, Core
Earth's Structure, Lower Mantle
Earth's Structure, Upper Mantle
Free Oscillations of the Earth
Geodynamics
Inverse Theory, Linear
Lithosphere, Continental
Mantle Convection
Mantle D″ Layer
Mantle Plumes
Plate Driving Forces
Propagation of Elastic Waves: Fundamentals
Seismic Anisotropy
Seismic Discontinuities in the Transition Zone
Seismic Imaging, Overview
Seismic Instrumentation
Seismic Structure at Mid-Ocean Ridges
Seismic Tomography
Seismic, Migration
Seismic, Ray Theory
Seismic Wave Propagation in Real Media: Numerical Modeling Approaches
Subduction Zones
Surface Waves

EARTH'S STRUCTURE, LOWER MANTLE

Edward J. Garnero, Allen K. McNamara, James A. Tyburczy
School of Earth and Space Exploration, Arizona State University, Tempe, AZ, USA

Synonyms

Deep mantle; Deep Earth

Definition

Seismic tomography. An inverse method that utilizes seismic wave travel times and/or waveforms to estimate Earth's seismic velocity and density structure, either regionally or globally.

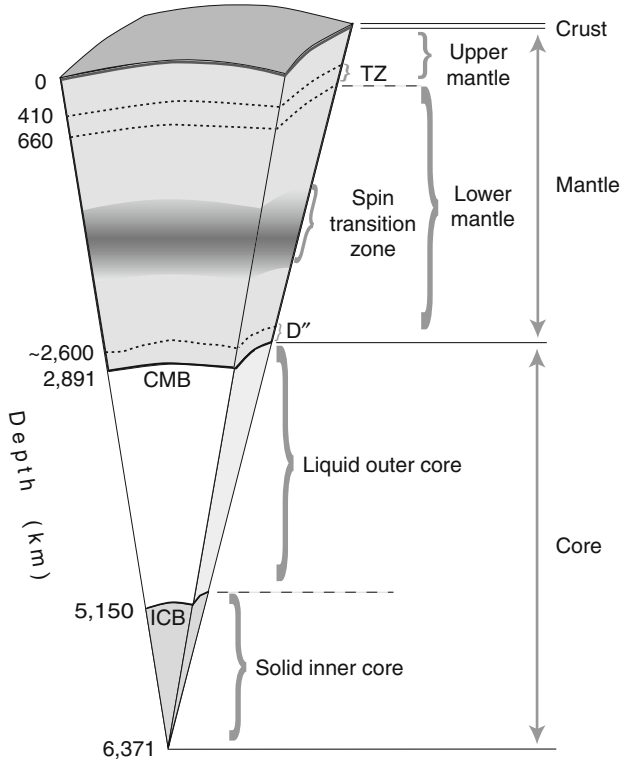
Forward modeling. The procedure involving the generation of models that predicts observations, which is frequently used in deep Earth seismology.

Phase transition. A crystallographic rearrangement of atoms in a mineral that occurs at high pressure.

D″. The depth shell of Earth occupying the lowermost 200–300 km of Earth's lower mantle. (Pronounced "D-double-prime").

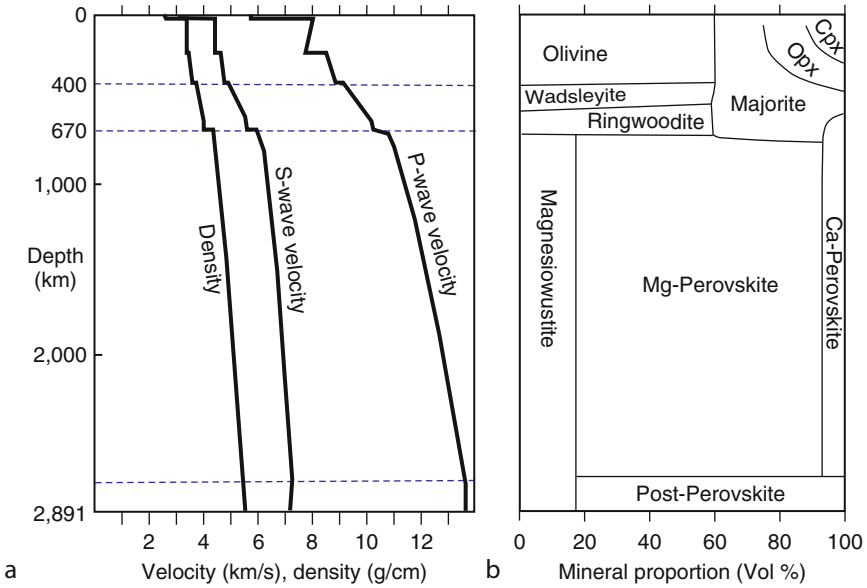
Introduction

Earth's interior has three fundamentally distinct divisions: (1) the relatively buoyant and thin outer crust, which averages roughly 6–7 km in thickness beneath oceans and 30–40 km in continents; (2) the mantle, which extends to nearly half way to Earth's center; and (3) the dominantly iron (Fe) core, which has a fluid outer part and a solid inner part. Figure 1 highlights these basic shells,



Earth's Structure, Lower Mantle, Figure 1 Major depth shells of Earth, along with a number of subdivisions. TZ = transition zone (the upper mantle between 410 and 660 km depth); CMB = core–mantle boundary; ICB = Inner core boundary.

along with a number of additional significant mantle subdivisions. The mantle is divided into an upper and lower part, with the boundary near 660 km depth. The upper mantle has another distinct discontinuity near 410 km depth. As shown in Figure 2, the dominant mineral in the upper mantle is olivine, a magnesium-iron silicate mineral with composition $(Mg,Fe)_2[SiO_4]$, that undergoes phase transitions to higher density mineral forms (or polymorphs) near 410, 520, and 660 km depths. The region between 410 and 660 km is referred to as the mantle transition zone. Below 660 km, mantle silicates undergo another phase transformation, compacting into the perovskite structure, with composition $(Mg,Fe)[SiO_3]$, plus magnesium-iron oxides, (Mg,Fe)O (known as ferropericlase and also referred to as magnesiowüstite in earlier literature). In the lowermost several hundred km of the mantle, a region referred to as D″, perovskite has been shown to undergo an additional phase transition into the "post-perovskite" structure (Murakami et al., 2004 [Science]), and depending on the temperature, post-perovskite might back-transform into perovskite at greater depth, remaining as perovskite down to the CMB (see *Mantle D″ Layer*). The lower mantle depth shell accounts for nearly 56% of Earth's volume, and extends from 660 km depth down to the core–mantle boundary (CMB) at 2,891 km depth (and thus includes the D″ layer).



Earth's Structure, Lower Mantle, Figure 2 (**a**) Variations in seismic properties of the Earth's mantle with depth, from the surface to the core–mantle boundary, showing the major discontinuities in P-wave velocity, S-wave velocity, and density. (**b**) Mineralogical makeup of the mantle. Cpx = clinopyroxene; Opx = orthopyroxene.

The major discontinuities throughout Earth's mantle are thus well-explained by phase transitions in the olivine-perovskite system.

While uncertainties are present in our understanding of the exact compositional make-up of Earth's mantle, it is generally believed that magnesium-iron silicate perovskite is the dominant mineral structure of the lower mantle, possibly accounting for ~80% of the lower mantle, and hence is the most abundant mineral on Earth. This perovskite mineral structure is the same as that occurring in well known electronic and piezoelectric materials, such as barium titanate. A fundamental difference between the silicate minerals of the upper mantle and lower mantle perovskite is that every silicon (Si) atom in perovskite is surrounded by six oxygen (O) atoms (silicon is octahedrally coordinated), whereas for upper mantle silicates the silicon is bonded to four O atoms (tetrahedral coordination). A primary effect of the higher pressure in the lower mantle is that O atoms occupy a smaller volume, thus more can be organized around (coordinated with) Si. This more compact rearrangement of atoms in lower mantle phases means that they have higher density than upper mantle materials, as well as differences in other properties such as the material's stiffness and resistance to shearing. These changes affect the velocities of seismic waves; thus, at depths where phase transitions occur, there are jumps (i.e., discontinuous increases) in seismic compressional and shear wave speeds (referred to as P-wave and S-wave velocities, or V_P and V_S, respectively). Velocity and density depth distributions in Earth's mantle are presented in Figure 2; these values represent an estimation of the globally averaged properties.

The exact depth of a phase transition in any given mineral group depends upon temperature and pressure (and chemical composition). Earlier conceptions of the lower mantle (Birch, 1952) postulated that the lower mantle had a constant composition, and hence the temperature and velocity variations could by described by simple adiabatic compression. However, in light of more recent studies, it is interesting to consider the fact that lateral temperature variations are likely in the lower mantle, because the mantle is convecting, and hence phase boundaries are not expected to be perfectly flat (e.g., Lebedev et al., 2002). Rather, they should have relief that directly relates to the temperature field (or to perturbations in chemical composition, since minerals may transform at different pressures if there are variations or impurities in the mineral composition). The Earth presumably was significantly hotter in the distant geologic past (from heat generated during formation and differentiation of the mantle and core), therefore we expect phase boundaries in the earliest Earth to have been at different depths than today.

Tools for studying the lower mantle

Earth's mass and moment of inertia, determined from orbital and rotational dynamics, provide important information about the planet's internal density distribution. These data, combined with assumptions about starting materials for Earth's formation from analyses of meteorites, have long indicated a predominantly iron core with a silicate rock mantle (Carlson, 2003). Seismic studies early in the twentieth century confirmed the existence of the core (outer and inner), as well as the fact that the outer core is liquid (from the absence of shear waves on the opposite side of the planet following an earthquake, because shear waves cannot propagate in fluids).

A number of Earth science disciplines contribute to our body of knowledge of the lower mantle (and the deep Earth in general). For example, the field of mineral physics includes high pressure laboratory experiments aimed at simulating the pressures and temperatures within the deep Earth. Predictions of the melting temperature of iron at the conditions of the boundary between the inner and outer core (at radius of ~1,220 km), as well as measurements delineating the temperature of the olivine to perovskite plus ferropericlase phase transition at 660 km depth provide two important temperature tie-in points that help to constrain Earth's geotherm; temperature must be extrapolated from these points toward the CMB, where a very large temperature jump occurs. Based on such studies, the top of the lower mantle at 660 km depth is at a pressure of about 23 GPa (1 Pa = 1 N m^{-2} and 10^5 Pa corresponds to approximately 1 atm of pressure) and a temperature of about 1,800 +/− 200 K (e.g., Ito and Katsura, 1989; Frost, 2008). The CMB is at a pressure of about 130 GPa; the temperature, while less constrained, has been recently estimated to be about 4,000 +/− 500 K (Van der Hilst et al., 2007). The temperature jump from the CMB to the outermost core may be as much as 1,000–1,500 K.

A number of notable uncertainties are present that currently preclude precise knowledge of lower mantle temperature, including the amount of heat-producing radiogenic materials, the vigor of mantle convection, the heat flux from the fluid core into the mantle, and important thermal properties of perovskite at lower mantle pressure and temperature conditions (e.g., thermal conductivity, heat capacity, and the coefficient of thermal expansion). However, using a diamond anvil cell high-pressure device to achieve pressures as great as those of the lower mantle and even the core, researchers have recently refined our understanding about two important characteristics of perovskite. As mentioned above, perovskite (*pv*) can transform into a higher pressure (and density) structure termed "post-perovskite" (*ppv*) at D″ pressures (around 120 GPa). Thus, if the systematics of this phase transition can be well defined from laboratory experiments or theory, then seismic mapping of the $pv \rightarrow ppv$ transition depth (as well as that of possible $ppv \rightarrow pv$ back-transformation) adds an additional temperature tie-in point for the geotherm (Hernlund et al., 2005). Another discovery describes a change with increased pressure in the electronic spin state of iron in magnesiowüstite and perovskite (a change from a high-spin configuration with a maximum number of unpaired d-shell electrons to a low-spin configuration in which d-shell electron pairing is maximized). Experiments indicate that this will occur between the depths of 1,000 and 2,200 km (Lin et al., 2007). Current research predicts that this spin transition may result in a softer material, which can be more easily compressed thus increasing the density of the deeper mantle, possibly affecting convective dynamics.

The vigor of mantle convection plays a central role in the cooling of Earth. The most commonly used parameter to determine if convection is occurring and to describe the strength of convection is the Rayleigh number, *Ra*:

$$Ra = \frac{\alpha g \rho \Delta T D^3}{\eta \kappa},$$

where α is the coefficient of thermal expansion, g is the acceleration due to gravity, ρ is density, ΔT is the non-adiabatic temperature change across the convecting layer (in this case, the whole mantle) of thickness D, η is the viscosity, and κ is the thermal diffusivity. *Ra* in the lower mantle is estimated to be on the order of 10^7; values greater than $\sim 10^3$ indicate that convection is occurring and values as high as 10^7 indicate vigorous convection. Factors such as the temperature dependence of viscosity can strongly affect the style of convection. Thus, numerical convection calculations incorporate our best understanding of Earth's properties and provide important information about the dynamics and evolution of the mantle (and the planet as a whole).

Erupted materials offer an opportunity to study mantle chemistry. Specifically, some hot spot volcanoes are thought to be caused by mantle plumes that originate in the lower mantle. There is currently debate regarding whether all hot spots tap the lower mantle, as well as the depth of origin of different plume sources (Courtillot et al., 2003). However, it is widely accepted that magmas from hot spot volcanoes contain isotopes that are different compared to lavas erupted at mid-ocean ridges. This has led to a view that the lower mantle represents a distinct reservoir that has undergone little or no mixing with the upper mantle (Hofmann, 1997). Subsequent and ongoing work continues to challenge that perspective, which we address in the three-dimensional structure and dynamics section below.

Seismic imaging is the primary tool that gives us the ability to map lateral variations of mantle properties from the upper mantle to the CMB. Seismic tomography provides an image of the full three-dimensional heterogeneity field, albeit at relatively long horizontal wavelengths, on the order of 3,000 km or so (e.g., Ishii and Tromp, 1999; Ritsema and van Heijst, 2000; Mégnin and Romanowicz, 2000; Grand, 2002). Variations in seismic velocity as well as density (Trampert et al., 2004; Ishii and Tromp, 2004) have been mapped, and when compared to surface tectonics, help to depict convective flow in the interior.

Information about the small-scale heterogeneity field is provided by a number of forward-modeling seismic studies. For example, some studies have suggested that small-scale scatterers exist throughout the lower mantle (Hedlin and Shearer, 2000). Other studies find evidence for discrete reflectors in some locations, particularly beneath subduction zones (Rost et al., 2008; Kito et al., 2008). The exact nature of such small-scale heterogeneity is unknown, but may indeed relate to an incompletely mixed mantle, or heterogeneities entrained from the top (dense material falling) or the bottom (buoyant material rising).

Three-dimensional structure and dynamics

For decades, it has been known that Earth's mantle contains significant lateral variations in seismic properties, which are due to variations in temperature and or chemical composition compared to the local average properties. Results from seismic tomography depict V_P, V_S, and ρ changes throughout the mantle. Variations are neither expected nor detected in the fluid outer core; the extremely low viscosity and relatively fast convection promote homogenization (Stevenson, 1987). Seismically detected mantle heterogeneity is the strongest in the outermost few hundred kilometers of the mantle (i.e., near the planet's surface), due to the strong variations in temperature and composition associated with plate tectonic and mantle convection processes (Masters et al., 2000). The next most seismically heterogeneous depth shell is the D″ region at the base of the mantle (see *Mantle D′′ Layer*). Figure 3 shows seismic velocity variations in the lower mantle. D″ silicate rock is juxtaposed against the fluid iron outer core, which represents roughly a 75% increase in density. Thus, the mantle side of the CMB can accommodate long-lived stable structures over a wide range of densities (between the mantle and core density); it is possible that cumulates denser than average lowest mantle silicates can survive entrainment by the convective currents in the overlying mantle (e.g., McNamara and Zhong, 2005). However, other possibilities are viable (Deschamps et al., 2007).

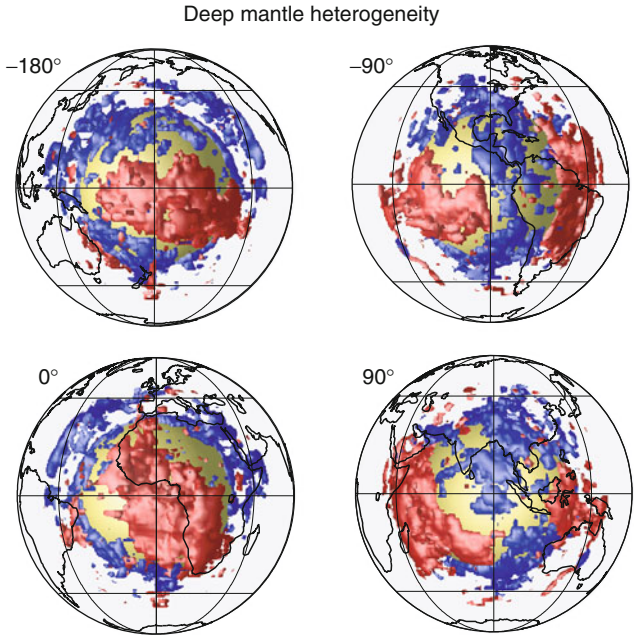
The lower mantle above D″ displays variability in seismic properties, but at a much weaker level than that imaged at the top and bottom of the mantle. In some regions it is plausible that low amplitudes of mid-mantle heterogeneity may reflect our limited ability to confidently image the mantle there (due to a lack of seismic information because earthquakes and seismic recorders are not uniformly distributed on Earth). Nonetheless, all seismic tomography studies to date agree that heterogeneity is greater at the top and bottom of the mantle relative to the middle of the mantle.

Plate tectonics involves the creation of plates at mid-ocean ridges, and destruction or consumption of plates at subduction zones, where the cold and dense oceanic crust and lithosphere (compared to the surrounding mantle) descend into the interior. Tomographic analyses have shown that planar features with high wave speeds beneath subduction zones (e.g., Grand et al., 1997) are consistent with cold material subducting deep into the lower mantle. While the penetration depth of subducting slabs is not adequately constrained beneath all subduction zones, it is widely accepted that most slabs descend through the 660 km discontinuity, in some cases reaching to the CMB (such as beneath the Caribbean). It is important to understand the large-scale long-term flux of material from Earth's surface to great depths in the interior because of the significance for long-term (geological time) effects on surface volatiles, the atmosphere, and global climate.

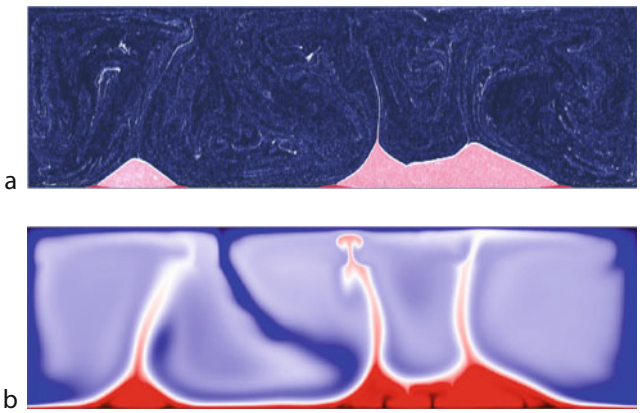
The amount and rate of mass flux between the upper mantle down into the lower mantle depends on several poorly known quantities such as the viscosity of the descending slab and mantle, the viscosity increase at the 660 km discontinuity, as well as the degree of viscous coupling between the oceanic lithosphere and underlying asthenosphere (before and after subduction). Time-integrated mass transfer into (and out of) the lower mantle plays a role in long-term mixing and lower mantle residence time of deep mantle minerals. For example, weak viscous coupling between slabs and the surrounding mantle predicts lower flux rates into (and out of) the lower mantle, and hence less mixing between the upper and lower mantle. A strong coupling between the slab and mantle predicts the opposite. Current data and predictions do not yet distinguish between these possibilities.

An analogous dependency between mixing and convective parameters exists for the bottom of the mantle, where recent work argues for a chemically distinct origin of tomographically identified large, low shear wave velocity regions (Garnero and McNamara, 2008). Convective currents associated with subduction sweep basal dense material to regions away from downwellings, forming thermochemical "piles" (Figure 4) (McNamara and Zhong, 2005). The contrast in properties (e.g., density, viscosity, temperature) between piles and surrounding mantle will determine the morphology and vertical extent of the pile material, as well as the degree of entrainment (and hence its evolutionary behavior). The two, nearly antipodal low shear velocity regions, one beneath the central Pacific and the other beneath the southern Atlantic and Africa (Figure 3), are accompanied (in certain areas) by low-velocity conduits that are associated with mantle plumes (Montelli et al., 2004). Thus, the present-day large-scale structure and circulation pattern of the lower mantle represent a time-integrated effect of anomalous temperature and/or chemistry in Earth's lowermost and uppermost mantle boundary layers (D″ and the lithosphere, respectively).

At smaller scale lengths, seismic studies have found evidence for seismic wave scattering. The imaged size of heterogeneities in the lower mantle ranges from a kilometer to tens of kilometers in scale (e.g., Vidale and Hedlin, 1998; Wen and Helmberger, 1998). This observation indicates chemical anomalies at a spectrum of physical scales and a mantle that is not fully mixed. Indeed, analyses of magmas from hot spot volcanoes have yielded characteristic isotopic signatures distinct from those of mid-ocean ridge basalts, arguing for unique, possibly isolated deep mantle reservoirs that feed mantle plumes. Irregular convective patterns (e.g., see Figure 4) are consistent with such heterogeneities spreading through the lower mantle as well as being entrained into mantle plumes, but not necessarily in a systematic fashion (since convection patterns appear multi-scaled and may be complex). Alternatively, scattering of seismic energy can occur from small pockets of partially molten mantle material. This scenario has been motivated by the detection of thin (a few km) laterally discontinuous layers right at the CMB, called ultra-low-velocity zones (ULVZs). ULVZs have been argued to



Earth's Structure, Lower Mantle, Figure 3 Heterogeneity in the deep mantle at a depth of 2,800 km is seen in the tomographically derived shear wave speed variations (Ritsema and van Heijst, 2000). Iso-surfacing at the + 0.7% (*blue*) and −0.7% (*red*) levels show the higher and lower velocity regions, respectively. Two large low shear velocity provinces are present, and high velocities tend to be located below present-day (or recent past) subduction.



Earth's Structure, Lower Mantle, Figure 4 Numerical geodynamics calculations for a model mantle consisting of three distinct chemistries are shown for the whole mantle, from Earth's surface to the core–mantle boundary. Panel (**a**) displays the chemical field, showing the bulk of the upper and lower mantle in blue, as one chemistry type, dense lower mantle piles (*pink*) as a second distinct chemistry, and thin ultra-low-velocity zones at the core mantle boundary (*red*) that accrue near the pile margins. (**b**) The temperature field associated with (**a**) is shown. Red and blue are hotter and colder regions, respectively. The dense piles can be stable for billions of years, though a low level of entrainment is constant, as evidenced by the white streaks throughout the mantle in panel (McNamara et al., 2010).

contain some degree of partial melt, and hence offer a plausible source of scatters composed of melt pockets (whether or not they are of a distinct chemical composition).

Future progress

As seismic sensors are deployed in new regions across Earth's surface, new geographical regions will be sampled with greater certainty, and thus our understanding of global processes will greatly improve. As sensor deployments become denser (with smaller distances between sensors), Earth's deep structure can be imaged with a much greater resolution. Thus small-scale structures and processes can be inferred with greater certainty and tied to estimations of global processes. The seismic work only maps present-day structure, which must be put in a dynamical and evolutionary framework with the predictions from geodynamical convection calculations, which in turn must first be informed by an understanding of the material properties. The latter comes from the field of mineral physics, both in laboratory experiments at high temperature and pressure, as well as with computer simulations of material behavior at the atomic scale. The time evolution of Earth is also informed by geochemical analyses of erupted materials and meteorites. Thus, future work that advances our understanding of the structure, dynamics, and evolution of Earth's interior will be multidisciplinary in nature.

Summary

Earth's lower mantle represents the largest volume of any depth shell on Earth. It is unique in that its chemistry, structure, dynamics, and evolution represent a time-integrated effect of the chemistry and dynamics of the surface (lithosphere, asthenosphere) and lowermost mantle (D″, ULVZ, CMB) boundary layers. The degree to which the lower mantle is recycled into the upper mantle depends upon many poorly known convective parameters, such as the viscosity and viscous coupling of descending slabs, as well as material properties, such as the nature and origin of dense deep mantle chemically distinct piles. Future work of improved seismic imaging coupled with continued advancements in other deep Earth disciplines, such as geodynamics, mineral physics, and geochemistry, will greatly help to reduce uncertainties in our understanding of Earth's evolutionary pathway, and present-day structure and dynamical state.

Bibliography

Birch, F., 1952. Elasticity and constitution of the Earth's interior. *Journal of Geophysical Research*, **57**(2), 227–286.

Carlson, R. W. (ed.), 2003. *Treatise on geochemistry, Vol. 2: Geochemistry of the Mantle and Core*. Amsterdam: Elsevier, 586 pp.

Courtillot, V., Davaille, A., Baesse, J., and Stock, J., 2003. Three distinct types of hotspots in the Earth's mantle. *Earth and Planetary Science Letters*, **205**, 295–308.

Deschamps, F., Trampert, J., and Tackley, P. J., 2007. Thermochemical structure of the lower mantle: seismological evidence and consequences for geodynamics. In: Yuen, D. A., et al. (ed.), *Superplume: Beyond Plate Tectonics*. Dordrecht: Springer, pp. 293–320.

Frost, D. J., 2008. The upper mantle and transition zone. *Elements*, **4**, 171–176.
Garnero, E. J., and McNamara, A. K., 2008. Structure and dynamics of Earth's lower mantle. *Science*, **320**, 626–628.
Grand, S. P., 2002. Mantle shear-wave tomography and the fate of subducted slabs. *Philosophical Transactions of the Royal Society of London A*, **360**, 2475–2491.
Grand, S. P., van der Hilst, R. D., and Widiyantoro, S., 1997. Global seismic tomography: a snapshot of convection in the Earth. *GSA Today*, **7**, 1–7.
Hedlin, M. A. H., and Shearer, P. M., 2000. An analysis of large scale variations in small-scale mantle heterogeneity using Global Seismic Network recordings of precursors to PKP. *Journal of Geophysical Research*, **105**, 13,655–13,673.
Hernlund, J. W., Thomas, C., and Tackley, P. J., 2005. A doubling of the post-perovskite phase boundary and structure of the Earth's lowermost mantle. *Nature*, **434**, 882–886.
Hofmann, A. W., 1997. Mantle geochemistry: the message from oceanic volcanism. *Nature*, **385**, 219–229.
Ishii, M., and Tromp, J., 1999. Normal-mode and free-air gravity constraints on lateral variations in velocity and density of Earth's mantle. *Science*, **285**, 1231–1236.
Ishii, M., and Tromp, J., 2004. Constraining large-scale mantle heterogeneity using mantle and inner-core sensitive normal modes. *Physics of the Earth and Planetary Interiors*, **146**, 113–124.
Ito, E., and Katsura, T., 1989. A temperature profile of the mantle transition zone. *Geophysical Research Letters*, **16**, 425–428.
Kito, T., Thomas, C., Rietbrock, A., Garnero, E., Nippress, S., and Heath, A., 2008. Detection of a continuous lower mantle slab beneath Central America from seismic wavefield migration. *Geophysical Journal International*, **174**, 1019–1028.
Lebedev, S., Chevrot, S., and Van der Hilst, R. D., 2002. Seismic evidence for olivine phase changes at the 410-and 660-kilometer discontinuities. *Science*, **296**, 1300–1302.
Lin, J.-F., Vanko, G., Jacobsen, S. D., Iota, V., Struzhkin, V. V., Prakapenka, V. B., Kuznetsov, A., and Yoo, C.-S., 2007. Spin transition zone in Earth's lower mantle. *Science*, **317**, 1740–1743.
Masters, G., Laske, G., Bolton, H., and Dziewonski, A. M., 2000. The relative behavior of shear velocity, bulk sound speed, and compressional velocity in the mantle: implications for chemical and thermal structure. In Karato, S.-I., Forte, A. M., Liebermann, R. C., Masters, G., and Stixrude, L. (eds.), *Earth's Deep Interior: Mineral Physics and Tomography From the Atomic to the Global Scale*. Washington, D.C: AGU, pp. 63–87.
McNamara, A. K., and Zhong, S., 2005. Thermochemical structures beneath Africa and the pacific ocean. *Nature*, **437**, 1136–1139.
McNamara, A. K., Garnero, E. J., and Rost, S., 2010. Tracking deep mantle reservoirs with ultra-low velocity zones. *Earth and Planetary Science Letters*, **299**, 1–9, doi:10.1016/j.epsl.2010.07.42.
Mégnin, C., and Romanowicz, B., 2000. The shear velocity structure of the mantle from the inversion of body, surface, and higher modes waveforms. *Geophysical Journal International*, **143**, 709–728.
Montelli, R., Nolet, G., Dahlen, F., Masters, G., Engdahl, E., and Hung, S., 2004. Finite-frequency tomography reveals a variety of plumes in the mantle. *Science*, **303**, 338–343.
Murakami, M., Hirose, K., Kawamura, K., Sata, Na, and Ohishi, Y., 2004. Post-perovskite phase transition in $MgSiO_3$. *Science*, **304**, 5672.
Ritsema, J., and van Heijst, H. J., 2000. Seismic imaging of structural heterogeneity in Earth's mantle: evidence for large-scale mantle flow. *Science Progress*, **83**, 243–259.
Rost, S., Garnero, E. J., and Williams, Q., 2008. Seismic array detection of subducted oceanic crust in the lower mantle. *Journal of Geophysical Research*, **113**, B06303, doi:10.1029/2007JB005263.
Stevenson, D. J., 1987. Limits on lateral density and velocity variations in the Earth's outer core. *Geophysical Journal of the Royal Astronomical Society*, **88**, 311–319.
Trampert, J., Deschamps, F., Resovsky, J., and Yuen, D. A., 2004. Probabilistic tomography maps chemical heterogeneities throughout the mantle. *Science*, **306**, 853–856.
Van der hilst, R. D., de Hoop, M. V., Wang, P., Shim, S.-H., Ma, P., and Tenorio, L., 2007. Seismostratigraphy and thermal structure of Earth's core-mantle boundary region. *Science*, **315**, 1813–1817.
Vidale, J. E., and Hedlin, M. A. H., 1998. Evidence for partial melt at the core-mantle boundary north of Tonga from the strong scattering of seismic waves. *Nature*, **391**, 682–685.
Wen, L., and Helmberger, D. V., 1998. Ultra-low velocity zones near the core-mantle boundary from broadband PKP precursors. *Science*, **279**, 1701–1703.

Cross-references

Core-Mantle Coupling
Earth's Structure, Core
Earth's Structure, Upper Mantle
Mantle Convection
Mantle D'' Layer
Mantle Plumes
Mantle Viscosity

EARTH'S STRUCTURE, UPPER MANTLE

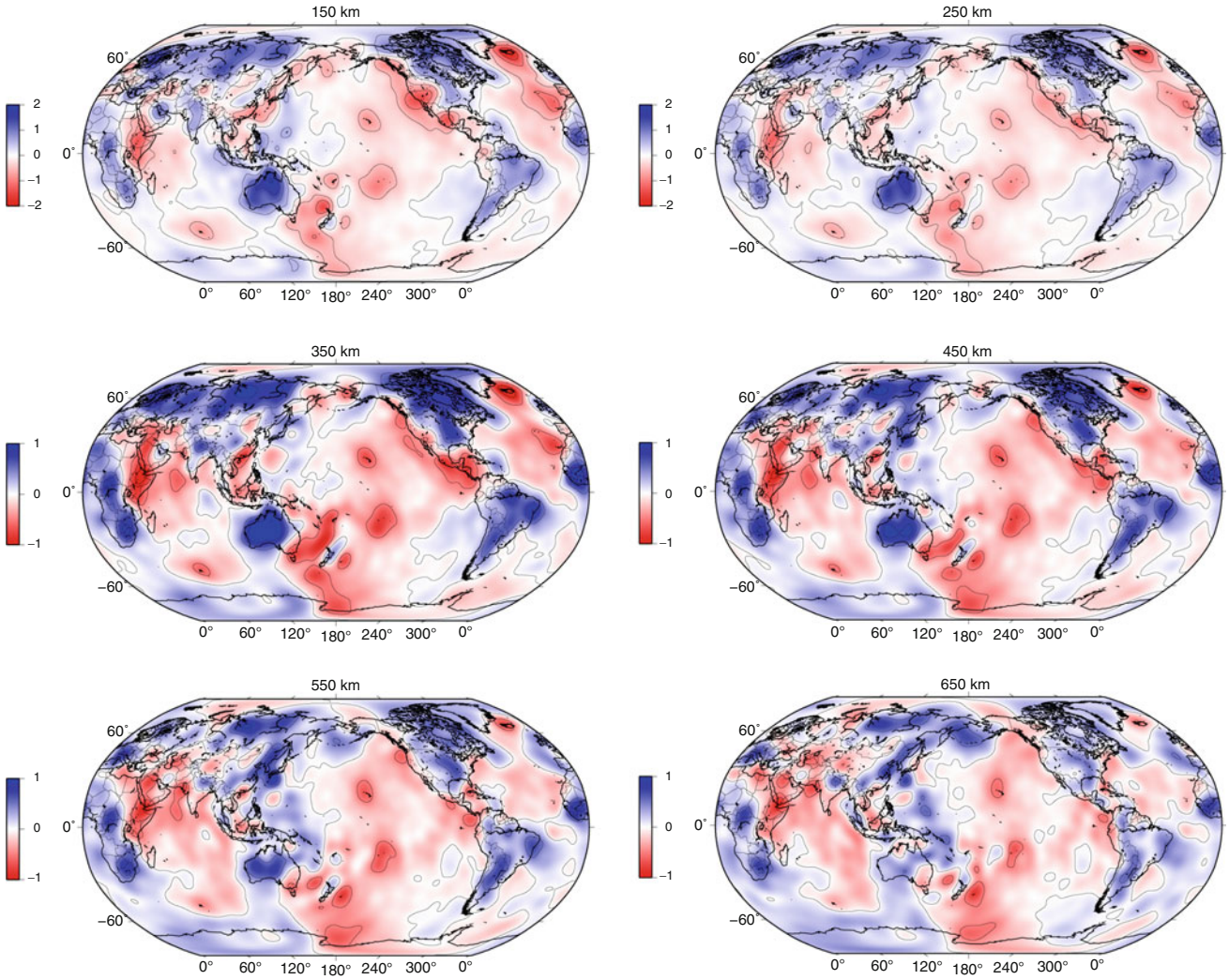
Guust Nolet
Geosciences Azur, Sophia Antipolis, France

Definition

The upper mantle is defined as that part of the mantle between the crust and the phase transition of γ-olivine to perovskite. The total mass of the upper mantle is 1.06×10^{24} kg, about a quarter of the total mass of the mantle. Its volume, 2.95×10^{11} km^3, is a third of the total volume of the mantle.

Until the discovery of a major transition near 660-km depth, the upper mantle was better known as Bullen's "layer B," extending to 400-km depth. A sharp gradient or discontinuity in seismic velocity causes a bend in the travel time curves of P-wave arrivals near 20°, such that the slowness drops from more than 12 s/deg to 10 s/deg. This phenomenon was observed as early as 1931 by Jeffreys, who correctly adopted a suggestion by J.D. Bernal that it represents a high-pressure modification of olivine. The modern notion of the upper mantle actually extends into Bullen's next layer, "layer C," which has now been abandoned as a physically meaningful subdivision.

The depths to the top and the bottom of the upper mantle cannot be given precisely because they depend on the region. The Mohorovičić discontinuity marks the transition from the crust to the upper mantle, with a density jump of the order of 10%. It is found at very shallow depth near ocean ridges where oceanic crust is formed, but extends to more than 70-km depth beneath the Himalayas



Earth's Structure, Upper Mantle, Figure 1 Model pmean, an averaged mantle model for the P velocity from Becker and Boschi (2002), plotted at six different depths in the upper mantle. The contour interval is 0.5%.

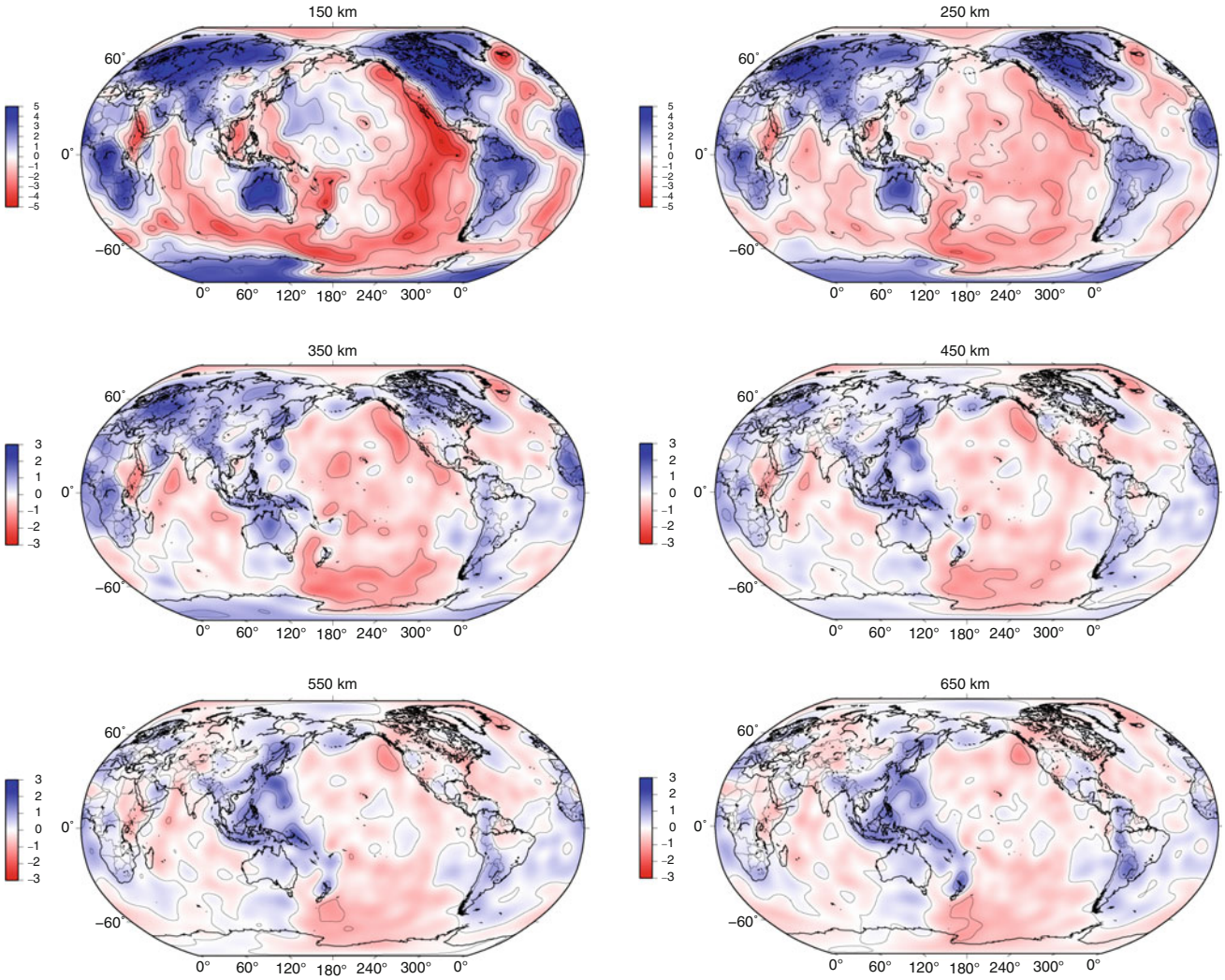
and the Andes. The "transition zone" between upper and lower mantle is marked by the transition of olivine (Mg_2SiO_4) to its polymorph β-olivine (Wadsleyite) near 410-km depth to γ-olivine near 520-km depth and finally to perovskite ($MgSiO_3$), which occurs near a depth of 660 km, but with variations of ±20 km depending on temperature and composition, and possibly with variations more than twice that large for narrow areas inside hot plumes and cold slabs. A typical temperature estimate for the phase transition gives a temperature of about 1,600°C. The upper mantle is one to two orders of magnitude less viscous than the lower mantle, and its electrical conductivity is much lower, with the possible exception of regions of partial melt.

The upper mantle itself is subdivided into several depth regions, characterized by their differences in density, mechanical properties, and/or seismic velocity:

Lithosphere: A strong layer with a viscosity of the order of 10^{21} Pa s and a thickness typically about 80–150 km, though it may exceed 200 km beneath old cratons, representing the rigid "plates." The temperature gradient is superadiabatic. The geotherm in the lithosphere connects the low temperature at the bottom of the crust, 600–900°C under continents but as low as 250°C in the oceans, to the (approximately adiabatic) temperature in the asthenosphere (1,200–1,300°C).

Asthenosphere: A mechanically weak layer with viscosity of the order of 10^{19} Pa s, characterized by a drop in S velocity and possibly also in P velocity. This layer is present under the oceans and the younger part of continents, but its existence beneath the oldest cratons is uncertain. It may contain a small degree of partial melt.

Lehmann discontinuity: A positive velocity jump named after Inge Lehmann who first observed this feature



Earth's Structure, Upper Mantle, Figure 2 Model smean, an averaged mantle model for the S velocity from Becker and Boschi (2002), plotted at six different depths in the upper mantle. The contour interval is 1%.

beneath Europe and North America near 150-km depth. It is generally interpreted as a sharp boundary representing the bottom of the asthenosphere, but not globally observed (Gu et al., 2001).

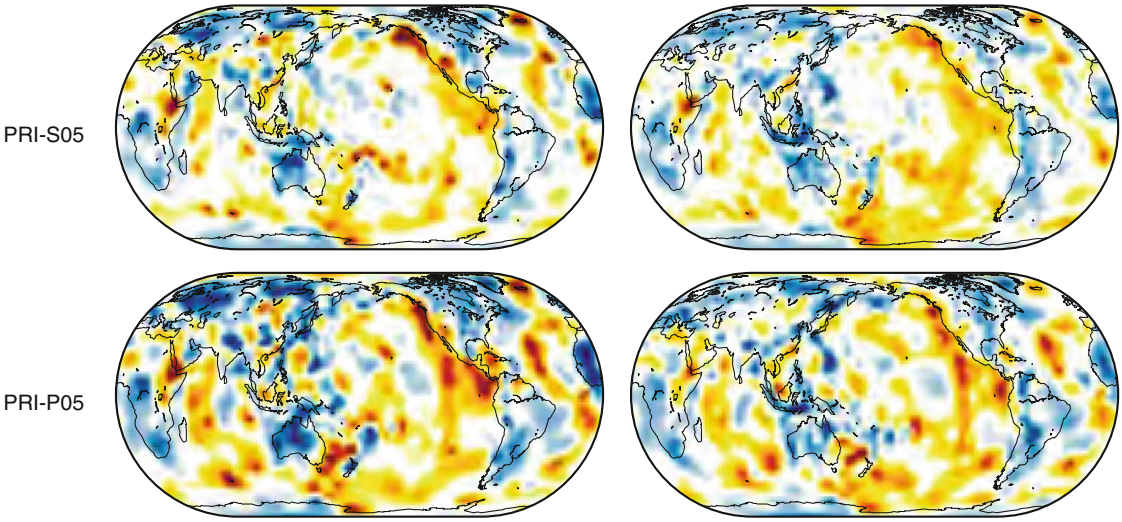
"410-*km*" *discontinuity*: A globally observed jump over a narrow depth interval in seismic velocity, marking the transition from the $\alpha \rightarrow \beta$ phase of olivine near 400-km depth. This transition occurs at a temperature of about 1,400–1,450°C.

"520-*km*" *discontinuity*: Predicted by the phase diagram for olivine, which transforms to the γ phase under certain temperature conditions, this gradient in seismic velocities is occasionally observed.

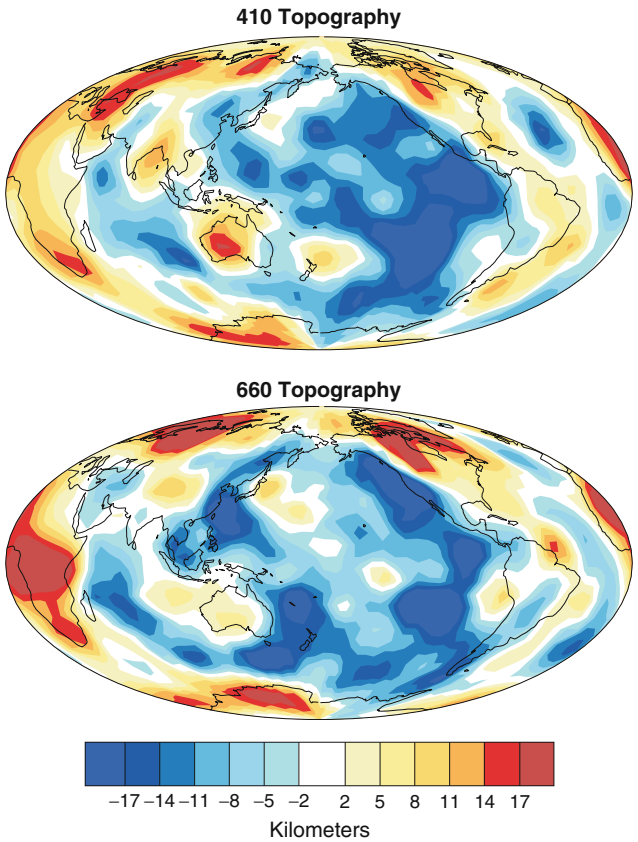
Reflections or refractions from other depths have been observed – such as the "Hales discontinuity" within the lithosphere – and are occasionally attributed to a global physical or chemical transition, but a lack of consistency between such observations makes it more likely that the observations should be explained as wave energy returned from three-dimensional structure rather than from a discontinuity that represents an omnipresent layering.

Early regionalized refraction studies of the upper mantle already showed a marked variability in P and S velocity with depth. Surface wave studies, which are the preferred tool to investigate layers of low velocity, indicate minimum S velocities as low as 4.05 km/s under the western USA, 4.09 km/s under the East-African Rift, and 4.14 km/s under the Pacific ocean, but body waves show a minimum V_s as high as 4.63 km/s under the North American craton.

The Lehmann discontinuity is observed in few regional models, but it is present in the Preliminary Reference Earth Model (PREM) (Dziewonski and Anderson, 1981), which shows a major velocity increase in V_p of



Earth's Structure, Upper Mantle, Figure 3 Models PRI-P05 (P velocity) and PRI-S05 (S-velocity anomaly) at a depth of 350 km (*left*) and 600 km (*right*). The color scale ranges from −1.5% (*red*) to +1.5% (*blue*) in P velocity and ±3% in S velocity. Upper mantle plumes show as narrow quasi-circular slow anomalies in *red* (From Montelli et al., 2006. With permission from the AGU).



Earth's Structure, Upper Mantle, Figure 4 The depth variations to the "410-" and "660-" km discontinuities, with respect to 410 and 650 km depth, respectively (From Houser et al., 2008. With permission from the AGU).

7.1% at 220 km. By way of compensation, PREM has only a minor increase of 2.6% from 8.90 to 9.13 km/s at the "410-km" discontinuity. This, however, disagrees with all regional observations and reveals a major shortcoming of the reference model PREM. In fact, the variability in the velocity jump at 410 km is not very pronounced in the early refraction models. Typically, V_p in such models jumps about 5%. An average over seven regional models gives a V_p jump from 8.78 ± 0.06 to 9.26 ± 0.03 km/s, where the uncertainty represents 1 standard deviation. The jump in V_p at 660 km over the same set of models is roughly 4%, but more variable than that at 410 km: from 10.22 ± 0.20 to 10.72 ± 0.10 km/s. Near 660-km depth model, PREM is in agreement with regional studies.

Though the regional models based on refracted waves provide the most accurate estimates of *absolute* velocities, the velocity–depth relationships represent horizontal averages over thousands of kilometers. We know that velocity variations exist over length scales much smaller than that. New tomographic techniques are able to delineate such variations and allow us to image them.

Tomographic techniques

Knowledge of the three-dimensional upper mantle structure comes from tomographic techniques, notably:

- Transmission tomography of body wave delay times for both P- and S-velocity structure
- Inversion of the splitting of normal mode frequencies by lateral heterogeneity, for long wavelength heterogeneity in both P and S velocity
- Phase or group delay tomography using surface waves, primarily sensitive to S velocity
- Imaging of waves reflected or converted at discontinuities ("receiver functions"), sensitive to the depth location of such discontinuities

For an extensive overview of transmission tomographic techniques, see Rawlinson et al. (2010). Each of these techniques has seen important progress in recent times. Ray theory is being replaced with first-order scattering theory that allows for the modeling of the frequency-dependent dispersion in delay times. Single station (or "single channel") stacking of converted or reflected phases is being replaced by multichannel techniques, in particular Kirchhoff-type stacking along diffraction hyperboles much as one does to migrate seismic sections in exploration seismics (see Bostock, 2009).

Seismic tomography allows the mapping of "anomalies," or deviations from a regular layering of the upper mantle. The most remarkable anomalies are listed here:

Subduction zones and the associated low velocity structure in the mantle wedge can be very clearly imaged because of increased seismicity when actively subducting. The observed velocity anomalies are consistent with average (cold) temperature anomalies of several hundred degrees.

Oceanic ridges are mapped best with surface waves, which leave the exact depth extent as well as the temperature anomaly somewhat uncertain.

Cratonic roots were originally detected through the large negative delay times they cause for S waves, consistent with a cold lithosphere (as much as 400° colder than oceanic lithosphere), but the depth extent is mapped using teleseismic surface waves. In the tectosphere model of Jordan (1981), the high density induced by the cooling is compensated by a chemical composition that is depleted in basaltic components.

Upper mantle plumes are rather narrow anomalies in the upper mantle, but can be imaged adequately using body wave delays as well as surface wave observations from field deployments of portable seismographs. The velocity anomalies are consistent with temperature anomalies of 200–300°C inferred from petrological studies.

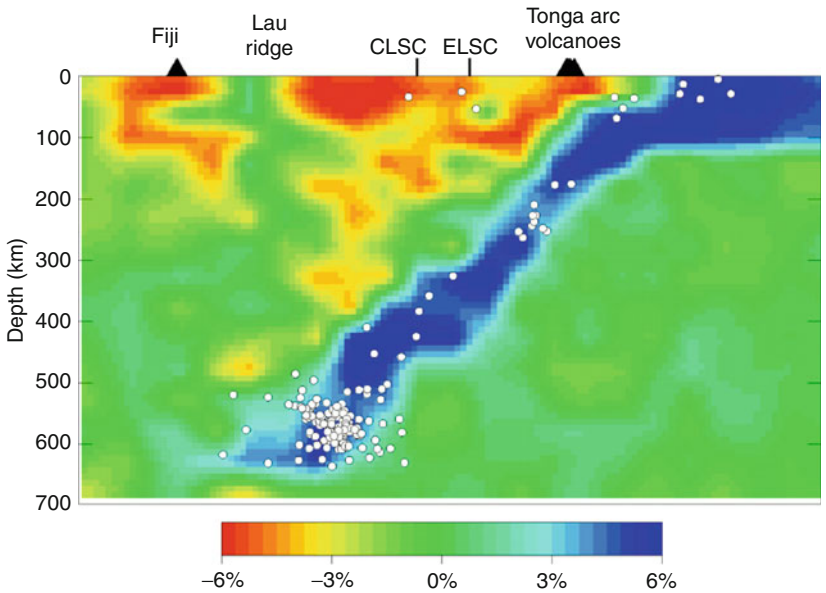
Topography of the "410"- *and* "660"-*km discontinuity* is globally imaged using underside reflections of waves such as SS. This involves some geographic averaging but topographic anomalies of 20 km or more are observed.

These anomalies are all visible in the variations of isotropic velocity variations. There is strong evidence, notably from the birefringence of SKS waves, that the seismic velocity in the shallow part of the upper mantle depends locally on the propagation direction of the wave; under favorable circumstances, this could in principle allow the mapping of mantle flow, since the crystal structure is assumed to align itself to the flow direction. However, at present there is little agreement between different tomographic models of anisotropy, partly because the SKS information is insensitive to depth, while the crustal correction, which is badly known, is in itself as large as the signal from anisotropy in surface waves (Ferreira et al., 2010).

Global models

Becker and Boschi (2002) review a large number of global tomography models of the mantle, and construct average models from a small selection of global models.

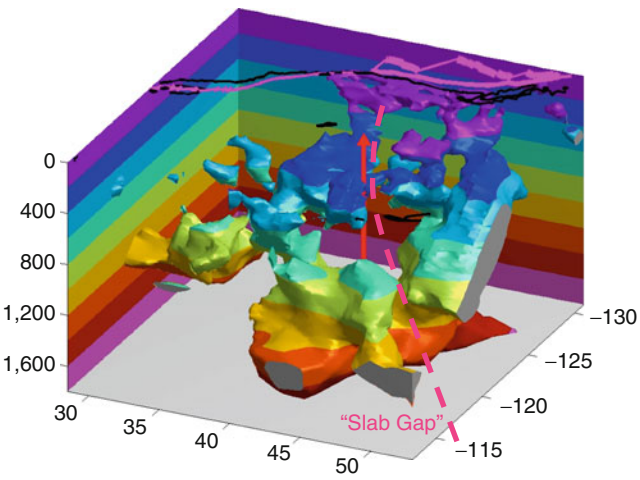
The map views displayed in Figures 1 and 2 illustrate the global features that are present in the upper mantle:



Earth's Structure, Upper Mantle, Figure 5 A west-to-east cross section through the Tonga subduction zone crossing the trench at 19S, 173W shows the subducting slab as a continuous feature down to 600 km depth (*blue positive anomaly*), and coinciding with the deep seismicity (*circles*). The negative velocity anomaly above the slab represents most likely the effect of dehydration. Active volcanism on Fiji and Tonga is indicated by *black triangles*. Also indicated are the Central and Eastern Lau Spreading Centers (Zhao et al., 1997).

old continental shields (Canada, Scandinavia, Western Australia, west and south Africa) are characterized by high S velocities down to depths of about 250 km. These anomalies extend to more than 400 km in the maps of P velocity, but P-wave delay times lack the depth resolving power that the surface wave data give to the S tomography. The oceanic ridges extend at least to 150-km depth in the S-velocity maps (Figure 2). The oceans are generally characterized by low S velocity throughout much of the upper mantle, with the exception of the S Atlantic and the westernmost Pacific. Plume-like anomalies, with a high degree of vertical continuity in the upper mantle, are visible in the P-velocity model beneath Hawaii, Society Islands (Tahiti), the East-African rift area (Afar). For a proper imaging of plumes, tomographic techniques that correct for the effects of diffraction around small or narrow anomalies are more suitable. Figure 3 shows models PRI-05 for P and S velocity in the upper mantle from Montelli et al. (2006), which clearly show a number of narrow upper mantle plumes.

Underside reflections from the upper mantle discontinuities allow the depth variations to the "410-" and "660-km" discontinuity to be mapped globally. This has most recently been done by Houser et al. (2008) using precursors to SS waves (Figure 4).



Earth's Structure, Upper Mantle, Figure 6 A view from the north-east at the subducted Farallon slab beneath western North America shows major ruptures that allow mantle rock to flow around it, such as the "slab gap" indicated by the *broken line*. The slab itself is delineated by a surface with an anomaly of +0.4% in P velocity (high velocity anomalies associated with the lithosphere are not mapped). The location of Yellowstone, another suspected passage of hot mantle rock toward the surface, is indicated by the *arrow*. Numbers along the axes indicate depth in kilometer, latitude, and longitude (negative for West). *Color shading* indicates depth. The plate boundaries between the North America, Pacific, and Juan de Fuca plates are indicated as *purple lines* at the surface (Figure courtesy Karen Sigloch).

Regional models

Regional studies are able to offer a better resolution than the global efforts, certainly when using data from large and dense seismic arrays such as Skippy (Australia), USArray (USA), or HiNet (Japan), but also using temporary deployments of portable seismic arrays. Efforts to image the upper mantle beneath them have led to very precise images of subduction zone structure in particular. Figure 5 shows the slab subducting in the Fiji–Tonga region.

Though this slab image is continuous down to 600-km depth, slab geometry in the upper mantle is very variable. Whereas some slabs remain for at least some time lodged in the transition zone others have no trouble sinking through the phase transition barrier at 660-km depth, with a negative Clapeyron slope and an increase of as much as two orders of magnitude in viscosity (Fukao et al., 2001).

A very different fate of subducting slabs is shown for the now inactive Farallon subduction beneath western North America, where the former Farallon plate is slowly disintegrating, letting mantle rock flow through major ruptures such as the "slab gap" in Figure 6 (Sigloch et al., 2008).

Outlook

It is now clear that the simple, layered, models of the upper mantle are inadequate to explain many of the dynamic processes we observe at the Earth's surface. The resolving power of tomographic techniques is still growing because of theoretical improvements and the densification of seismic networks. The coming decades will undoubtedly reduce the high degree of nonuniqueness that still plagues tomographic images, and allow for more reliable estimates of temperature and anisotropy. New instrumentation will open up the oceans for dense seismic observations.

Bibliography

Becker, T. W., and Boschi, L., 2002. A comparison of tomographic and geodynamic mantle models. *Geochemistry Geophysics Geosystems*, **3**(1), 1003.

Bostock, M. G., 2009. Teleseismic body-wave scattering and receiver-side structure. In Romanowicz, B., and Dziewonski, A. (eds.), *Seismology and Structure of the Earth*. Amsterdam: Elsevier, pp. 219–246.

Dziewonski, A. M., and Anderson, D. L., 1981. Preliminary reference Earth model. *Physics of the Earth and Planetary Interiors*, **25**, 297–356.

Ferreira, A. M. G., Woodhouse, J. H., Visser, K., and Trampert, J., 2010. On the robustness of global radially anisotropic surface wave tomography. *Journal of Geophysical Research*, **115**, B04313, doi:10.1029/2009JB006716.

Fukao, Y., Widiyantoro, S., and Obayashi, M., 2001. Stagnant slabs in the upper and lower mantle transition region. *Reviews of Geophysics*, **39**, 291–323.

Gu, Y. J., Dziewonski, A. M., and Ekström, G., 2001. Preferential detection of the Lehmann discontinuity beneath continents. *Geophysical Research Letters*, **28**, 4655–4658.

Houser, C., Masters, G., Flanagan, M., and Shearer, P., 2008. Determination and analysis of long-wavelength transition zone structure using precursors. *Geophysical Journal International*, **174**, 178–194.
Jordan, T. H., 1981. Continents as a chemical boundary layer. *Philosophical Transactions of the Royal Society of London*, **301A**, 359–373.
Montelli, R., Nolet, G., Dahlen, F. A., and Masters, G., 2006. A catalogue of deep mantle plumes: new results from finite-frequency tomography. *Geochemistry Geophysics Geosystems*, **7**, Q11007.
Rawlinson, N., Pozgay, S., and Fishwick, S., 2010. Seismic tomography: a window into deep earth. *Physics of the Earth and Planetary Interiors*, **178**, 101–135.
Sigloch, K., McQuarrie, N., and Nolet, G., 2008. Two-stage subduction history under North America inferred from multiple-frequency tomography. *Nature Geoscience*, **1**, 458–462.
Zhao, D., Xu, Y., Wiens, D. A., Dorman, L., Hildebrand, J., and Webb, S., 1997. Depth extent of the Lau back-arc spreading center and its relation to subduction processes. *Science*, **278**, 254–257.

Cross-references

Body Waves
Geodynamics
Gravity Field of the Earth
Lithosphere, Continental
Mantle Convection
Mantle Plumes
Plate Driving Forces
Seismic Tomography
Seismic, Receiver Function Technique
Surface Waves
Traveltime Tomography Using Controlled-Source Seismic Data

EARTHQUAKE LIGHTS

John S. Derr[1], France St-Laurent[2], Friedemann T. Freund[3], Robert Thériault[4]
[1]Tijeras, NM, USA
[2]LaSalle, QC, Canada
[3]NASA Ames Research Center/San Jose State University, Moffett Field, CA, USA
[4]Quebec Ministry of Natural Resources and Wildlife, Québec, QC, Canada

Synonyms

Luminous phenomena associated with earthquakes

Definition

Earthquake lights are anomalous luminosities associated with and presumably caused by the accumulation and release of stress of the earthquake process, observed prior to or during a seismic event and/or during the aftershock sequence. There is some disagreement over whether the classification Earthquake Lights (EQL) applies to similar luminous phenomena observed in earthquake-prone areas or along faults without immediately associated seismic activity. These may be due to locally high stress levels in the Earth's crust that are subsequently relieved without catastrophic rock failure, or by earthquakes that are too distant or late relative to the observed EQL.

Classification of earthquake lights

The earliest known report of earthquake lights dates to 373 BCE when the Greek cities Helice and Buris were destroyed by an earthquake accompanied by "immense columns of fire" (Seneca). Ancient references will always be questionable, especially when the surviving accounts are written later, but such references indicate that the people at that time were aware that lights might accompany earthquakes. Many sightings from various countries around the world have since been reported. They have been summarized in a number of papers (see, e.g., St-Laurent et al., 2006; Derr, 1973). As a result of extensive observations associated with the Saguenay earthquake sequence from November 1988 to January 1989, St-Laurent (2000) examined and extended earlier EQL classifications (e.g., Montandon, 1948). Six types of luminous phenomena have been documented:

1. "Seismic lightning," similar to ordinary lightning, but seismic lightning is typically of longer duration, sometimes illuminating large areas of the sky from below without thunder. "Seismic lightning" can be mistaken for "sheet" or "heat" lightning at night, which is too distant for thunder to be heard.
2. Luminous bands in the atmosphere at indeterminate height, sometimes horizontally or vertically and sometimes in a bundle, similar in appearance to some polar auroras although at much lower altitude, possibly even extending from the ground level upward.
3. Globular incandescences (moving or static), sometimes attached to luminous bands, called "orbs," "globes," or "meteors." They have the appearance of ball lightning, for example, of luminous spheres floating in midair, sometimes coalescing, lasting for up to a few minutes.
4. Fire tongues, small "flames" flickering or creeping along or near the ground, or like ignis fatuus.
5. Seismic "flames" seen emerging from the ground like an evanescent gas flame but very rarely causing visible burns.
6. Coronal and point-discharge-like lights.

A possible seventh type might be "luminous clouds" like those filmed shortly before the M7.9 Wenchuan, China, earthquake of May 12, 2008 (Panoramio photo, 2008). These are phenomena in the high atmosphere, similar to "fire rainbows," possibly linked to processes at the upper edge of the atmosphere, and/or to ionospheric perturbations.

Examples of EQL

While reliable pictures of EQL are scarce, one example of type 1 comes from the Matsushiro earthquake swarm from 1965 to 1967, Figure 1 (Yasui, 1968, 1971; discussed in Derr, 1973). An example of type 3 was photographed at Tagish Lake, Yukon Territory, Canada, Figure 2 (Jasek, 1998). While it is not possible to say that this picture is definitely EQL, these orbs are typical of the most-

frequently observed EQL, as described by Montandon (1948). The only known photo of "flames," type 5, was taken during the aftershock sequence to the Vrancea, Romania, earthquake on March 4, 1977, Figure 3 (Hedervari, 1983; discussed in St-Laurent, 2000). A sketch of type 6 comes from the account by Joseph A. Dallaire during the Saguenay 1988 mainshock, Figure 4 (presented in St-Laurent, 2000). These examples are notable for the differences they exhibit: illumination of a broad area of the sky, discrete glowing plasmas, evanescent "flames," and an electric discharge bursting through the surface of the Earth. They are representative of a number of other sightings, although the last one, the rapidly moving electric discharge, is so far unique in the detail of its description.

Earthquake Lights, Figure 1 Earthquake lights from Mt. Kimyo, Matsushiro area, Japan, September 26, 1966, 0325 (JST). Luminosity lasted 96 sec (Derr, 1973; Yasui, 1968). (© Seismological Society of America).

Example of type 3 EQL

Schmidt and Mack (1912; Translation by Steven Soter, personal communication) give a number of eye-witness descriptions of type 3 EQL. The most detailed was by Friedrich Konzelmann, foreman in an Ebingen factory:

> On the evening of 16 November between 10 and 10:30 I went home with my wife. My house is about 200 m from the city on the national road to Strassberg. As we were about 30 m from our house, I heard a distant noise like thunder from Sigmaringen, followed by a faint vibration. As I glanced in the direction from which the noise came, I suddenly noticed at some distance a bright flash from the ground, which then at a considerable height turned into a ball of light, about the size of 20 suns. This ball maintained its brightness about 2–3 seconds and then divided itself like lightning in the direction of Ebingen. In my opinion the earthquake came from southeast to northwest. The earthquake began with the lighting of the fireball. After the rolling went past, I looked backwards to the city and saw above the Katzenbuckel a similar but somewhat smaller fireball. The whole surroundings were brightly illuminated.

Source of EQL and associated electromagnetic (EM) phenomena

The leading theory of how these various light phenomena are generated, based on extensive laboratory experiments, attributes them to effects by mobile positive hole charge

Earthquake Lights, Figure 2 Earthquake lights from Tagish Lake, Yukon-Alaska border region, around July 1, probably 1972 or 1973 (exact date unknown). Estimated size: 1 m diameter. Closest orbs slowly drifted up the mountain to join the more distant ones (Jasek, 1998). (Photo credit: Jim Conacher, used with permission).

Earthquake Lights, Figure 3 One of five pictures taken around March 23, 1977 near Brasov, Romania, about 100 km NW from the Vrancea epicenter and during the aftershock sequence of the M 7.2 mainshock that occurred on March 4. The image corresponds to what the witness observed. (St-Laurent, 2000; Hedervari, 1983) (© Seismological Society of America).

Earthquake Lights, Figure 4 Drawing by artist from witness' description of a bright, fast moving bluish-white light, preceded and accompanied by crackling noise emitted by the trees. Laterrière, Québec, 19 km N of the epicenter, observed a few seconds before the M 5.9 mainshock on November 25, 1988. Not to scale. (St-Laurent, 2000) (© Seismological Society of America).

carriers, which become activated when rocks are subjected to high levels of stress (Freund, 2002, 2010). Positive holes exist in common rocks in a dormant state as peroxy links, $O_3Si/^{OO}\backslash SiO_3$. Under stress, dislocations sweeping through the mineral grains cause the peroxy links to break. They thereby activate positive holes, which flow down stress gradients, constituting an electric current with attendant magnetic field variations and electromagnetic (EM) emissions. As these electronic charge carriers accumulate at the Earth surface, they create electric microfields strong enough to field-ionize air molecules and produce EQL. Positive holes can also recombine at the surface, leading to a spectroscopically distinct IR emission identified in laboratory experiments and night-time infrared satellite images. The positive hole theory accounts not only for EQL but also for other pre-earthquake phenomena such as:

1. Air ionization at the ground-to-air interface.
2. Changes in the electrical conductivity of the soil.
3. Geo-electric and geomagnetic anomalies in the Earth's crust.
4. Ionospheric perturbations.
5. Ultralow and extremely low frequency (ULF/ELF) and radio frequency (RF) emissions.
6. Anomalous infrared emissions from around a future epicentral area.
7. Anomalous fog/haze/cloud formation and unusual animal behavior.

Conclusion

At present there is no consensus among seismologists on the phenomenology and physical mechanism of EQL and associated EM phenomena. However, if the positive hole theory is correct, the entire suite of pre-earthquake phenomena, including EQL, may become useful in forecasting earthquakes.

Bibliography

Derr, J. S., 1973. Earthquake lights: a review of observations and present theories. *Bulletin of the Seismological Society of America*, **63**, 2177–2187.

Freund, F. T., 2002. Charge generation and propagation in rocks. *Journal of Geodynamics*, **33**, 545–572.

Freund, F. T., 2010. Toward a unified theory of pre-earthquake signals. *Acta Geophysica*, **58**, 719–766.

Hedervari, P., 1983. *Second Circular, Project on Collection and Evaluation of Data on Earthquake Light Phenomena, Georgiana Observatory*. Budapest: Center for Cosmic and terrestrial Physics.

Jasek, M., 1998. Tagish Lake UFO photo report. http://www.ufobc.ca/yukon/tagish.htm.

Montandon, F., 1948. Lueurs et malaises d'origine séismique. *Geographica Helvetica*, **3**(1), 157–178.

Panoramio photo, 2008. http://www.panoramio.com/photo/19149469. Also available at http://www.youtube.com/watch?v=KKMTSDzU1Z4.

Schmidt, A. V., and Mack, K., 1912. Das Süddeutsches erdbeben vom 16 November 1911. Abschnitt VII: Lichterscheinung. *Württembergische Jahrbücher fur Statistik und Landskunde*, Jahrg. part 1, 131–139.

Seneca, *Natural Questions*, VI, 26.2–4.

St-Laurent, F., 2000. The Saguenay, Québec, earthquake lights of November 1988–January 1989. *Seismological Research Letters*, **71**, 160–174.

St-Laurent, F., Derr, J. S., and Freund, F. T., 2006. Earthquake lights and the stress-activation of positive hole charge carriers in rocks. *Physics and Chemistry of the Earth*, **31**, 305–312.

Yasui, Y., 1968. A study of the luminous phenomena accompanied with earthquake (part 1). *Memoirs of the Kakioka Magnetic Observatory*, **13**, 25–61.

Yasui, Y., 1971. A study of the luminous phenomena accompanied with earthquake (part 2). *Memoirs of the Kakioka Magnetic Observatory*, **14**, 67–68.

Cross-references

Earthquake Precursors and Prediction
Electrical Properties of Rocks
Electrical Resistivity Surveys and Data Interpretation
Geoelectromagnetism

EARTHQUAKE PRECURSORS AND PREDICTION

Seiya Uyeda[1], Toshiyasu Nagao[2], Masashi Kamogawa[3]
[1]Japan Academy, Taito-ku, Tokyo, Japan
[2]Earthquake Prediction Research Center, Institute of Oceanic Research and Development, Tokai University, Shimizu-ku, Shizuoka-shi, Shizuoka, Japan
[3]Department of Physics, Tokyo Gakugei University, Koganei-shi, Tokyo, Japan

Definitions and introduction

Earthquake (EQ) prediction or forecast is to specify the source location, magnitude M, and occurrence time of EQ with certain accuracy before its occurrence. The terms prediction and forecast are often used with different nuances, the latter allowing for more stochastic nature. We will mainly use the former term in this article. EQ prediction is often roughly classified by the length of concerned time into Long Term (10^2–10^1 years), Intermediate Term (10^1–1 years) and Short Term (<1 year). The latter of course includes "Imminent" or "Very Short Term" ones like weeks to hours prediction. EQ precursors are anomalous phenomena, which take place before EQs so that prediction is made on their information. This article will focus on the precursors for Short Term prediction. At this stage, however, the existence itself of the precursors has to be more firmly established as a scientific fact and further details of them must be thoroughly investigated.

Various EQ precursors have been noticed throughout human history until present day (Tributsch, 1982). They are mostly short term precursors on animals, plants, clouds, and sounds (see *Earthquake Sounds*). Precursors perceived without aid of scientific instruments are called macro- or macroscopic anomalies.

Precursors instrumentally recorded are classified into long/intermediate and short-term ones. Long/intermediate Term precursors: Active faults serve as geological precursors. Trenching of active seismic faults provides the dates and displacements of past large EQs, from which those of the next large EQ may be inferred. Other long/intermediate precursors include (1) geodetic: land deformation, changes in sea-level, tilt, strain; (2) seismological: seismic gaps (see *Characteristic Earthquakes and Seismic Gaps*), quiescence (see *Seismic Quiescence and Activation*), S wave splitting (see *Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting*), changes in micro-seismicity, seismic wave velocity, *b*-values, characteristic EQs (see *Characteristic Earthquakes and Seismic Gaps*); (3) changes in electrical resistivity; and (4) geochemical/hydrological: changes in radon, CO_2 and other gas discharge (earthquakes, thermal anomalies), and level and temperature of underground water (see *Seismic Signals in Well Observations: pre, co, post*). Some, in particular, those in (3) and (4) serve for short-term EQ prediction also.

Short-term precursors which are of major concern of this article are mainly geochemical/hydrological and electric, magnetic, and electromagnetic (EM). The latter are changes in (1) geoelectric field (telluric current), geomagnetic field, EM emissions at various frequencies (DC to VLF), (2) transmission properties of man-made EM waves of VLF to VHF range, (3) Ionosphere (foF2 and total electron contents, TEC), and (4) infrared (IR) radiation. These changes are measured by apparatus on the ground or those mounted on artificial satellite.

History

Zechariah of Israel (ca. 520 BC) is said to have issued the first long-term EQ forecast, based on the surface rupture of an EQ in 759 BC. Precursors were recorded as early as in 373 B.C in Greece. Many followed from many parts of the world in the form of documents and folklore. In later years, even larger number of macro-anomalies are kept accumulated (Tributsch, 1982), now including such items as radio and TV noise, e.g., at 1995 M7.3 Kobe, 1999 M7.4 Izmit, 1999 M7.7 Chichi EQs (Ikeya, 2004). Most of the macro-anomalies, however, have not been well studied systematically, although they apparently played significant roles, e.g., in the success of 1975 M7.3 Haicheng EQ prediction (Raleigh et al., 1977).

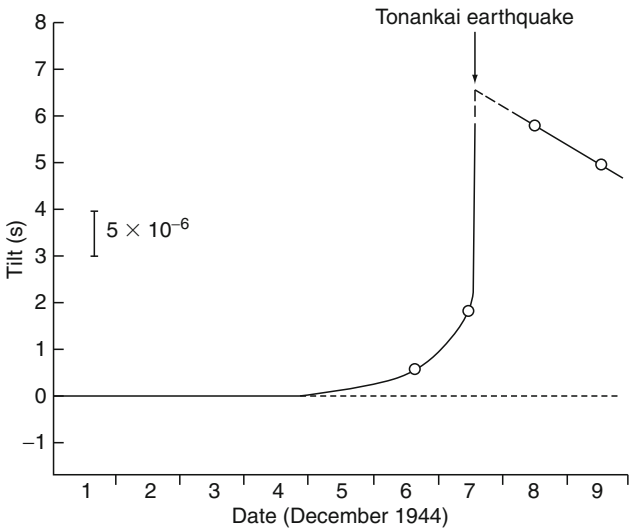
After the 1906 M8.3 San Francisco EQ, through investigation of geodetic data on faults, the elastic rebound theory (see *Earthquakes and Crustal Deformation*) was developed, laying the theoretical foundation for EQ prediction. From the 1960s to the 1970s, countries including USSR, Japan, China, and USA embarked on national-scale EQ prediction projects. Projects consisted mainly of statistical approach for long-/intermediate-term prediction (see *Statistical Seismology*) but there was also deterministic approach to short-term prediction. In early 1970s, the dilatancy-diffusion model (Sholtz et al., 1973) appeared. This model took advantage of results of field and laboratory studies obtained in various countries, like 10–20% Vp/Vs change in Garm, USSR. (Semyenov, 1969) and USA (Aggarwal et al., 1973), crustal uplift in Japan (Danbara, 1981), radon emission and electrical resistivity (Sadovsky et al., 1972) and so forth. All the quoted precursors, long and short term, appeared to support the physical picture of the dilatancy-diffusion model. The famous prediction of the 1975 M7.3 Haisheng EQ also was based on all kinds of precursors (Raleigh et al., 1977). Intermediate term measures were taken in June 1974, and short term in December, 1974. Emergency Alert to the public was issued in the early afternoon of February 4, 1975, several hours before the main shock to evacuate the public in time. These and other encouraging works in the 1960s–1970s made the whole prediction community highly optimistic (e.g., Press, 1975).

However, the period of optimism did not last long. First, later studies using controlled sources (McEvilly and Johnson, 1974) did not confirm the significant change of Vp/Vs, which was the most critical background of the dilatancy-diffusion model. Second, for the 1976 M7.8

Tangshan EQ, no timely alert was issued because there was no foreshock activity (see *Earthquake, Foreshocks*), despite many other precursors, causing the loss of over 240,000 lives (Chen et al., 1988).

In fact, the traditional seismological/geodetic methods were unable to short term predict any great EQs (see *Great Earthquakes*) other than the Haisheng EQ. For none of these EQs, any pre-slip, the most legitimate short-term precursor from seismological point of view (e.g., Shibazaki and Matsu'ura, 1995; Kato and Hirasawa, 1999) has been detected. Yet, the main strategy of Japan Meteorological Agency (JMA) for short-term prediction of the Tokai EQ is to catch pre-slips as reported for 1944 M7.9 Tonankai EQ (Figure 1, Mogi, 1985). The Tokai EQ has already been at a forecasted status for well over 30 years (Ishibashi, 1981). Another well-publicized, unsuccessful example is the Parkfield case, where M6 class event was predicted in 1985 to come within 5 years and 72 hour; alert was issued in 1992 (Roeloffs, 1994). It did not come until 2004, 12 years later (Bakun et al., 2005).

As a matter of fact, seismological precursors and methods, such as seismic gap (see *Characteristic Earthquakes and Seismic Gaps*), M8 algorithm (see *Earthquake Prediction, M8 Algorithm*), ZMAP (Wiemer and Wyss, 1994), hotspots (Holliday et al., 2006), etc., were not aiming at short-term prediction. The new development and discoveries in seismology such the asperity models (Yamanaka and Kikuchi, 2004), slow/silent EQs (see *Slow Earthquake*), and low-frequency nonvolcanic tremors (Obara, 2002) greatly contributed to the understanding of EQ mechanisms, but their direct relevance to short-term prediction has not been made clear. Characteristic EQ (see Characteristic Earthquakes and Seismic Gaps) may be the only case that can be quasi-short-term predictable by seismic monitoring, but they are rare cases. By the late 1970s, the general view of the main stream community became entirely pessimistic (e.g., Evernden, 1982). This pessimism has persisted until now, making the word prediction almost "forbidden."



Earthquake Precursors and Prediction, Figure 1 Precursory crustal deformation observed just before the M7.9 Tonankai EQ in 1944 (Mogi, 1985).

It may be noted, however, that numerous undeniable short-term precursors were observed mainly in non-seismic/geodetic (non-mechanical) fields. On top of many macro-anomalies, they were radon, CO_2 and other gas discharges, water level and electric, magnetic and EM anomalous changes.

Short-term EQ prediction may be realized by a more intensive search for short-term precursors not only through traditional, e.g., GPS detected pre-seismic deformation in Kamchatka as shown in Figure 2 (Gordeev et al., 2001), pre-seismic velocity changes at the Parkfield SAFOD drill site (Niu et al. 2008), but also through developing non-mechanical approach.

Non-mechanical short-term precursors

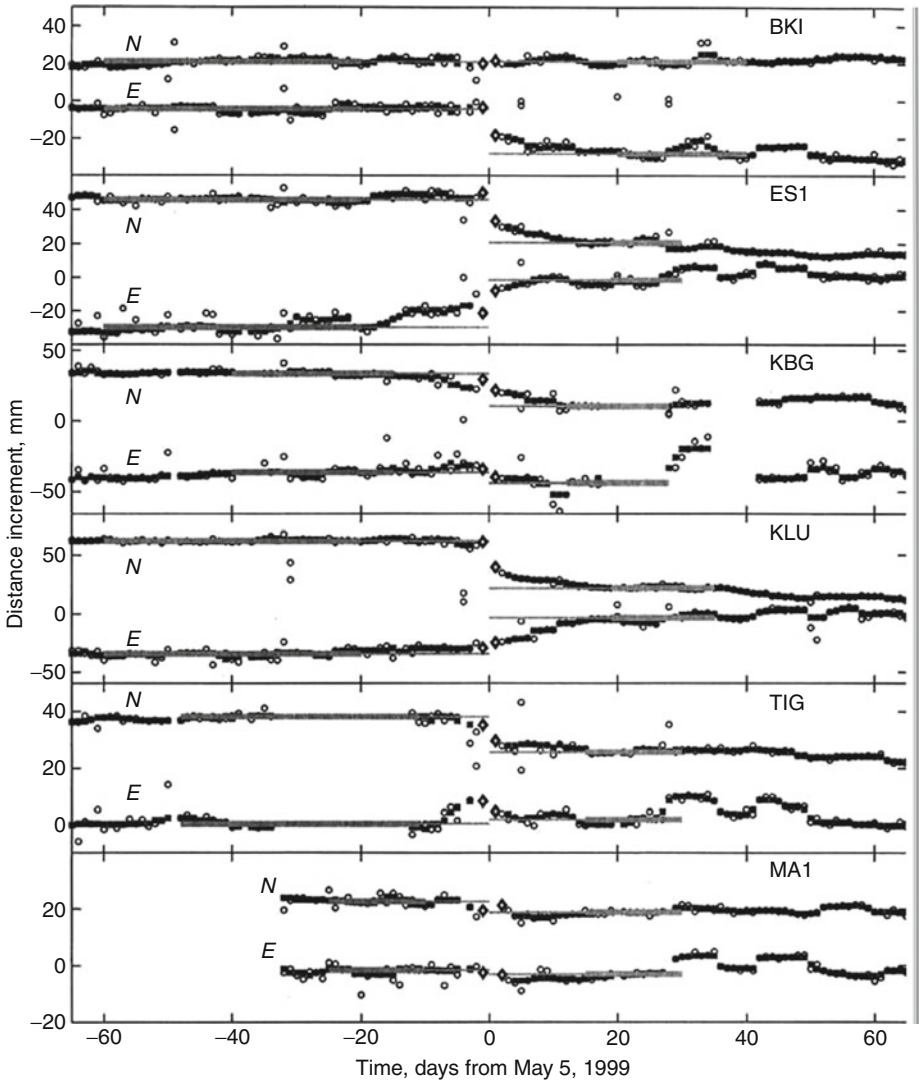
Difficulties inherent to short-term precursors

Short-term EQ prediction by standard mechanical approach seems to have scarce chance because the instant the stress reaches critical state can hardly be pinpointed by the present seismic or geodetic observations. Non-mechanical precursors make use of some phenomena which necessarily occur at that instant as by-products of the same cause, the stress build-up. The only necessary condition to make them useful for prediction is that their critical stress is slightly smaller than that of EQ. EQ would occur soon after the detection of them, perhaps aided by the last push of some triggers.

There are two rates to evaluate precursors. One is success rate, which is the number of successful predictions divided by the number of issued predictions, and the other is alarm rate, which is the number of successfully predicted EQs divided by the number of target EQs. In applying these evaluating rates, criteria of success and failure, i.e., the allowable errors in the three prediction targets, the dates, location and magnitude, must be clearly stated. When they are presented properly, however, criticizing them as causal relation is not clear or there are no other supporting geophysical simultaneous phenomena seems unreasonable.

Macro-anomalies

Macro-anomalies are numerous and the well-known examples, after the Haisheng EQ, are 1995 M7.3 Kobe, 1999 M7.4 Izmit, 1999 M7.7 Chichi EQs (Ikeya, 2004), and 2008 M8.0 Shuchuan EQ (Figure 3, Li et al., 2009). The causes of modern macro-anomaly precursors, such as radio, TV noise, malfunctioning of electrical appliances, are most probably related to pre-seismic EM emissions.



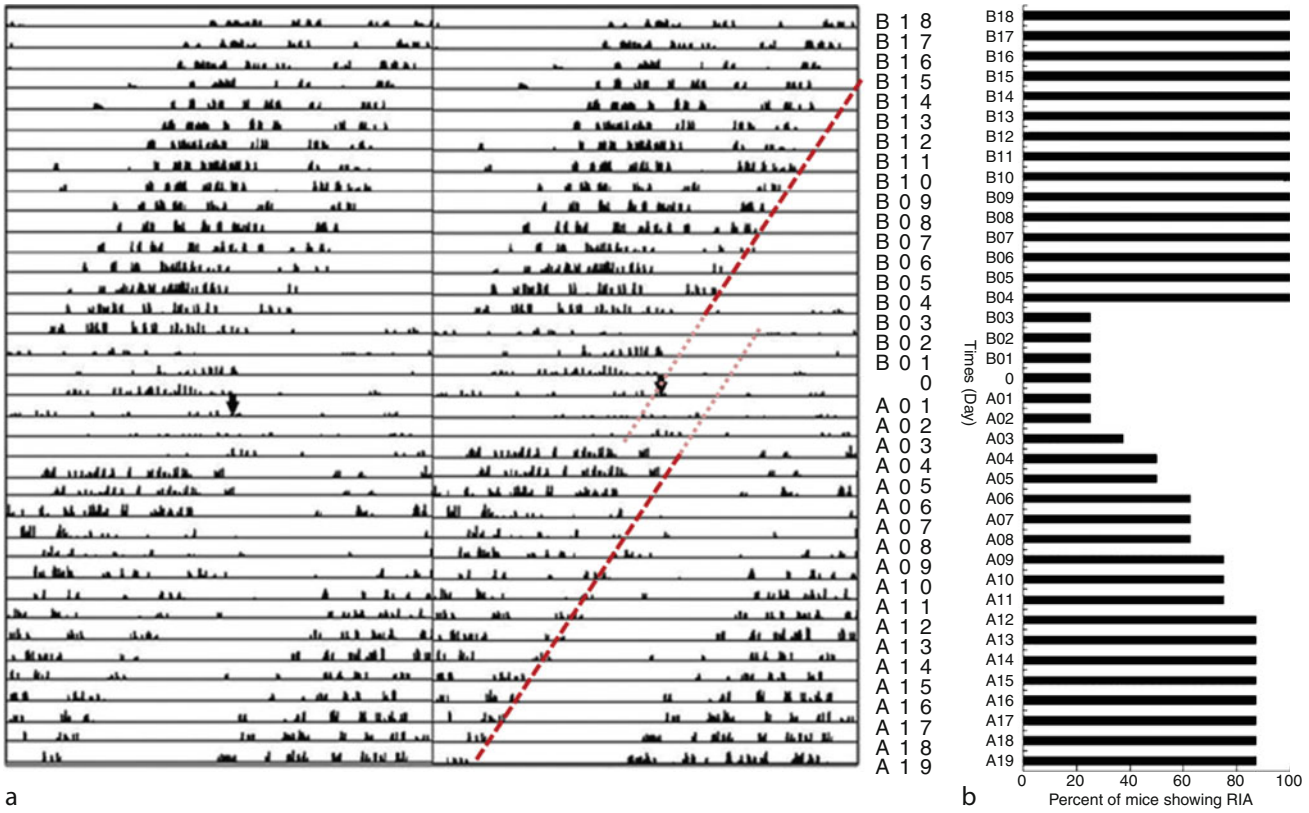
Earthquake Precursors and Prediction, Figure 2 The record of possible pre-slip before the M7.8 Kronotskoe EQ in 1997 (Gordeev et al., 2001).

Geochemical/hydrological precursors

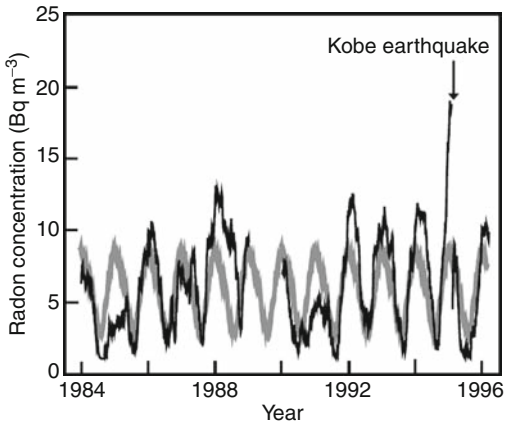
Geochemical/hydrological precursors are also diverse. Examples are discharges of radon, carbon dioxide, helium, H_2, and CH_4 (see Toutain and Baurbron 1999). Among them, radon emission has been most widely monitored, e.g., the 1976 M7.1 Gazli EQ, Uzbekistan, the 1985 M8.1 Mexico EQ, and the 1995 M7.2 Kobe EQ shown in Figure 4 (Yasuoka et al., 2009). Level changes of underground water are also potential precursors (e.g., Sidorenko et al., 1984 for the 1978 M7.8 Kuril EQ, Wang et al., 1984 for the Tangshan EQ, Igarashi et al., 1992 for the 1992 M5.9 Tokyo Bay EQ.). Geochemical/hydrological precursors are often observed at localities far (100 km or more) away from epicenters. Perhaps their appearances are heavily controlled by local geological situations such as faults and poroelastic conditions (see *Seismicity, Intraplate*). For water level and temperature, co-seismic changes are much more eminent and commonly observed than pre-seismic ones, in contrast to EM precursors (see later).

Satellite IR images as possible surface temperature

Anomalous satellite IR images as possible ground surface temperature precursor have been reported over seismic regions (earthquakes, thermal anomalies). The meaning of satellite-observed IR anomalies, however, is not clear. An alternative interpretation is that the satellite-observed IR image is not a temperature anomaly. When the p-holes (see later) arrive at the earth's surface, they recombine



Earthquake Precursors and Prediction, Figure 3 Abnormal animal behavior observed just before the M8.0 Wenchuan EQ in 2008 (Li et al., 2009). (**a**) Under free-running conditions, mouse locomotor activity presents circadian rhythm from B18 (18 days before EQ) to B04 (4 days before), and no rhythm from B03 (3 days before EQ) to A02 (2 days after EQ) including the day of EQ. Horizontal axis is 48 h (2 days). *Diagonal solid lines* mean the circadian rhythm does not coincide with 24 h. (**b**) The percentage of mice showing RLA (circadian rhythm of mouse locomotor activity) among the eight mice, indicating that some mice lost RLA around the EQ.



Earthquake Precursors and Prediction, Figure 4 Time series of the daily radon concentration observed from January 1984 to February 1996. The daily minimum variation is shown by a *black line* and the normal variation by a *gray line* (Yasuoka et al., 2009). Clear enhancement of radon concentration can be seen before the M7.3 Kobe EQ.

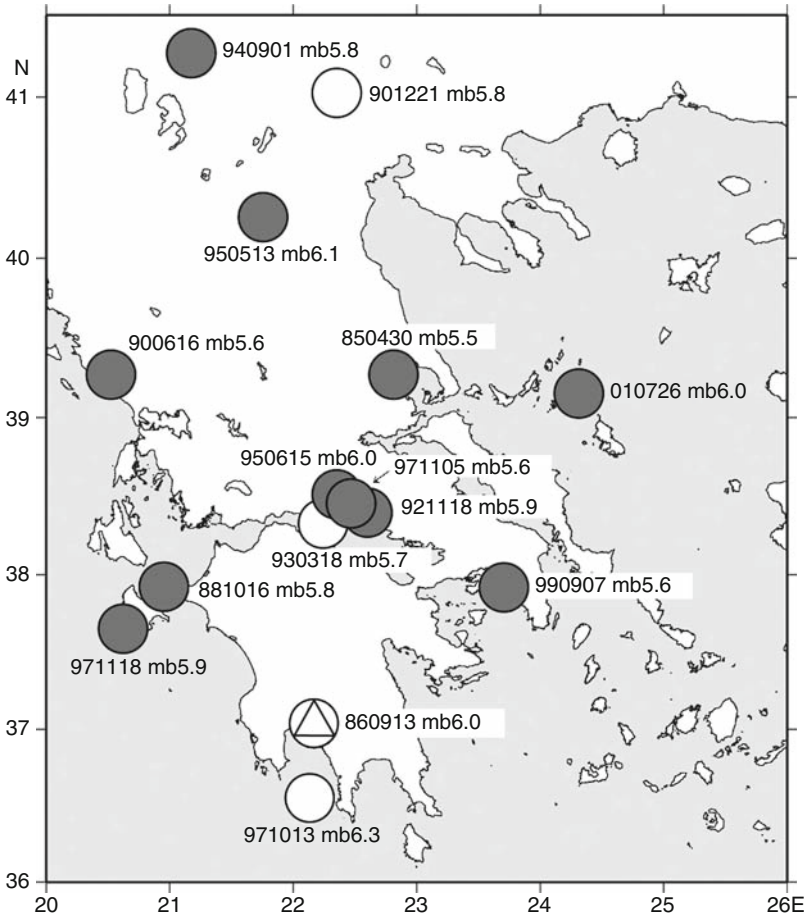
to cause a spectroscopically distinct non-thermal IR emission (Freund, 2009, 2010).

Electric, magnetic and EM precursors

Systematic research on EQ-related electric, magnetic, and EM phenomena were initiated more or less simultaneously in varied parts of the world in the 1980s. This relatively new branch of science is now called seismo-electromagnetics.

Telluric current precursors: the VAN method

The VAN method, developed in Greece since early 1980s, is named after initials of the founding scientists, P. Varotsos, K. Alexopoulos, and K. Nomikos (Varotsos and Alexopoulos, 1984a, b). This has been the only working system for real short-term prediction for more than two decades (Figure 5). For $M \geq 5$ Greek EQs, the criteria for successful prediction imposed by themselves are: $<$ a few weeks in time, <0.7 units in M, and <100 km in epicentral location. SES (seismic electric signals) are transient DC geopotential variations observed before EQs by dipoles of buried electrodes at separate sites and continuously monitored at many



Earthquake Precursors and Prediction, Figure 5 Evaluation of VAN prediction. All EQs with USGS PDE magnitude larger than 5.5 for 1985–2003 (Uyeda and Meguro, 2004). Shaded, white with triangle, and white circles represent "successfully" predicted, unsuccessfully predicted, and missed EQs.

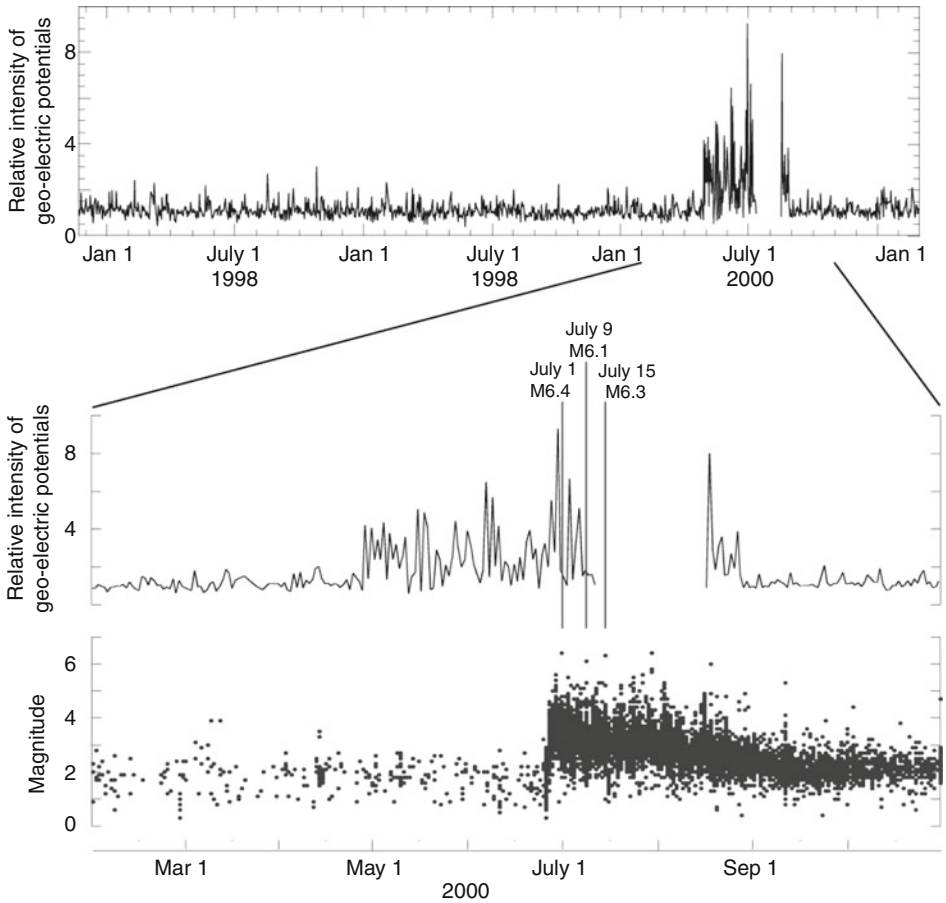
stations (e.g., Varotsos, 2005). At each station, several short (50~200 m) dipoles in both EW and NS directions and a few long dipoles (2~20 km) were installed. By adoption of multiple dipole system they accomplished successful noise rejection. Single SES precedes single EQ, whereas the so-called SES Activity, which consists of a number of SES in a short time, is followed by a series of EQs.

The VAN group made two major discoveries. One is the "Selectivity", which means that only some sites are sensitive to SES, and a sensitive site is sensitive only to SES from some specific focal area(s). This gives information for estimating the epicenter area. The selectivity is considered to be originated from the inhomogeneity of the subterranean electrical structures, i.e., SES goes only through underground conductive channels to sensitive sites. The other is the "VAN-relation" which means that the focal distance r, EQ magnitude M and the potential difference ΔV for dipole of length L are related by $\log(\Delta V/L \times r) = aM + b$ where a is a constant 0.34~0.37 and b is a site-dependent constant. Once the epicentral location is estimated from the selectivity data, M can be assessed since both $\Delta V/L$ and r are known. The VAN method has been a contentious subject (e.g., Geller, 1996; Lighthill, 1996). However, it has well survived all the criticism.

In Japan, VAN-type monitoring has been tried since the late 1980s (Uyeda et al., 2000). Despite the high-level noise of DC trains, the existence of the VAN type SES has been confirmed. In the year 2000, a two-month-long seismic swarm, with 7,000 $M \geq 3$ shocks and 5 $M \geq 6$ shocks, occurred in the Izu Island region. For this swarm activity, significant pre-seismic electric disturbances were observed (Figure 6, Uyeda et al., 2002).

Ultralow-frequency (ULF) precursors

A well-known example is the case of the 1989 M7.1 Loma Prieta EQ (Fraser-Smith et al., 1990). The horizontal component measured at a site 7 km from the epicenter started anomalous enhancement about 2 weeks prior to and a sharp increase a few hours before the EQ, as shown in Figure 7. The disturbance lasted for about 3 months after the EQ. These have never been observed at any other time

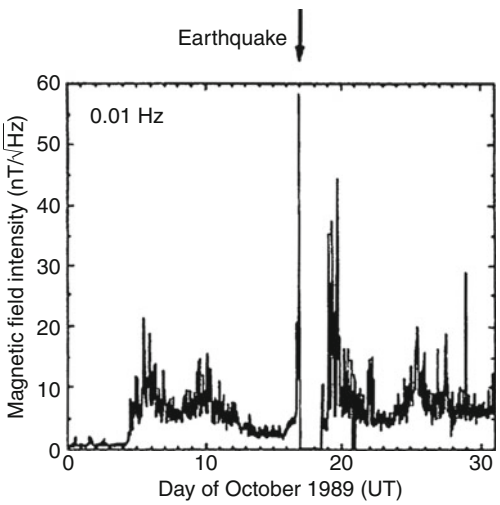


Earthquake Precursors and Prediction, Figure 6 Anomalous geo-electric potential changes in ULF (0.01 Hz) band prior to 2000 seismic swarm activity in Izu island region, Japan. The bottom panel shows seismicity. (Modified from Uyeda et al., 2002.)

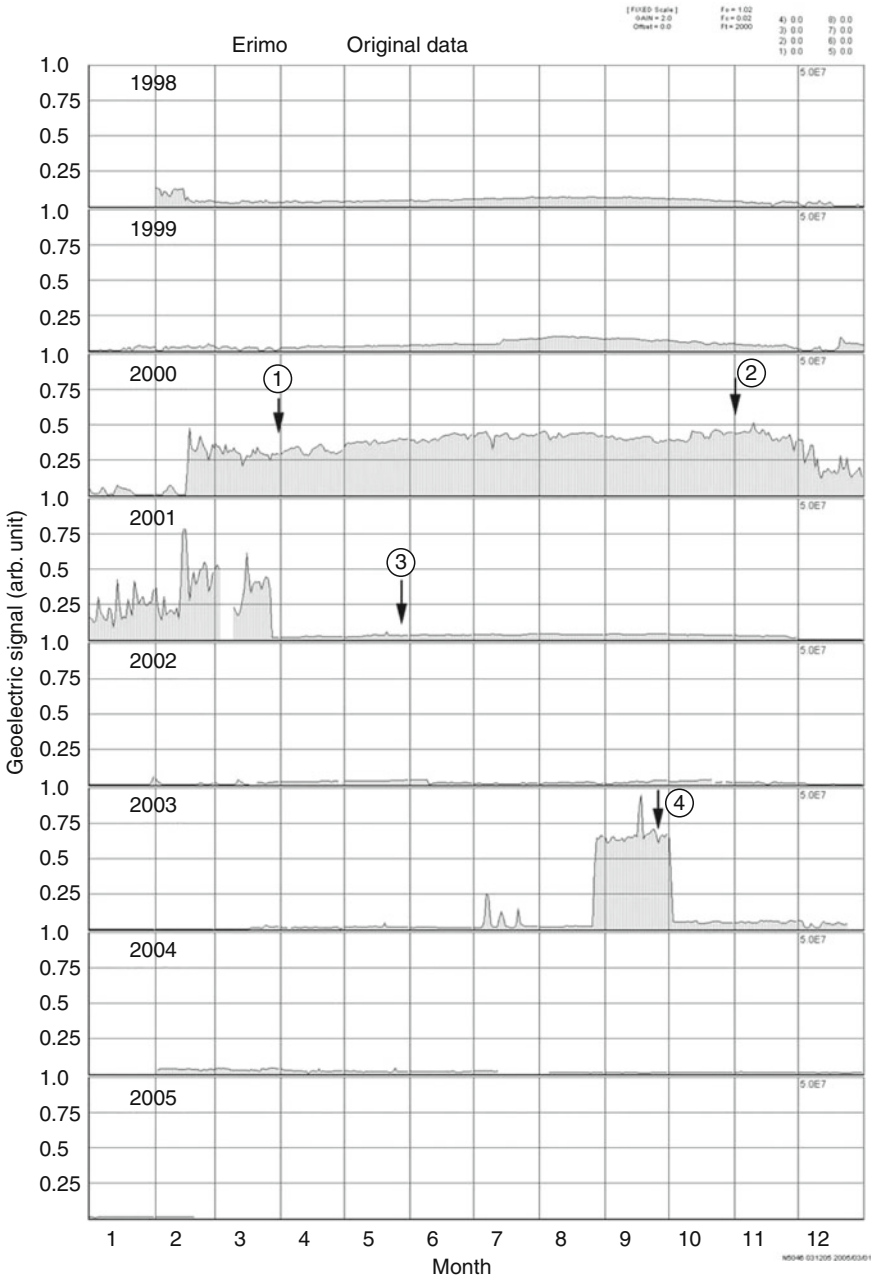
during the observation of more than 15 years. Reports of observing pre-seismic ULF geomagnetic anomalies have been made also for 1988 M6.9 Spitak (Armenia) EQ (Kopytenko et al., 1993) and 1993 M8.0 Guam (Marianas) EQ (Hayakawa et al., 1996). There are some discussions pointing out that these ULF signals might have been artifact or magnet-telluric origin (e.g., Thomas et al., 2009).

Higher-frequency EM emission

Gokhberg et al. (1982) was a pioneering observation of pre-seismic EM wave emission in the LF range. Asada and his group in Japan monitored the wave forms of two horizontal magnetic components of VLF EM waves, through which apparent incoming directions of VLF pulses were assessed (Asada et al., 2001). They found that, a few days before M5 class land EQs within 100 km of their stations, pulses with fixed incoming direction appeared and the EQ actually occurred in that direction, whereas the sources of overwhelmingly more numerous and stronger noises were moving along with lightning sources.



Earthquake Precursors and Prediction, Figure 7 Amplitude of the geomagnetic horizontal component at 0.01 Hz band observed before and after the M6.9 Loma-Prieta EQ (Fraser-Smith et al., 1990).



Earthquake Precursors and Prediction, Figure 8 Envelope of the geoelectric signal recorded at the Erimo site over 7 years since 1998. Japan Meteorological Agency (JMA) announcement of (1) the first eruption of Mt. Usu volcano, (2) decline in magma activity, (3) complete cessation of the magma activity, and (4) the main shock of M8.0 Tokachi-Oki EQ (Enomoto et al., 2006).

Warwick et al. (1982) reported on the observation of high-frequency radio waves possibly related to the precursory activity of 1960 Great M9.5 Chilean EQ. This was probably the first report of this kind. Enomoto et al. (2006) recorded two events of anomalous geo-electric current pulses (HF-band) at Erimo station, Hokkaido, Japan (Figure 8). One in 2000–2001 occurred before and during the volcanic activity of Mt. Usu (200 km away) and the second started one month before 2003 M8.0 Tokachi-Oki EQ (80 km away). These were the only anomalies during their 10-year observation period.

Possible mechanism of pre-seismic EM emissions

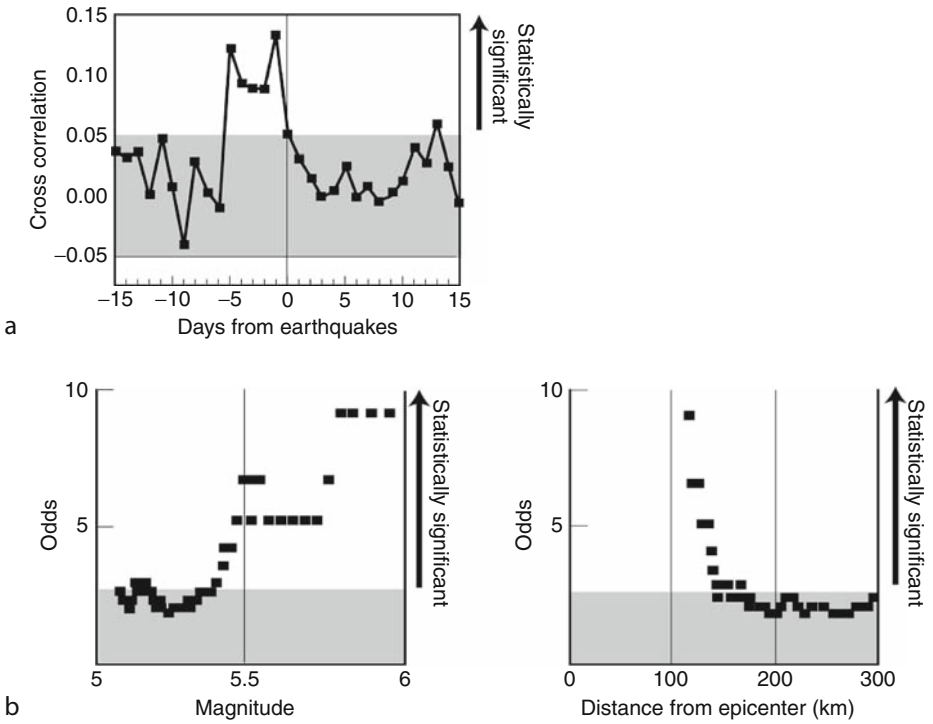
It is often questioned why EM signals, such as SES, appear only pre-seismically and not co-seismically. This point makes scientific community dubious about

EQ-related EM precursors in general because EQ is by far the most energetic event. Actually, co-seismic signals are routinely observed but they are not useful as precursors. Moreover, all the so called co-seismic EM signals observed so far were found to occur at the time of the arrival of seismic waves. They are, therefore, "co-seismic wave" and not "true" co-seismic. The fact that no true co-seismic signals are observed may be an important clue in exploring the physical mechanism of generation of both EM signals and EQs themselves. In fact, the slow pre-seismic stress buildup and the instantaneous stress release at EQ are physically very different processes and there is no compelling reason why they generate similar EM signals.

The generation mechanism is different for different frequencies. It may be the electro-kinetic effects (e.g., Mizutani et al. 1976; Fitterman, 1978) or pressure stimulated current (PSC) effects for DC to low-frequency signals, and piezo-electric effects and exo-electron emission (Enomoto and Hashimoto, 1990) for higher frequency signals. A mechanism involving the positive holes (p-holes) in rock forming minerals under stress has also been proposed.

The PSC was proposed by Varotsos and Alexopoulos (1986). The impurities and vacancies in crystals form randomly oriented local electric dipoles, which align their orientation in avalanche following an activation process to generate the electric current. PSC model is unique in generating currents during gradual stress build-up without requiring any sudden change of stress. Freund and his colleagues are proposing a different mechanism for ULF electric signals (Freund et al., 2006). They discovered in laboratory that when a block of igneous rock is put under local stress, the rock turns into a battery without external electric field. The charge carriers in this case are semiconductor-type positive holes (p-holes) which is oxygen in one-valence state. They attribute the satellite image of pre-seismic IR anomalies to the result of recombination of two p-hole charge carriers at earth's surface as mentioned earlier. However, whether their model based mainly on experiments on dry rocks works in nature where rocks are mostly wet is not so far clear.

Although models related to micro-cracking based on laboratory experiments such as discharge of screening charge of piezo-electric polarization (Ikeya, 2004; Yoshida et al., 1997), electrification of fresh crack surfaces (Yamada et al., 1989), and exo-electron have been proposed, there have been no reliable field reports on the occurrence of pre-main shock micro-cracking. Furthermore, it might be pointed out that in all these models much stronger co-seismic signals would be expected.



Earthquake Precursors and Prediction, Figure 9 Statistical significance of pre-seismic ionospheric anomaly (PEIA). (After Liu et al., 2006.) (**a**) Cross-correlation coefficient between EQs and anomalies with time lag. The shaded area denotes the rejection bounds at significance level 0.01. (**b**) Odds of the shallow EQs with the PEIA against the magnitude and distance. Odds versus related mean magnitude (*left*), and odds versus related mean distance (*right*). The *shaded areas* in the two plots denote the rejection value of odds as 2.5 under significance level of 0.01.

Lithosphere-atmosphere-ionosphere coupling

Sub-ionospheric anomalies before large EQs were reported (e.g., Molchanov and Hayakawa, 1998) through monitoring 10–20 kHz VLF ship-navigation waves, although Clilverd et al. (1999) did not obtain similar results. A decrease in electron density of ionosphere within 5 days before $M \geq 5$ class EQ was found in Taiwan by the ionosonde measurements of critical plasma frequency, foF2 (see Figure 9, Liu et al., 2006). Fujiwara et al. (2004) found pre-seismic atmospheric anomalies causing FM radio transmission anomalies (Kushida and Kushida, 2002) within 5 days before $M \geq 4.8$ EQs. Moriya et al. (2010) further promoted this study and obtained statistical relations between the total time length of anomalous reception and EQ magnitude. The concept of pre-seismic lithosphere-atmosphere-ionosphere coupling (LAI coupling) arising from these investigations became an important issue in precursor studies (Pulinets and Boyarchuk, 2005; Kamogawa, 2006), because some causative factor should exist in the lithosphere and there should be some mechanism to transport its energy to the high atmosphere. Recently, the French satellite DEMETER observed anomalous depression of VLF EM wave intensity at ionospheric height 0–4 h before EQs, which implies that the ionospheric disturbance may modulate the VLF propagation path (Němec et al., 2009). Possible LAI coupling processes have been proposed by many researchers but they are still at hypothetical stage.

Critical phenomena

EQ may be a critical phenomenon defined in statistical physics (e.g., Bak and Tang, 1989) which means that the short time EQ prediction is synonym for indentifying the approach to criticality. Although this is difficult to achieve (see Main, 1995), it has been recently shown that by analyzing time-series of seismicity in a newly introduced time domain "natural time," the approach to the state of criticality can be clearly identified (Sarlis et al., 2008). This way, they seem to have succeeded in shortening the lead-time of VAN prediction to only a few days (Uyeda and Kamogawa, 2008). This means, seismic data may play an amazing role of short-term precursor when combined with SES data.

Summary and future outlook

Most of the precursors mentioned above are possibilities, but so far they (except VAN's SES) were recognized or documented only after main shocks and they gave scarce constraints on the source location and magnitude M. Much more enhanced research is needed to make them practically useful precursors.

1. Some macro-anomaly short-term precursors are real. Scientific research on them is inadequate so far.
2. Seismic and geodetic precursors can be useful for long/intermediate EQ prediction. However, it is unlikely that they can be useful short-term precursors in near future.
3. Non-mechanical precursors, such as some geochemical, hydrological, and EM precursors, are plausible candidates for short-term EQ prediction. For sooner realization of practical short-term prediction, more research on them should be promoted.

Bibliography

Aggarwal, Y., Sykes, L., Armbruster, J., and Sbar, M., 1973. Premonitary changes in seismic velocities and prediction of earthquakes. *Nature*, **241**(5385), 101–104.

Asada, T., Baba, H., Kawazoe, K., and Sugiura, M., 2001. An attempt to delineate very low frequency electromagnetic signals associated with earthquakes. *Earth Planets and Space*, **53**, 55–62.

Bak, P., and Tang, C., 1989. Earthquakes as a self-organized critical phenomena. *Journal of Geophysical Research*, **94**(B11), 15,635–15,637.

Bakun, W. H., Aagaard, B., Dost, B., Ellsworth, W. L., Hardebeck, J. L., Harris, R. A., Ji, C., Johnston, M. J. S., Langbein, J., Lienkaemper, J. J., Michael, A. J., Murray, J. R., Nadeau, R. M., Reasenberg, P. A., Reichle, M. S., Roeloffs, E. A., Shakal, A., Simpson, R. W., and Waldhauser, F., 2005. Implications for prediction and hazard assessment from the 2004 Parkfield earthquake. *Nature*, **437**, 969–974.

Chen, Y., Tsoi, K., Chen, F., Gao, Z., Q. Zou, and Chen, Z., 1988. *The Great Tangshan Earthquake of 1976*. Tarrytown: Pergamon, 153 pp.

Clilverd, M. A., Rodger, C. J., and Thomson, N. R., 1999. Investigating seismo-ionospheric effects on a long subionospheric path. *Journal of Geophysical Research*, **104** (A12), 28171–28179.

Danbara, T., 1981. Geodesy and earthquake prediction. In Rikitake, T. (ed.), *Current Research in Earthquake Prediction I*. Tokyo: Center for Academic; Dordrecht: D. Reidel, pp. 167–220.

Enomoto, Y., and Hashimoto, H., 1990. Emission of charged particles from indentation fracture of rocks. *Nature*, **346**, 641–643.

Enomoto, Y., Hashimoto, H., Shirai, N., Murakami, Y., Mogi, T., Takada, M., and Kasahara, M., 2006. Anomalous geoelectric signals possibly related to the 2000 Mt. Usu eruption and 2003 Tokachi-oki earthquake. *Physics and Chemistry of the Earth*, **31**, 319–324.

Evernden, J. F., 1982. Earthquake prediction: what we have learned and what we should do now. *Bulletin of the Seismological Society of America*, **72**, 343–349.

Fitterman, D. V., 1978. Electrokinetic and magnetic anomalies associated with dilatant regions in a layered earth. *Journal of Geophysical Research*, **83**, 5923–5928.

Fraser-Smith, A. C., Bernardi, A., McGill, P. R., Ladd, M. E., Helliwell, R. A., and Villard, O. G., Jr., 1990. Low-frequency magnetic field measurements near the epicenter of the Ms 7.1 Loma Prieta earthquake. *Geophysical Research Letters*, **17**, 1465–1468.

Freund, F., 2009. Stress-activated positive hole charge carriers in rocks and the generation of pre-earthquake signals. In Hayakawa, M. (ed.), *Electromagnetic Phenomena Associated with Earthquakes*. New Delhi: Research Signpost, pp. 41–96. ISBN: 978-81-7895-297-0.

Freund, F., 2010. Toward a unified solid state theory for pre-earthquake signals. *Acta Geophysica*, **58**, 719–766, doi:10.2478/s11600-009-0066-x.

Freund, F., Takeuchi, A., and Lau, B. E. S., 2006. Electric currents streaming out of stressed igneous rocks – A step towards understanding pre-earthquake low frequency EM emissions. *Physics and Chemistry of the Earth*, **31**, 389–396.

Fujiwara, H., Kamogawa, M., Ikeda, M., Liu, J. Y., Sakata, H., Chen, Y. I., Ofuruton, H., Muramatsu, S., Chuo, Y. J., and

Ohtsuki, Y. H., 2004. Atmospheric anomalies observed during earthquake occurrences. *Geophysical Research Letters*, **31**, L17110, doi:10.1029/2004GL019865.
Geller, R. (ed.), 1996. Debate on VAN. *Geophysical Research Letters*. **23**(11) (Special Issue), 1291–1452.
Gokhberg, M. B., Morgounov, V. A., Yoshino, T., and Tomizawa, I., 1982. Experimental measurement of electromagnetic emissions possibly related to earthquakes in Japan. *Journal of Geophysical Research*, **87**(B9), 7824–7828.
Gordeev, E., Gusev, A., Levin, V., Bakhtiarov, V., Pavlov, M., Chevrov, V., and Kasahara, M., 2001. Preliminary analysis of deformation at the Eurasia-Pacific-North America plate junction from GPS data. *Geophysical Journal International*, **147**, 189–198.
Hayakawa, M., Kawate, R., Molchanov, O. A., and Yumoto, K., 1996. Results of ultra-low frequency magnetic field measurements during Guam earthquake of 8 August 1993. *Geophysical Research Letters*, **23**, 241–244.
Holliday, J. R., Rundle, J. B., Turcotte, D. L., Kristy, W. K., Tiampo, F., and Donnellan, A., 2006. Space-time clustering and correlations of major earthquakes. *Physical Review Letters*, **97**, 238501.
Igarashi, G., Wakita, H., and Sano, T., 1992. Precursory and coseismic anomalies in well water levels observed for February 2, 1992, Tokyo Bay earthquake. *Geophysical Research Letters*, **19**, 1583–1586.
Ikeya, M., 2004. *Earthquakes and Animals: From Folk Legends to Science*. Singapore: World Scientific, 316 pp. ISBN-13: 978–9812385918.
Ishibashi, K., 1981. Specification of a soon-to-occur seismic faulting in the Tokai district, central Japan, based upon seismotectonics. In Simpson, D. W., and Richards, P. G. (ed.), *Earthquake Prediction: An International Review.* Maurice Ewing Series. Washington: AGU, Vol. 4, pp. 297–332.
Kamogawa, M., 2006. Preseismic lithosphere-atmosphere-ionosphere coupling. *EOS*, **87**(40), 417–424.
Kato, N., and Hirasawa, T., 1999. A model for possible crustal deformation prior to a coming large interpolate earthquake in the Tokai district, central Japan. *Bulletin of the Seismological Society of America*, **89**, 1401–1417.
Kopytenko, Y. A., Matishvili, T. G., Voronov, P. M., Kopytenko, E. A., and Molchanov, O. A., 1993. Detection of ultra-low-frequency emissions connected with the Spitak earthquake and its aftershock activity, based on geomagnetic pulsation data at Dusheti and Vardzia observatories. *Physics of the Earth and Planetary Interiors*, **77**, 85–95.
Kushida, Y., and Kushida, R., 2002. Possibility of earthquake forecast by radio observations in the VHF band. *Journal of Atmospheric Electricity*, **22**, 239–255.
Li, Y., Liu, Y., Jiang, Z., Guan, J., Yi, G., Cheng, S., Wang, B., Fu, T., and Wang, Z., 2009. Behavior change related to Wenchuan devastating earthquake in mice. *Bioelectromagnetics*, doi:10.1002/bem.20520.
Lighthill, J. (ed.), 1996. *A Critical Review of VAN – Earthquake Prediction from Seismic Electrical Signals*. Singapore: World Scientific, 376 pp.
Liu, J. Y., Chen, Y. I., and Chuo, Y. J., 2006. A statistical investigation of pre-earthquake ionospheric anomaly. *Journal of Geophysical Research*, **111**, A05304, doi:10.1029/2005JA011333.
Main, I., 1995. Statistical physics, seismogenesis, and seismic hazard. *Reviews of Geophysics*, **34**, 433–462.
McEvilly, T. V., and Johnson, L. R., 1974. Stability of P and S velocities from central California quarry blasts. *Bulletin of the Seismological Society of America*, **64**, 342–353.
Mizutani, H., Ishido, T., Yokokura, T., and Ohnishi, S., 1976. Electrokinetic phenomena associated with earthquakes. *Geophysical Research Letters*, **3**, 365–368.
Mogi, K., 1985. Temporal variation of crustal deformation during the days preceding a thrust-type great earthquake – The 1944 Tonankai earthquake of magnitude 8.1, Japan. *PAGEOPH*, **122**, 765–780.
Molchanov, O. A., and Hayakawa, M., 1998. Subionospheric VLF signal perturbations possibly related to earthquakes. *Journal of Geophysical Research*, **100**, 1691–1712.
Moriya, T., Mogi, T., and Takada, M., 2010. Anomalous pre-seismic transmission of VHF-band radio waves resulting from large earthquakes, and its statistical relationship to magnitude of impending earthquakes. *Geophysical Journal International*, **180**(2), 858–870.
Němec, F., Santolík, O., and Parrot, M., 2009. Decrease of intensity of ELF/VLF waves observed in the upper ionosphere close to earthquakes: a statistical study. *Journal of Geophysical Research*, **114**, A04303, doi:10.1029/2008JA013972.
Niu, F., Silver, P., Daley, T., Cheng, X., and Majer, E., 2008. Preseismic velocity changes observed from active source monitoring at the Parkfield SAFOD drill site. *Nature*, **454**, 204–208.
Obara, K., 2002. Nonvolcanic deep tremor associated with subduction in southwest Japan. *Science*, **296**, 1679–1681.
Press, F., 1975. Earthquake prediction. *Scientific American*, **232**(5), 14–23.
Pulinets, S., and Boyarchuk, K., 2005. *Ionospheric Precursors of Earthquakes*, Berlin: Springer, 316 pp.
Raleigh, B., Benett, D., Craig, H., Hanks, T., Molnar, P., Nur, A., Savage, J., Scholz, C., Turner, R., and Wu, F., 1977. Prediction of Haisheng earthquake. *EOS*. **58**(5), 2.
Roeloffs, E., 1994. The earthquake prediction experiment at Parkfield, California, *Reviews of Geophysics*, **32**(3), 315–336.
Sadovsky, M., Nersesov, I., Nigumatullaev, S., Latynina, L., Lukk, A., Semenov, A., Simbireva, I., and Ulmov, V., 1972. The processes preceding strong earthquakes in some regions of Middle Asia. *Tectonophysics*, **14**, 295–307.
Sarlis, N. V., Skordas, E. S., Lazaridou, M. S., and Varotsos, P. A., 2008. Investigation of seismicity after the initiation of a seismic electric signal activity until the main shock. *Proceedings of the Japan Academy Series B*, **84**, 331–343.
Semyenov, A. N., 1969. Variation in the travel time of traverse and longitudinal waves before violent earthquakes. *Izvestiya Academy of Sciences USSR (Physics of the Solid Earth)*. **4**, 245–248 (English transl.)
Shibazaki, B., and Matsu'ura, M., 1995. Foreshocks and pre-events associated with the nucleation of large earthquakes. *Geophysical Research Letters*, **22**, 1305–1308.
Sholtz, C., Sykes, L. R., and Aggarwal, Y. P., 1973. Earthquake prediction: a physical basis. *Science*, **181**, 803–810.
Sidorenko, A., Sadovsky, M., Nersesov, I., Popov, E., and Soloviev, S., 1984. Soviet experience of prediction of earthquakes in U.S. S.R. and the prospect for its development. In *Proceedings of the International Symposium on Earthquake Prediction*. Unesco, pp. 565–573.
Thomas, J. N., Love, J. J., Johnston, M. J. S., and Yumoto, K., 2009. On the reported magnetic precursor of the 1993 Guam earthquake. *Geophysical Research Letters*, **36**, L16301, doi:10.1029/2009GL039020.
Toutain, J. P., and Baurbron, J. C., 1999. Gas geochemistry and seismotectonics: a review. *Tectonophysics*, **304**, 1–27.
Tributsch, H., 1982. *When the Snakes Awake: Animals and Earthquake Prediction*. Cambridge: MIT Press, 264 pp.
Uyeda, S., and Kamogawa, M., 2008. The Prediction of Two Large Earthquakes in Greece. *Eos TransAGU*, **89**, 39, doi:10.1029/2008EO390002.
Uyeda, S., and Meguro, K., 2004. Earthquake prediction, seismic hazard and vulnerability, in The State of the Planet: Frontiers and Challenges in Geophysics. *Geophysical Monograph*, **150**(19), 349–358.

Uyeda, S., Nagao, T., Orihara, Y., Yamaguchi, Y., and Takahashi, I., 2000. Geoelectric potential changes: possible precursors to earthquakes in Japan. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 4561–4566.
Uyeda, S., Hayakawa, M., Nagao, T., Molchanov, O., Hattori, K., Orihara, Y., Gotoh, K., Akinaga, Y., and Tanaka, H., 2002. Electric and magnetic phenomena observed before the volcano-seismic activity in 2000 in the Izu Island Region, Japan. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 7352–7355.
Varotsos, P. A., 2005. *The Physics of Seismic Electric Signals*. Tokyo: TerraPub, 358 pp.
Varotsos, P., and Alexopoulos, K., 1984a. Physical properties of the variations of the electric field of the earth preceding earthquakes, I. *Tectonophysics*, **110**, 73–98.
Varotsos, P., and Alexopoulos, K., 1984b. Physical properties of the variations of the electric field of the earth preceding earthquakes, II. *Tectonophysics*, **110**, 99–125.
Varotsos, P., and Alexopoulos, K., 1986. *Thermodynamics of Point Defects and their Relation with Bulk Properties*. Amsterdam: North Holland, 474 pp.
Varotsos, P., Sarlis, N., and Skordas, E., 2002. Long range correlations in the electric signals that precede rupture. *Physical Rev E*, **66**, 011902(7).
Wang, C., Wang, Y., Zhang, H., Li, Y., and Zhao, S., 1984. Characteristics of water-level variation in deep wells before and after the Tangshan earthquake of 1976, earthquake Pre4diction. In *Proceedings of the International Symposium on Earthquake Prediction*. Unesco, pp. 215–232.
Warwick, J. W., Stoker, C., and Meyer, T. R., 1982. Radio emission associated with rock fracture: possible application to the great Chilean earthquake of May 22, 1960. *Journal of Geophysical Research*, **87**, 2851–2859.
Wiemer, S., and Wyss, M., 1994. Seismic quiescence before the Landers (M = 7.5) and Big bear (M = 6.5) earthquakes. *Bulletin of the Seismological Society of America*, **84**(3), 900–916.
Yamada, I., Masuda, K., and Mizutani, H., 1989. Electromagnetic and acoustic emission associated with rock fracture. *Physics of the Earth and Planetary Intriors*, **57**, 157–168.
Yamanaka, Y., and Kikuchi, M., 2004. Asperity map along the subduction zone in northeastern Japan inferred from regional seismic data. *Journal of Geophysical Research*, **109**, B07307, doi:10, 1029/2003JB002683.
Yasuoka, Y., Kawada, Y., Nagahama, H., Omori, Y., Ishikawa, T., Tokonami, S., and Shinogi, M., 2009. Preseismic changes in atmospheric radon concentration and crustal strain. *Physics and Chemistry of the Earth*, **34**, 431–434.
Yokoi, S., Ikeya, M., Yagi, T., and Nagai, K., 2003. Mouse circadian rhythm before the Kobe earthquake in 1995. *Bioelectromagnetics*, **24**(4), 289–291.
Yoshida, S. M., Uyeshima, M., and Nakatani, M., 1997, Electric potential changes associated with slip failure of granite: Preseismic and coseismic signals., *Journal of Geophysical Research*, **102**, 14,883–14,897.

Cross-references

Characteristic Earthquakes and Seismic Gaps
Earthquake Prediction, M8 Algorithm
Earthquake Sounds
Earthquake, Foreshocks
Earthquakes and Crustal Deformation
Gravity Anomalies, Interpretation
Great Earthquakes
Seismic Quiescence and Activation
Seismic Signals in Well Observations: Pre, Co, Post
Seismicity, Intraplate
Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting
Slow Earthquake
Statistical Seismology

EARTHQUAKE PREDICTION, M8 ALGORITHM

Alik Ismail-Zadeh[1,2,3], Vladimir Kossobokov[2,3]
[1]Geophysical Institute, Karlsruhe Institute of Technology, Karlsruhe, Germany
[2]International Institute of Earthquake Prediction Theory and Mathematical Geophysics, Russian Academy of Sciences, Moscow, Russia
[3]Institut de Physique du Globe, Paris, France

Definition

An *earthquake* is a sudden movement within the Earth's crust or the upper mantle that releases tectonic stress and, usually, originates seismic waves. The hierarchy of movable lithospheric volumes composes a large nonlinear dynamical system. Prediction of such a system in a sense of extrapolation of trajectory into the future is futile. However, upon coarse-graining the integral empirical regularities emerge opening possibilities of earthquake prediction (Keilis-Borok et al., 2001).

Earthquake prediction is a statement about future earthquake occurrence based on the information, data, and scientific methods that are available now. To predict an earthquake, someone "must specify the expected magnitude range, the geographical area within which it will occur, and the time interval within which it will happen with sufficient precision so that the ultimate success or failure of the prediction can readily be judged. Only by careful recording and analysis of failures as well as successes can the eventual success of the total effort be evaluated and future directions charted. Moreover, scientists should also assign a confidence level to each prediction" (Allen et al., 1976).

Introduction

According to the definition of earthquake prediction one can classify an earthquake prediction of certain magnitude range by duration of time interval and/or by territorial specificity. Four major stages can be distinguished in temporal earthquake predictions: (1) long-term, (2) intermediate-term, (3) short-term, and (4) immediate (see Table 1).

Long-term prediction is essentially based on the determination of probabilities of active fault segments to rupture for the next few decades (e.g., Working group on California earthquake probabilities, 1999). This kind of prediction can guide engineering and emergency planning measures to mitigate the impact of a target earthquake.

An intermediate-term prediction for the next few years is an update of the long-term prediction by some indicators (e.g., an increase in background seismicity, clustering of events in space and time, transformation of magnitude

distribution, and some others). The M8 algorithm is designed for an intermediate-term prediction of earthquakes (Keilis-Borok and Kossobokov, 1990).

In the 1980s, P. Varotsos, K. Alexopoulos and K. Nomicos proposed the VAN method (named after the researchers' initials) of short-term earthquake prediction for the next few months/weeks, which is based on detection of characteristic changes in the geoelectric potential via a telemetric network of conductive metal rods inserted in the ground (Lighthill, 1996).

Immediate earthquake prediction for the next few hours is sometimes mixed with "early warning," which is usually based on the first arrival of seismic waves and transmission of an electronic alert within a lead-time of seconds. Early warning is used (e.g., in Japan) to shut down gas and electricity grids and to stop high-speed trains in the event of a strong earthquake.

Following common perception, some investigators concentrate their efforts on predicting the "exact" fault segment to rupture, e.g., the Parkfield earthquake prediction experiment (Bakun and Lindh, 1985), which is by far more difficult and might be an unsolvable problem (Main, 1999). Being related to the rupture size L of the incipient earthquake, predictions could be summarized in a classification that distinguishes several options of wider spatial ranges of certainty and not only "exact" mode of prediction (Table 1). The Gutenberg–Richter power law relation between the number of earthquakes and their magnitude suggests limiting magnitude range of prediction, $Mm+ = [m, m + \Delta m]$, to about one unit or less (i.e., $\Delta m \leq 1$).

From a viewpoint of such a classification, the earthquake prediction problem might be approached by a hierarchical, step-by-step prediction technique, which accounts for multi-scale escalation of seismic activity to the main rupture (e.g., Keilis-Borok, 1990; Kossobokov et al., 1999). It starts with term-less recognition of the earthquake-prone zones (Gorshkov et al., 2003) and then follows with determination of the areas of different spatial ranges, where long- and intermediate-term *times of increased probability* (TIP) of a target earthquake occurrence are estimated. Finally, a forecast may come out with short- or immediate-term alerts at specific sites.

Earthquake Prediction, M8 Algorithm, Table 1 Classification of earthquake predictions

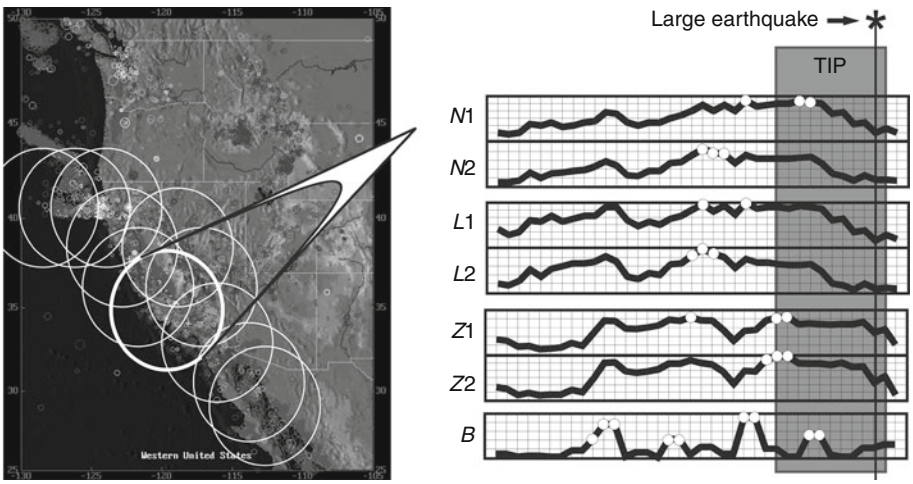
Temporal, in years		Spatial, in source zone size L	
Long-term	10	Long-range	Up to 100 L
Intermediate-term	1	Middle-range	5–10 L
Short-term	0.01–0.1	Narrow	2–3 L
Immediate	0.001	Exact	1 L

Spatial accuracy of predictions depends on the source size L of target earthquakes.

The M8 algorithm

The intermediate-term middle-range earthquake prediction method, called briefly the *M8 algorithm*, was designed by retroactive analysis of dynamics in seismicity preceding the great, magnitude 8.0+, earthquakes worldwide, hence its name. The exact definitions and computer code of the M8 algorithm can be found in Keilis-Borok and Kossobokov (1990), Healy et al. (1992), and Kossobokov (1997).

Prediction is aimed at the earthquakes of magnitude M, where $M \in [M_0, M_0 + \Delta m] = M_0+$. The territory of the seismic region under study is analyzed by moving overlapping circles of the fixed size (by a factor of about 5–10 larger than a source size of target magnitude M_0 events) (Figure 1). The sequence of earthquakes with



Earthquake Prediction, M8 Algorithm, Figure 1 General scheme of the M8 algorithm. *Left panel*: Circles of investigation overlay the targeted seismic region; seismic sequences in each circle give a description of seismic activity, which is then used to diagnose an alert (TIP). *Right panel* presents the plots of the M8 algorithm's functions (explained below) with time. The interval of TIP is shaded by gray; the asterisk denotes the large event occurred; and the white dots show anomalously high values of the functions.

aftershocks removed is considered within each circle. Sequences in different circles are normalized to the annual number of earthquake $\bar{N}$ by selecting the lower magnitude cutoff $\underline{M} = M_{\min}(\bar{N})$.

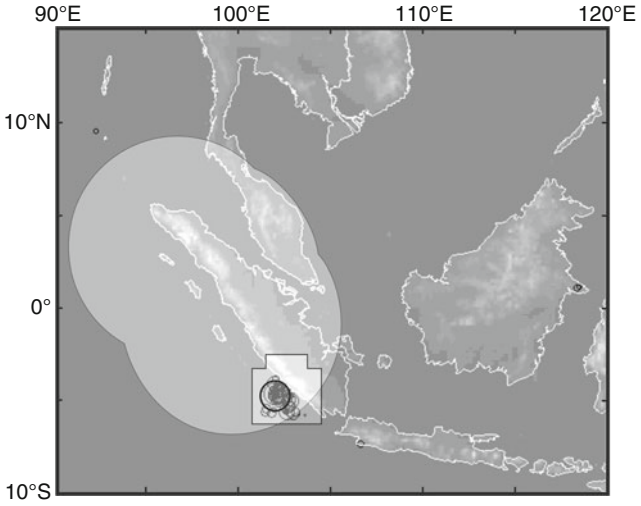
For this sequence, several functions are computed in the trailing time window (*t-s, t*) and the magnitude range ($\underline{M} \leq$ *magnitude* $< M_0$). These functions include (1) the number of earthquakes *N(t)* of magnitude *M* or greater; (2) the deviation of *N(t)* from longer-term trend *L*(*t*); (3) concentration *Z*(*t*) estimated as the ratio of the average source diameter to the average distance between sources; and (4) the maximum number of aftershocks *B*(*t*). Each of the functions *N*, *L*, and *Z* is calculated twice with $\underline{M} = M_{\min}(\bar{N})$ for $\bar{N} = 20$ and $\bar{N} = 10$. As a result, the earthquake sequence is given a robust averaged description by seven functions *N*1, *N*2, *L*1, *L*2, *Z*1, *Z*2, and *B* (Figure 1). An alarm or a TIP is declared for 5 years when at least six out of seven functions, including *B*, become "anomalously large" within a time window of 3 years.

Retrospectively the M8 algorithm was applied to predict earthquakes from different magnitude ranges in a number of regions worldwide. Its modified versions targeting earthquakes from magnitude ranges down to M5.5+ were found efficient in regions of seismic activity lower than required by the original version (Kossobokov and Shebalin, 2003; and references therein).

The MSc algorithm

The second approximation prediction method, called briefly the *MSc algorithm* (Kossobokov et al., 1990), was designed by retroactive analysis of the detailed regional seismic catalog prior to the Eureka earthquake (1980, M = 7.2) near Cape Mendocino in California, hence its name, *Mendocino Scenario*, and an abbreviation. The MSc algorithm is applied to reduce the area of alarm by analyzing dynamics at lower magnitude levels of seismic hierarchy. Sometimes, the data is enough to get a near-perfect outline of the incipient large earthquake. More often the catalog of earthquakes is just enough for the M8 analysis but insufficient for the MSc application.

Qualitatively, the MSc algorithm determines an area of the territory of alarm, where the activity recognized by the first approximation prediction algorithm (e.g., by the M8 algorithm) within a sufficient temporal and/or spatial span is continuously high and infrequently drops for a short time. The phenomenon, which is used in the MSc algorithm, might reflect the second (possibly, shorter-term and, definitely, narrow-range) stage of the premonitory rise of seismic activity near the incipient source of the main shock. The MSc algorithm reduces area of the M8 predictions much better (by at least a factor of 2) than a few simple alternatives like all earthquake-prone or the most active cells that contain certain part of the recent seismic activity (Kossobokov et al., 1990). Figure 2 shows, as an example of the M8-MSc forecast in the real time, the case history of June 4, 2000, M8.0 Sumatra earthquake.



Earthquake Prediction, M8 Algorithm, Figure 2 Prediction of June 4, 2000, M = 8.0 Sumatra earthquake. The highlighted overlapping circular areas of alarm determined by the M8 algorithm and the highlighted overlapping rectangular areas of alarm by the MSc algorithm. The main event and its aftershocks are marked by circles.

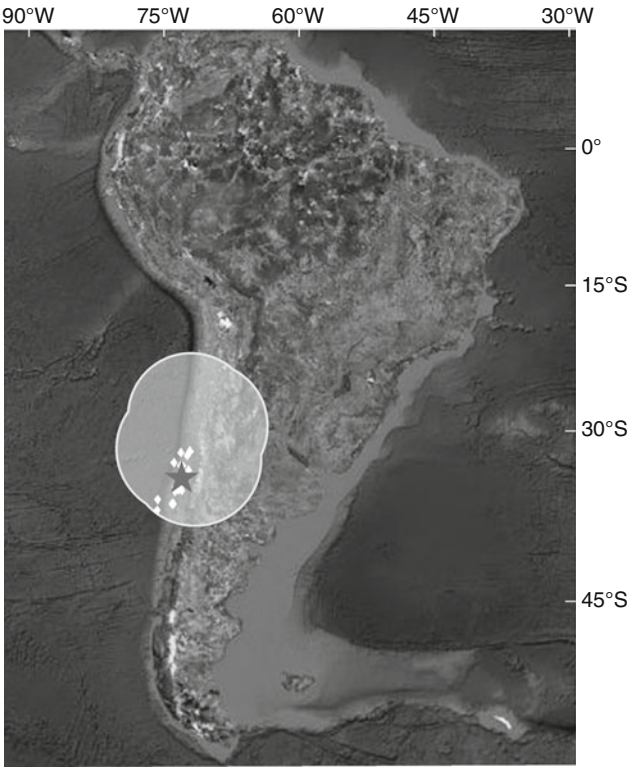
The earthquake prediction algorithms M8 and MSc receive a fair amount of attention due to ongoing real-time experimental testing unprecedented in rigor and global coverage (Healy et al., 1992; Kossobokov et al., 1999). Relevant predictions are communicated, with due discretion, to several dozens of leading scientists and administrators in many countries. The accumulated statistics of this experiment confirm intermediate-term predictability of large earthquakes with middle- to exact-range of location (Kossobokov, 2006). Table 2 summaries the prediction outcomes for the great (magnitude 8.0+) earthquakes. For the last 25 years, 13 out of 18 great earthquakes were predicted by the M8 algorithm, while the M8-MSc algorithm narrows the area of alarm at the cost of additional three failures-to-predict. The ratio of the *space-time volume alerted by the algorithms* (i.e., alarms) to the total space-time volume is about 33% in the case of the M8 algorithm and 17% in the case of the M8-MSc algorithm. A confidence level estimate tells how sure one can be that the achieved performance of the algorithm is not arisen by chance. To bring the confidence level down to generally accepted value of 95%, the ongoing experimental testing has to encounter eight failures-to-predict in a row, which appears absolutely unlikely to happen.

Targeting mega-earthquakes

The December 26, 2004, M = 9.2 Sumatra-Andaman (S-A), March 28, 2005, M = 8.5 Nias, and February 27, 2010, M = 8.8 Chile mega-earthquakes, which ruptured the 1300-, 400-, and 600-km portions of subducting margins of the Indian and Pacific Oceans, are not included

Earthquake Prediction, M8 Algorithm, Table 2 Worldwide performance of earthquake prediction algorithms M8 and M8-MSc

	Large earthquakes						
	Predicted by			Alarms, %		Confidence level, %	
Test period	M8	M8-MSc	Total	M8	M8-MSc	M8	M8-MSc
1985–2009	13	10	18	32.93	16.78	99.93	99.98



Earthquake Prediction, M8 Algorithm, Figure 3 Prediction of February 27, 2010, offshore Maule, Chile M_W8.8 mega-earthquake. The epicenters of the main shock (star) and its aftershocks (white diamonds) fill the 600-km portion of the South American subduction zone, which is about one half of the major seismic segment recognized (gray outline) as capable of producing a magnitude M8.0+ event before mid-2012 in the ongoing real-time experimental testing.

in Table 2 because of their rupture size. According to the M8 methodology, much larger circles of investigation (CI) with diameter of several thousand kilometers should then be employed to search for a precursory seismic activation in advance of mega-earthquakes. An application of the M8 algorithm targeting M9.0+ earthquakes could have predicted the S-A earthquake (see http://www.mitp.ru, section "Predictions").

Actually, the retrospective application of the M8 algorithm (with the 6,000-km diameter CI) showed that in July 2004 the precursory seismic activation was widespread over the most of the earthquake-prone areas of the Earth (Romashkova, 2008). This activation reduced to four clusters of TIPs by 2008. The largest cluster includes S-A and adjacent major seismic zones from Kashmir in the North-West to Timor Islands in the East, where March 28, 2005, Nias, M_W8.7 M-earthquake and coupled September 12, 2007, Southern Sumatra, M_W8.5 and M_W7.9 great earthquakes have already confirmed the M8 algorithm diagnosis (see http://www.mitp.ru; *note that an access to the current predictions is available only for experts; contact volodya@mitp.ru for the access code*). One of the other three smaller clusters of TIPs (i.e., the circle centered at 31°S and 70°W) got the confirmation with the occurrence of February 27, 2010, offshore Maule, Chile M_W8.8 M-earthquake (Figure 3).

Figure 3 does not show this CI, but the cluster of TIPs in four smaller (667-km diameter) CIs of the regular M8 algorithm diagnosis aimed at M8.0+. Evidently, this provides a narrow-range reduction of the prediction uncertainty acting as the second approximation to diagnosis aimed at M9.0+.

Conclusion

High statistical significance of intermediate-term earthquake prediction methods, achieved in the ongoing experimental testing worldwide, confirms the following paradigms: (1) precursory seismicity patterns exist; (2) the dimension of an area where precursory seismicity pattern appear is by far (about ten times) larger than that of source zone of the incipient large earthquake; (3) many precursory seismicity patterns are similar in regions of different seismic and tectonic environment; and (4) some precursory seismicity patterns are universal. Their analogs take place in advance extreme catastrophic events in other complex nonlinear systems.

Our knowledge of earthquake physics and earthquake dynamics is not sufficient to predict strong earthquakes with a high accuracy. We do not know well (1) how earthquakes, especially large events, originate; (2) when an earthquake starts, when it stops, and what magnitude could it be; (3) how and why earthquakes cluster; and (4) what were the initial conditions of stress state before a large event in terms of stress transfer.

The scientific community should use the full potential of mathematics, statistics, statistical physics, and computational modeling and the data derived from seismological

(monitoring of physical parameters of earthquakes and tectonic stress, fluid migration, etc.), geodetic (GPS, InSAR, and other measurements of the crustal deformation), and geological (e.g., determination of the time intervals between large earthquakes using paleo-seismological tools) studies to improve intermediate- and short-term earthquake predictions (Ismail-Zadeh, 2010).

Though the current accuracy of earthquake prediction is limited, any scientifically validated prediction can be useful for earthquake preparedness and disaster management, if the accuracy of the prediction is known, even though it is not high. An inexpensive low-key response to intermediate-term middle-range predictions described here is well justified, if even a little part of the total damage due to a strong event is prevented.

Bibliography

Allen, C. R., Edwards, W., Hall, W. J., Knopoff, L., Raleigh, C. B., Savit, C. H., Toksoz, M. N., and Turner, R. H., 1976. *Predicting Earthquakes: A Scientific and Technical Evaluation – With Implications for Society*. Panel on Earthquake Prediction of the Committee on Seismology, Assembly of Mathematical and Physical Sciences, National Research Council. Washington: U.S. National Academy of Sciences.

Bakun, W. H., and Lindh, A. G., 1985. The Parkfield, California, earthquake prediction experiment. *Science*, **229**, 619–624.

Gorshkov, A. I., Kossobokov, V., and Soloviev, A., 2003. Recognition of earthquake-prone areas. In Keilis-Borok, V. I., and Soloviev, A. A. (eds.), *Nonlinear Dynamics of the Lithosphere and Earthquake Prediction*. Heidelberg: Springer, pp. 239–310.

Healy, J. H., Kossobokov, V. G., and Dewey, J. W., 1992. A test to evaluate the earthquake prediction algorithm, M8. *U.S. Geol. Surv. Open-File Report*, 92–401, 23 p. with 6 Appendices.

Ismail-Zadeh, A., 2010. Computational geodynamics as a component of comprehensive seismic hazards analysis. In Beer, T. (ed.), *Geophysical Hazards: Minimizing Risk and Maximizing Awareness*. Amsterdam: Springer, pp. 161–178.

Keilis-Borok, V. I., 1990. The lithosphere of the Earth as a nonlinear system with implications for earthquake prediction. *Reviews of Geophysics*, **28**, 19–34.

Keilis-Borok, V. I., and Kossobokov, V. G., 1990. Premonitory activation of earthquake flow: algorithm M8. *Physics of the Earth and Planetary Interiors*, **61**, 73–83.

Keilis-Borok, V. I., Ismail-Zadeh, A. T., Kossobokov, V. G., and Shebalin, P. N., 2001. Non-linear dynamics of the lithosphere and intermediate-term earthquake prediction. *Tectonophysics*, **338**(3–4), 247–259.

Kossobokov, V., 1997. User Manual for M8. In Healy, J. H., Keilis-Borok, V. I., and Lee, W. H. K. (eds.), *Algorithms for Earthquake Statistics and Prediction. IASPEI Software Library*. El Cerrito: Seismological Society of America, Vol. 6, pp. 1–167.

Kossobokov, V., 2006. Quantitative earthquake prediction on global and regional scales. In Ismail-Zadeh, A. (ed.), *Recent Geodynamics, Georisk and Sustainable Development in the Black Sea to Caspian Sea Region. Conference Proceedings*. New York: American Institute of Physics, Vol. 825, pp. 32–50.

Kossobokov, V., and Shebalin, P., 2003. Earthquake prediction. In Keilis-Borok, V. I., and Soloviev, A. A. (eds.), *Nonlinear Dynamics of the Lithosphere and Earthquake Prediction*. Heidelberg: Springer, pp. 141–207.

Kossobokov, V. G., Keilis-Borok, V. I., and Smith, S. W., 1990. Localization of intermediate-term earthquake prediction. *Journal of Geophysical Research*, **95**(B12), 19763–19772.

Kossobokov, V. G., Maeda, K., and Uyeda, S., 1999. Precursory activation of seismicity in advance of Kobe, 1995, M = 7.2 earthquake. *Pure and Applied Geophysics*, **155**, 409–423.

Lighthill, J. (ed.), 1996. *A Critical Review of VAN: Earthquake Prediction from Seismic Electric Signals*. Singapore: World Scientific Publication.

Main, I., 1999. Is the reliable prediction of individual earthquakes a realistic scientific goal? Nature debates at http://www.nature.com/nature/debates/earthquake. Accessed 18 June 2010.

Romashkova, L. L., 2008. Global-scale analysis of seismic activity prior to 2004 Sumatra-Andaman mega-earthquake. *Tectonophysics*, **470**, 329–344.

Working group on California earthquake probabilities, 1999. *Earthquake probabilities in the San Francisco Bay region: 2000–2030 – A summary of findings*. U.S. Geological Survey Open-file Report 99–517.

Cross-references

Earthquake, Aftershocks
Earthquake, Foreshocks
Earthquake, Magnitude
Earthquake Precursors and Prediction
Earthquakes, Early and Strong Motion Warning
Earthquakes, Strong-Ground Motion
Great Earthquakes
Seismic Hazard

EARTHQUAKE RUPTURE: INVERSE PROBLEM

Shamita Das
Department of Earth Sciences, University of Oxford, Oxford, UK

In seismology, as in all geophysics, an area of major interest is inverse problems. That is, given the effect, one would like to determine the cause. The inverse problem of earthquake source mechanics consists of analyzing seismograms to obtain detailed information on the earthquake rupturing process. In particular, scientists are interested in how the slip develops over the fault area during the time period that the two sides of the fault are moving past one another, at what speed do these two sides slide past, what the resulting stress drop is, as well as the speed at which rupture itself spreads out over the fault. Such information is essential for reliable seismic hazard assessment.

The solution of this problem is far from trivial, because it is unstable (Kostrov and Das, 1988), and from the computational point of view, this instability manifests itself as non-uniqueness of the solution. Adding constraints to the solution of the relevant equations has been shown to stabilize the problem. Such constraints appear best when based on the physics of problem, that is, the physics of the earthquake faulting process. Unfortunately, our knowledge of the physics of the earthquake process is still rather limited. So, seismologists have had to use those few properties that are considered to be reliably known. For example, forward modeling showed that it is not likely that there is major

back-slip on the fault. So, a constraint that is often used is the requirement that the slip rate vector be positive, that is, no back-slip is permitted on the fault. While, in principle, slip in the opposite direction is possible due to interference of waves on the fault, especially at the fault edges, experience from three-dimensional forward modeling shows that this is unlikely. This constraint has been used by Olson and Apsel (1982), by Hartzell and Heaton (1983), and by Das and Kostrov (1990, 1994), among others, and has been shown to stabilize the solution. Other possible constraints, such as requiring the solution to be in agreement with the seismic moment obtained from the centroid moment tensor solution or from geodetic measurements, or, requiring the fault rupture speed to be limited by some wave velocity, can also been used. Further constraints such as requiring the maximum slip rate on the fault to be limited by some considerations from fracture mechanics, could be considered. It has been found that minimizing the seismic moment often provides stable solutions.

As stated by Olson and Apsel (1982), in order for a particular rupture process to be an acceptable solution to the inverse problem it must satisfy the following three conditions:

1. The solution must explain the data.
2. The solution must be physically reasonable (consistent with independent constraints).
3. If more than one solution fits the data equally well, additional information must be supplied to uniquely define which solution is being obtained.

The rupture process is solved for by determining the slip rate distribution over the fault with time, and the entire rupture process is reconstructed from it. For example, the final slip distribution is found simply by integrating the slip rate distribution over time. The rupture speed can be found simply by tracking the rupture front, and so on.

To invert for the three-dimensional slip rate distribution (two spatial dimensions on the fault and time) is a computationally difficult task. Not surprisingly, simplified methods of solving the problem have been used. For example, in order to reduce the number of unknowns, one often restricts for how long a point on the fault can slip. But, in reality, if some region of the fault re-slips, or has delayed slip, both shown to be feasible in forward models, this may not be found by the solution. This restriction would inhibit discovering fault complexity. The discretization itself may implicitly impose some constraint on the solution. It is thus important to ensure that the solution methods do not impose any implicit (and unknown) constraints on the solution, and if implicit constraints are unavoidable, they must be kept in mind when interpreting such solutions in geophysical terms.

To formulate the problem, we start from the representation theorem (e.g., Aki and Richards, 1980), where the ground displacement at a station located at a point $\mathbf{x}$, on the Earth's surface can be expressed in terms of the slip distribution over a fault Σ as

$$u_k(\mathbf{x}_1, t_1) = \int_0^{t_1} dt \iint_\Sigma K_{ik}(\mathbf{x}_1, \mathbf{x}, t_1, t) a_i(\mathbf{x}, t) dS \qquad (1)$$

where $i, k = 1, 2, 3$, $u_k(\mathbf{x}_1, t_1)$ are the components of the displacement vector, $a_i(\mathbf{x}, t)$ are the components of the slip and $K_{ik}(\mathbf{x}_1, \mathbf{x}, t_1, t)$ are the components of the impulse response of the medium at $(\mathbf{x}_1, t_1)$, due to a dislocation point source at $(\mathbf{x}, t)$. Assuming a planar fault and the slip direction on the fault to be constant, we can write, after some transformations:

$$S_j(t_1) = \int_0^{t_1} dt \iint_\Sigma W_j(\xi, t_1 - t) \dot{a}(\xi, t) dS \qquad (2)$$

where j identifies the station and ground displacement components of the seismogram $S(t_1)$, $W_j(\xi, t)$ is the Heaviside response of the medium at (ξ, t) corresponding to a fixed slip direction, $\dot{a}(\xi, t)$ is the fault slip rate, the two-dimensional vector ξ gives the position on the fault relative to some reference point (for example, the earthquake hypocenter) and t is the time measured from the origin time of the earthquake. With a continuous distribution of stations, Equation 2 would represent an integral equation of the first kind (and such equations are known to be unstable).

The following additional constraints are often used:

$$\dot{a}(\xi, t) \geq 0 \text{ for all} (\xi, t) \qquad (3a)$$

$$\dot{a}(\xi, t) = 0 \quad \text{for} \quad t < T(\xi) \qquad (3b)$$

where, $t = T(\xi)$ gives the boundary of the area where slip is permitted (due to some chosen causality condition) at time t. Note that this is not the actual rupture front, which should not generally by assigned a priori, and should be determined as part of the solution.

$$\int_0^\infty dt \iint_\Sigma \mu(\xi) \dot{a}(\xi, t) dS = M_0 \qquad (3c)$$

where M_0 is the seismic moment and μ is the modulus of rigidity of the medium. The constraint (3*a*) is the "no-back-slip" or the "positivity" constraint, the constraint (3*b*) is a "causality constraint," which could be taken as the *P*- or *S*-wave causal front, or any other chosen region. The constraint (3*c*) is the "seismic moment constraint."

We restrict this discussion to long-period body waves at teleseismic distances ("far-field") because they are less sensitive to the details of Earth structure. This problem has been looked at using ground motion data from seismograms located at regional distances from earthquakes, for example, for the 1985 Mexican earthquake (Mendoza and Hartzell, 1989) and the 1992 Landers, California earthquake (Cohee and Beroza, 1994; Wald and Heaton, 1994), but the problem has not yet been considered for the case of true "near-field" data where the higher order terms are kept in the Green function calculations. The (Green) functions W_j for the teleseismic problem can be calculated using, for example, the ray

approximation and the Thomson–Haskell technique or ray-tracing methods to account for the layered crustal structure at the source and the station. These techniques take into account all surface reflections near the source and at the receiver, such as *pP*, *sP*, *pS*, and *sS*, as well as all reflections and transmissions at every layer of the crustal structures. For the numerical solution, the integrals in (2) must be discretized. For this, the area Σ where slip is permitted must be divided into a number of cells (usually square or rectangular) and $\dot{a}$ within each cell approximated by linear functions in time and along strike and dip (if the width of the cells are small, the variation along dip can be taken as constant). The W_j's must then be integrated over each cell (analytically, if possible, as done by Das and Kostrov (1990, 1994)), and the integrals over the fault are replaced by sums. The time at the source is discretized by taking a fixed time step, Δt say, and it could be assumed that the slip rate $\dot{a}$ during the time step is a constant, or varies in some prescribed way, say, linearly with time. Both the fault area and the total source duration should be determined as part of the inversion process. One can use the aftershock area as starting area, and then refine it further during the inversion. The initial source duration cannot be longer than the longest record used, in other words, the time length of the seismograms needs to be long enough to study the entire source process. It is important to keep in mind that at many distances, other phases, such as core reflections, arrive within this time window, and either the seismogram used has to be terminated at this time, or these additional phases correctly modeled. Unfortunately, often, neither is done in some studies, based on the incorrect assumption that such reflections are always small compared to the direct arrivals. The seismograms are sampled at some chosen time step Δt_1, not inconsistent with the spatial cell size of the fault and time step of the source duration. Then Equation 2 takes the form

$$A\mathbf{x} \approx \mathbf{b} \tag{4}$$

where A is the matrix obtained by integration of W_j, each column of A being a set of Green function seismograms for all stations corresponding to different cells and time instants of the source duration, ordered in the same way as the observed seismograms. The arrangement of Equation 4 is shown in Figure 1. Different stations and components can be weighted differently in Equation 2. This is very useful when inverting P- and S-waves simultaneously.

For the discrete problem, condition (3*a*) becomes:

$$\mathbf{x} \geq 0, \tag{5a}$$

the inequality meaning that every component of $\mathbf{x}$ is non-negative. The condition (3b) can be replaced by

$$x_i = 0 \tag{5b}$$

for those i corresponding to cells and time samples outside the "rupture front" and the condition (3*c*) becomes:

$$\Sigma c_i x_i \approx M_0 \tag{5c}$$

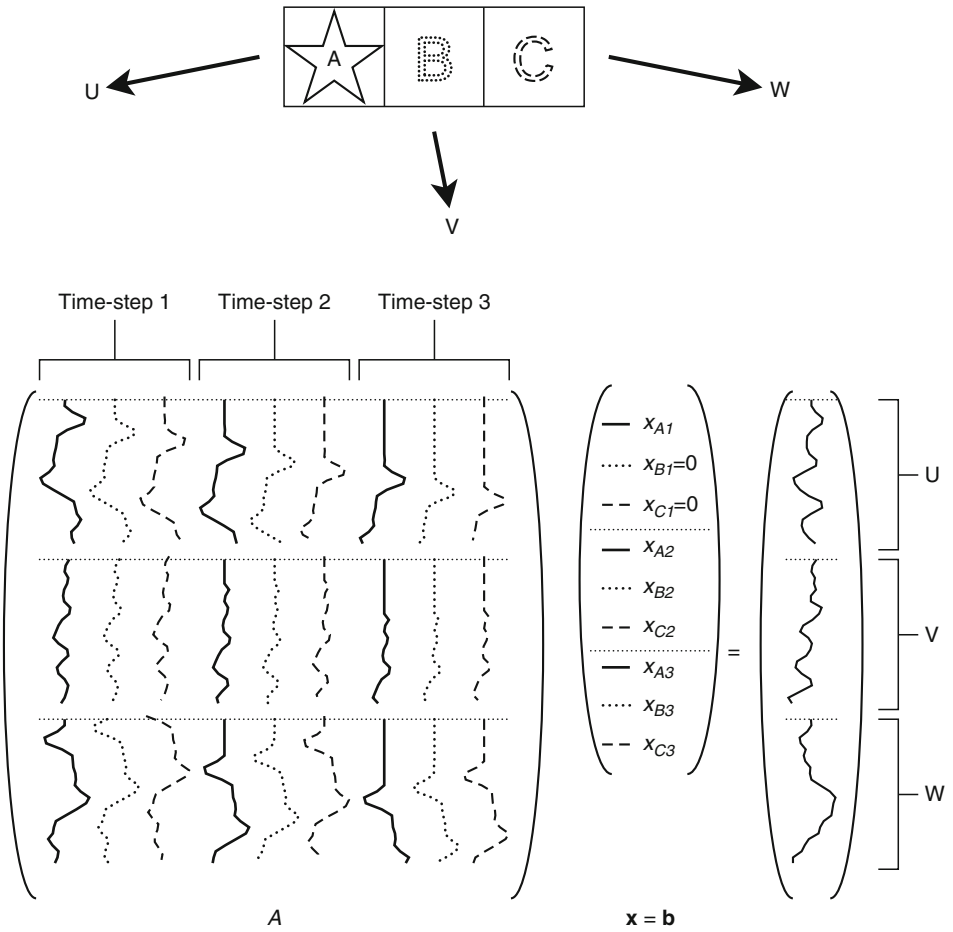
where c_i is time-independent and for each cell is equal to the product of the average rigidity multiplied by the area of the cell. So, the inverse problem has been reduced to the solution of the linear system (4) under one or more of the constraints (5). In the system (4), the number of equations, m, is equal to the total number of samples taken from all the records involved and the number of unknowns, n, is equal to the number of cells multiplied by the number of time steps at the source. We shall take $m > n$ to reduce the influence of the noise contained in the observations $\mathbf{b}$ on the solution. Then, the system (4) is overdetermined and we can only obtain a solution $\mathbf{x}$ that provides a best fit to the observations, under constraints (5). It is important to note that, in theory, for the constrained system, it is not necessary for $m > n$, though it has been so in all actual studies of earthquakes.

It is now well understood that the matrix A can be ill-conditioned, which implies that the system (4) admits more than one solution, equally well fitting the observations. The constraints (5) are introduced just for the purpose of reducing the set of permissible (feasible) solutions. When a "best-fitting" solution is obtained, other solutions almost satisfying the equations are also of great interest. This is because the data used in geophysical applications often contains noise and the models used are themselves approximations to reality, so that the "best-fitting" solution may not represent actuality. In addition, though the different solutions may be very close together in data-space, it is possible that they may be somewhat different in model space. Thus, more than one solution must be found and the set of (almost) equally well-fitting solutions examined to find features common to the solutions. To examine the stability of the inverse problem, Das and Kostrov (1990, 1994) obtained not only a "best-fitting" solution but also investigated the solution space around it. These other solutions, though they may lie very close to one another in model space, may physically represent rather different solutions. Das and Kostrov (1990, 1994) suggested additional tests to determine the robustness of such features. Only features clearly required by the solutions should be interpreted in a geophysical context.

For the system of Equations 4 together with the constraints (5) to comprise a complete mathematical problem, it is necessary to formulate the exact form of what the "best-fit" to observations means. We include only constraint (5*a*) in the mathematical formulation for the sake of simplicity, the inclusion (5*b*) and (5*c*) being simple, and is discussed, for example, by Das and Kostrov (1990, 1994). We have to minimize the vector of residuals:

$$\mathbf{r} = \mathbf{b} - A\mathbf{x} \tag{6}$$

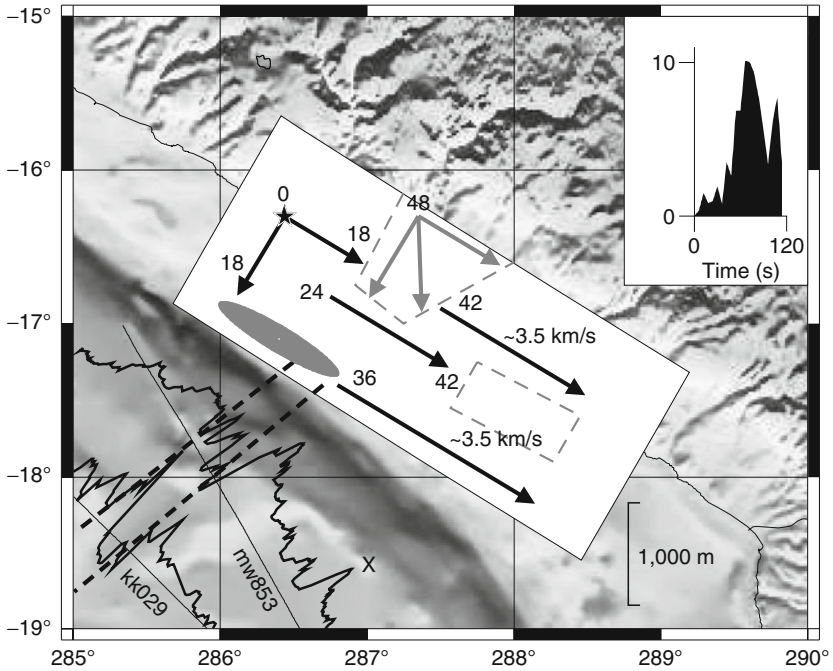
For this purpose, some norm of the vector $\mathbf{r}$ must be adopted. One may choose to minimize the ℓ_1, the ℓ_2, or the ℓ_∞ norm, all three being equivalent in the sense that they tend to zero simultaneously. Hartzell and Heaton



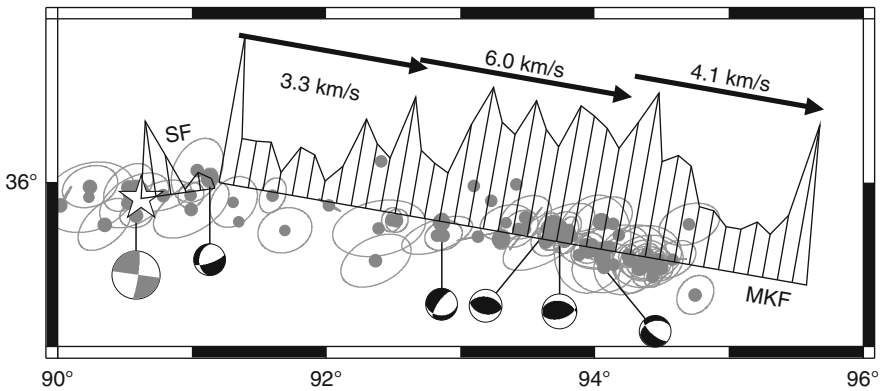
Earthquake Rupture: Inverse Problem, Figure 1 Schematic showing the arrangement of the linear system $A\mathbf{x} = \mathbf{b}$.
A simplified faulting model is shown at the top, with three spatial cells along the fault length and one cell along its width. The rupture initiates at time step 1, in cell A, marked by the star. Rupture proceeds to the left, breaking cell B and then cell C. The resulting seismograms are recorded at three stations named U, V, and W. In the matrix A, containing the Green functions, portions relating to the cells A, B, and C are coded with solid, dotted, and dashed lines, respectively. The vector $\mathbf{x}$ is the slip rate in each cell A, B, or C, at time steps 1, 2, and 3. In reality, there will be many more time steps, and A will have many more columns. If only cell A is allowed to slip at the first time step, by invoking the causality condition, then the unknown slip rates $\mathbf{x}_{B1}$ and $\mathbf{x}_{C1}$ will vanish. The vector of seismograms $\mathbf{b}$ simply consists of the seismograms at the stations, placed one after another, *but having the appropriate sample length* consistent with the l.h.s. of the equation.

(1983) used the ℓ_2 norm to solve the optimization problem with the positivity constraint, employing computer programs in Lawson and Hanson (1974). Das and Kostrov (1990, 1994) have used the linear programming method (Press et al., 1986) to solve the system (4) and minimize the ℓ_1 norm subject to the condition (5*a*). The ℓ_1 norm is considered more robust (Tarantola, 1987), and Hartzell et al. (1991) have investigated the effect of minimizing different norms on the solution, and shown that the most robust features are properly obtained in the solutions, whichever norm is used. Das and Kostrov (1990, 1994) have checked that when the ℓ_1 norm is minimized, the ℓ_2 and the ℓ_∞ norms for the solution are also small.

In reality, the actual scale at which the rupture propagation takes place in the Earth is microscopic, whereas the inverse problem must be solved using finite grids. Therefore, it is important to consider how well the inverse solution obtained gives a picture of the actual solution. To do this, it is necessary to decide how large the cells in the inversion should be. Clearly, the smaller the cells, the better the approximation to the integral equation being solved (Equation 1), but the larger the number of unknowns. On the other hand, though the approximation to the integral equation improves by going to smaller and smaller cells, the condition number of the matrix increases (due to columns becoming very similar) making the problem more and more unstable. As pointed out by Backus and Gilbert (1967), since the amount of data is finite, the problem of finding a continuous function from it is indeterminate. In other words, the stability of the system of equations



Earthquake Rupture: Inverse Problem, Figure 2 Tectonic setting of the June 23, 2001 great Peru earthquake (M_w8.4), together with bathymetry constructed from a GEBCO (2003) 1′ × 1′ grid. The rupture zone of the earthquake is shown as a white rectangle and the epicenter by a star. The black and gray lines show the rupture process, the numbers indicating the seconds after rupture initiation. The gray ellipse shows the region where large slip was induced by delayed rupture of the trapezoid-shaped barrier, shown by dashed lines. A second rectangular barrier is also seen. Figure based on Robinson et al. (2006a).



Earthquake Rupture: Inverse Problem, Figure 3 The rupture process of the M_w 7.8 Kunlun, Tibet earthquake of 2001. Relocated aftershocks for the 6-month period following the earthquake are shown by dots, together with their 90% confidence error ellipses. The two faults on which rupture occurred are marked as MKF (Main Kunlun Fault) and SF (Secondary Fault), the star showing the rupture initiation region. Schematic graphs showing the final slip on the two faults are shown, with the rupture speeds in the three phases of rupture indicated at the top. The central portion is where rupture reached the compressional wave speed. Figure based on Robinson et al. (2006b).

decreases (that is, the condition number of the matrix A increases) as we use finer and finer discretization of the fault for the inversion. Das and Suhadolc (1996) and Das et al. (1996) have suggested that synthetic tests using artificially created data should be used to best estimate the spatial and temporal grid sizes. It is important to keep in mind that the spatial and temporal grid sizes should be consistent. Henry et al. (2000) have also compared the Green functions obtained using various grid sizes to aid in the grid size selection. Sarao et al. (1998) have studied the effect of nonuniform seismic station distribution around the earthquake epicenter.

Finally, we discuss some examples of what we have learned regarding the mechanics of the earthquake source from the solution of such inverse problems. The first great earthquake for which the slip history and distribution was obtained was the great M_w 8.0 Michoacan, Mexico earthquake in 1985 (Mendoza and Hartzell, 1988). It was found that the main slip took place in two patches on the fault, and that aftershocks occurred in the regions of low fault slip or regions of transition from high to low slip, that is, where stresses were increased due to the earthquake. Many studies since then are summarized by Das and Henry (2003), and from them we have learned important features of earthquake mechanics. For example, the M_w 8.1 Antarctic plate earthquake in 1998 showed that ruptures can jump over barriers as large as 70–100 km and then continue propagating significant distances. The M_w 7.8 Wharton Basin earthquake in 2000 in the Indian Ocean showed that rupture could occur simultaneously on both nodal planes.

Some very recent examples include the great M_w 8.4 Peru earthquake in 2001 (Figure 2), which showed how a barrier on the fault could stall the rupture process (Robinson et al., 2006a). Based on marine geophysical data from the portion of the Nazca plate that was subducting in this region, this barrier was identified as being due to bathymetry (such as seamounts located on the inner corner of a fracture zone) related to a subducting fracture zone. Another important example was the large M_w 7.8 Kunlun, Tibet, earthquake in 2001 (Figure 3), where it was seen that the rupture speed during this strike-slip not only exceeded the local shear wave speed but actually reached the compressional wave speed of nearly 6 km/s, and propagated for nearly 100 km at this speed. The fault geometry was shown to be a controlling factor for the changes in rupture speed along the nearly 400 km long fault (Robinson et al., 2006b). This has important implications for earthquake hazard mitigation (Das, 2007).

Summary

In the last 2 decades, seismologists have solved the "inverse problem" of using broadband digitally recorded seismograms to infer the details of the earthquake rupturing process. The theory and some examples of particular earthquakes studied are given.

Bibliography

Aki, K., and Richards, P. G., 1980. *Quantitative Seismology: Theory and Methods*. New York: W.H. Freeman.

Backus, G., and Gilbert, F., 1967. Numerical applications of a formalism for geophysical inverse problems. *Geophysical Journal of the Royal Astronomical Society*, **13**, 247–276.

Backus, G., and Gilbert, F., 1968. The resolving power of gross earth data. *Geophysical Journal of the Royal Astronomical Society*, **16**, 169–205.

Cohee, B. P., and Beroza, G. C., 1994. Slip distribution of the 1992 Landers earthquake and its implications for earthquake source mechanism. *Bulletin. Seismological Society of America*, **84**, 692–712.

Das, S., 2007. The need to study speed. *Science*, **317**(5840), 905–906.

Das, S., and Henry, C., 2003. Spatial relation between main earthquake slip and its aftershock distribution. *Reviews of Geophysics*, **41**(3), 1013, doi:10.1029/2002RG000119, 2003.

Das, S., and Kostrov, B. V., 1990. Inversion for seismic slip rate and distribution with stabilizing constraints: application to the 1986 Andreanof Islands earthquake. *Journal of Geophysical Research*, **95**, 6899–6913.

Das, S., and Kostrov, B. V., 1994. Diversity of solutions of the problem of earthquake faulting inversion. Application to SH waves for the great Macquarie Ridge earthquake. *Physics of the Earth and Planetary Interiors*, **85**, 293–318.

Das, S., and Suhadolc, P., 1996. On the inverse problem for earthquake rupture. The Haskell-type source model. *Journal of Geophysical Research*, **101**, 5725–5738.

Das, S., Suhadolc, P., and Kostrov, B. V., 1996. Realistic inversions to obtain gross properties of the earthquake faulting process, Special issue entitled Seismic Source Parameters: from Microearthquakes to Large Events, ed. C. Trifu. *Tectonophysics*, **261**, 165–177.

GEBCO, 2003. *The GEBCO digital atlas, Centenary edition (CD-ROM)*. Liverpool: British Oceanographic Data Centre.

Hartzell, S. H., and Heaton, T. H., 1983. Inversion of strong-ground motion and teleseismic waveform data for the fault rupture history of the 1979 Imperial Valley, California earthquake. *Bulletin. Seismological Society of America*, **73**, 1553–1583.

Hartzell, S., Stewart, G. S., and Mendoza, C., 1991. Comparison of L_1 and L_2 norms in a teleseismic waveform inversion for the slip history of the Loma Prieta, California, earthquake. *Bulletin of the Seismological Society of America*, **81**, 1518–1539.

Henry, C., Das, S., and Woodhouse, J. H., 2000. The great March 25, 1998 Antarctic Plate earthquake: moment tensor and rupture history. *Journal of Geophysical Research*, **105**, 16097–16119.

Kostrov, B. V., and Das, S., 1988. Principles of Earthquake Source Mechanics. *Appl. Math. Mech. Ser,* **30**, 1241–1248. New York: Cambridge University Press.

Lawson, C. L., and Hanson, R. J., 1974. *Solving Least Squares Problems*. Englewood Cliffs: Prentice-Hall.

Mendoza, C., and Hartzell, S. H., 1988. Inversion for slip distribution using teleseismic *P* waveforms: North Palm Springs, Borah Peak, and Michoacan earthquakes. *Bulletin. Seismological Society of America*, **78**, 1092–1111.

Mendoza, C., and Hartzell, S. H., 1989. Slip distribution of the 19 September 1985, Michoacan, Mexico, earthquake: near-source and teleseismic constraints. *Bulletin. Seismological Society of America*, **79**, 655–669.

Olson, A. H., and Apsel, R. J., 1982. Finite faults and inverse theory with applications to the 1979 Imperial Valley earthquake. *Bulletin. Seismological Society of America*, **72**, 1969–2001.

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T., 1986. *Numerical Recipes: The Art of Scientific Computing*. New York: Cambridge University Press.

Robinson, D. P., Brough, C., and Das, S., 2006a. The M_w 7.8 2001 Kunlunshan earthquake: extreme rupture speed variability and effect of fault geometry. *Journal of Geophysical Research*, **111**, B08303, doi:10.1029/2005JB004137.

Robinson, D. P., Das, S., and Watts, A. B., 2006b. Earthquake rupture stalled by subducting fracture zone. *Science*, **312**(5777), 1203–1205.

Sarao, A., Das, S., and Suhadolc, P., 1998. Effect of non-uniform station coverage on the inversion for seismic moment release history and distribution for a Haskell-type rupture model. *Journal of Seismology*, **2**, 1–25.

Tarantola, A., 1987. *Inverse Problem Theory. Methods for Data Fitting and Model Parameter Estimation*. New York: Elsevier.

Wald, D. J., and Heaton, T. H., 1994. Spatial and temporal distribution of slip for the 1992 Landers, California, earthquake. *Bulletin. Seismological Society of America*, **84**, 668–691.

Cross-references

Earthquake, Focal Mechanism
Earthquakes and Crustal Deformation
Earthquakes, Source Theory

EARTHQUAKE SOUNDS

Andrew J. Michael
U.S. Geological Survey, Menlo Park, CA, USA

Definition

Earthquake sounds. Atmospheric pressure waves associated with earthquakes, especially those audible to humans.

Introduction

". . . previous to an earthquake, a roaring is usually heard," wrote Lucius Annaeus Seneca, circa 65 CE, in the sixth volume of "Naturales Quaestiones" (translation by Clarke and Geike, 1910). While Seneca's explanations for earthquakes now seem fanciful, this statement remains accurate and historical records of earthquake sounds have been joined over the last few decades by instrumental studies of this phenomenon.

A particularly interesting case of earthquake sounds concerns the long reported noises near Moodus, Connecticut, whose town name is derived from a Wangunk word meaning "where noises come from the ground." Reports of these noises predate the arrival of European settlers who described them as accompanying shaking "as though in an earthquake" (Davis, 1897). Seismographic studies of the earthquakes that accompany the Moodus Noises demonstrate that these sounds are caused by swarms of magnitude −2 to 2.1 events with shallow (less than 2.4 km) hypocenters (Ebel, 1982). Similarly, explosive sounds similar to cannon fire have been documented during earthquake swarms in the lower Rhone Valley of France since 1772 and have now been explained by events with ultrashallow focal depths of only 200 m. With magnitudes as small as −0.7 many of these events are heard even if the shaking cannot be felt (Thouvenot et al., 2009).

Because human hearing is an effective spectral analyzer, the sonification of seismograms provides a useful educational tool. And because sound is commonly associated with earthquakes, audio can be an effective element of artistic works about earthquakes. This entry reviews the role of earthquake sounds in science, education, and art.

Records of sounds

People have reported hearing earthquakes for as long as we have earthquake reports. *The Earthquake Catalog of the British Association* (Mallet and Mallet, 1858) lists reports of earthquake sounds back to 122 BCE, although these are referenced to histories written by Julius Obsequens in the fourth century CE.

Catalogs of intensity data are good sources of reports on earthquake sounds. Davison (1938) explored a global set of catalogs, including the Mallet and Mallet catalog mentioned above, and the references in his paper provide a guide to these classic sources. Most of his data came from Britain and he classified about 20,000 descriptions into seven classifications that are largely consistent with low rumbling sounds but also include some descriptions of impulsive sounds, which are heard more often at short epicentral distances. A more recent intensity catalog with special emphasis on sounds was collected by Pierre Stahl in the French Pyrenees (Sylvander and Mogos, 2005), and the USGS "Did You Feel It?" catalog of Internet-collected intensity data includes information on sounds in its 1.25 million observations (Wald et al., 1999).

A problem with sounds described in intensity catalogs is that they depend on peoples' ability to describe an ephemeral experience in words. In the 1979 M4.0 earthquake near Bath, Maine, sounds were widely heard and mostly described as high frequency booms (Pulli et al., 1980). Many respondents described the sound in terms of the explosion of a heating system boiler. One exception was a resident who had actually heard a boiler explode and said the earthquake sounded nothing like that. Regardless of the words used to describe the sound, some residents who had previously lived in southern California explained that they were different than the longer-lasting, low frequency rumbles they remembered from earthquakes there. Those descriptions are consistent with Hill et al. (1976) who described earthquake sounds heard in southern California as "ranging from the rumble of distant thunder to the rushing of a wind." The high frequency content of the sounds in Maine could be due to high frequency ground motions created by a combination of shallow hypocenters, low crustal attenuation, and relatively high stress drops in that region (Pulli et al., 1980). The Moodus area shares similar characteristics and thus the shaking should also have substantial high frequency energy. However, many observers there described the noises in 1981 as "distant thunder" (Ebel, 1982). Thus, the subjective observations in intensity catalogs can make interpretation difficult.

The first known recording of earthquake sounds is from 1954 when a recording that was being made inside a two-story wood-frame dwelling in Eureka, California, accidently captured an M6.6 earthquake. It is likely that the recording, which was cut short by a power outage, largely captured sounds created by the building (Steinbrugge, 1974). The catalog of earthquake sound recordings compiled by Steinbrugge is available from the Seismological Society of America. All of these recordings mix together natural earthquake sounds with those produced by man-made structures and human reactions to the shaking. A notable entry in this catalog is a 5-min recording done by Robert Pate during the 1964 M9.2 Alaska earthquake, including his narration of several minutes of felt shaking.

The first purposeful recordings of earthquake sounds away from man-made structures were made in studies of the 1965–1967 Matsushiro swarm in Japan (Japan

Meteorological Agency, 1968) and the 1975 earthquake swarm near Brawley, California (Hill et al., 1976). Hill et al. recorded the audio on one stereo channel of an audio cassette tape and recorded both a time signal and the seismic sensor on the other channel by using FM multiplexing. Unfortunately, the seismograms recorded by this analog system were clipped and that made some of their analysis more uncertain. Sylvander et al. (2007) made digital audio and seismic recordings in the Pyrenees that stayed on scale.

How are sounds produced?

The existence of earthquake sounds is hardly surprising given that seismic waves from the solid earth will refract, albeit weakly, into the atmosphere. This was first demonstrated by Ewing, who considered the simple case of a P-wave in a body of water being transmitted into the atmosphere (Ewing, 1883).

As described by Seneca and in the observations cataloged by Mallet and Mallet (1858) and Davison (1938), earthquake sounds are often heard slightly before the shaking is felt. Because people often feel the stronger, but later arriving S-wave, this suggests that the sounds are produced by the P-wave. The instrumental studies referenced above used collocated seismometers and microphones to demonstrate that the sounds coincide with the arrival of the P-wave. The coincidence of the sounds with the arrival of the P-wave demonstrates that the onset of the sounds is generated by transmission of the seismic waves into the atmosphere very close to the listener. This is consistent with Davison's conclusion that the audible sound and the felt shaking travel at the same velocity. Hill et al. (1976) extended the simple analysis done by Ewing and explored the transmission of sound from a realistic shallow crustal structure to the atmosphere, the frequency content of the shaking from the seismic source, and the sensitivity of human hearing. That analysis showed that the higher frequencies of the seismic shaking (up to 50 Hz) overlap with the lower range of human hearing (over 20 Hz) and are transmitted into the atmosphere with sufficient amplitude to be heard.

Thus, what people hear is in the frequency range of 20–50 Hz. These low frequency sound waves have wavelengths much longer than the size of a human head and therefore people cannot localize the source of such sounds. The description of earthquake sounds as distant is based on association with other low frequency sounds that actually are distant, despite that the earthquake sounds are generated close to the listener. It is noteworthy that because earthquake sounds are at such low frequencies, humans are at least as sensitive to these sounds as many animals (Hill, 1976).

At Brawley, the later and stronger S-wave arrival was not associated with an audio signal while at Matsushiro and in the Pyrenees the S-wave was coincident with sound in some cases. The relative rarity of sound associated with the stronger S-wave shaking can be explained by the attenuation of higher seismic, and therefore audible, frequencies in the S-wave (Hill et al., 1976).

Applications of sounds

Mallet's classic study of the Great Neapolitan earthquake of 1857 includes many reports of people hearing the earthquake and he devotes Part III, Chapter VIII, titled "Of the Sounds that Attended the Shock," to discussing these observations and attempting to interpret them in terms of a finite source model by drawing an analogy to a long line of troops firing their rifles in succession (Mallet, 1862).

Davison's (1938) study of earthquake sounds attempted to relate his seven classifications to the maximum intensity on the Rossi–Forel scale, the distance from the event, and the region within which the event occurred. He found that earthquakes were heard over the entire felt area for small earthquakes but only half of the felt area for larger events and that at greater distances the sound was described as being smoother and more monotonous. The smoother and more monotonous sound with increasing distance is consistent with the attenuation of high frequency shaking at greater distances from the event. Due to this attenuation, low frequency motion from large earthquakes is felt at distances beyond the point where shaking in the higher frequency human hearing range is transmitted. Thus, for large earthquakes the area of felt shaking is larger than the area where the earthquake can be heard.

In a more recent study, Sylvander and Mogos (2005) developed relationships between the locations of sounds heard, earthquake magnitude, and epicentral distance and considered sound as a factor in human detection of small earthquakes. Souriau (2006) also explored relationships between heard sounds and shaking and points out that sound may be responsible for waking people and thus could contribute to intensity values at the low end of the scales. That builds on Hill et al.'s (1976) observation that the sound can be audible even when the P-wave is not felt. Tosi et al. (2000) examined whether records of sounds could be used to infer focal mechanisms. Partially because the sound observations do not include polarity information, they concluded that the uncertainties were too large to use such data to infer unknown focal mechanisms.

Despite these efforts, it is not clear that earthquake sounds make a significant contribution to the analysis of pre-instrumental earthquakes beyond helping us understand how such sounds contribute to observations of lower intensities. As noted above, there are problems with interpreting subjective reports of sounds that mix together both the natural transmission of shaking into the atmosphere and the shaking of man-made objects. These objections could be removed for future earthquakes by making audio recordings, which could easily be integrated into existing digital seismic acquisition systems (Sylvander et al., 2007). Unfortunately, audio recordings may not be a very useful addition to the collocated seismograms because little of the seismic energy is transmitted into

the atmosphere and because audio captures only a scalar recording of the 3-component ground motion.

While microphones may not be an effective way to record seismic data, seismometers can produce useful records of acoustic waves in the atmosphere that are strong enough to shake the ground. For instance, seismic records of acoustic waves have contributed to forensic seismology investigations into the 1995 Oklahoma City bombing (Holzer et al., 1996), the 1998 truck bombing of the American embassy in Nairobi (Koper et al., 1999), a chemical plant explosion in Nevada (Ichinose et al., 1999), and a pipeline explosion in New Mexico (Koper et al., 2003).

Infrasound

Very low frequency acoustic waves known as infrasound can travel great distances through the atmosphere, although with frequencies below 20 Hz these waves are not audible to humans. Bolt (1964) demonstrated that micropressure recordings after the 1964 Mw 9.2 Alaska earthquake included acoustic waves with periods of 20–40 s (frequencies of 0.025–0.050 Hz) generated near the earthquake and recorded 3,100 km away at Berkeley, California. Donn and Posmentier (1964) further explored these recordings and showed that earlier infrasound arrivals were generated near the recording station by the passing Rayleigh waves. Away from the earthquake, infrasound can also be generated when surface waves travel through mountainous regions. A good review of these mechanisms is Mutschlecner and Whitaker (2005) and the references therein. More recently Le Pichon et al. (2006) suggested that infrasound could be used to study the amplification of ground displacement by topography, especially in regions with inadequate seismic networks. And Green et al. (2009) demonstrated that coastal cliffs could also be a source of infrasound. While infrasound arrays have recorded many earthquakes, their primary applications continue to be the study of atmospheric phenomena such as meteorites and the study of explosions whether chemical or nuclear tests.

Sonification, education, and art

Sonification is the process of producing sound from nonauditory data. Most seismic data contains frequencies below human hearing and the simplest way to sonify seismograms is to speed them up, which increases the frequency content, and play them back through an audio system. Steinbrugge (1974) mentions that, in 1952 or 1953, Hugo Benioff and Cook Laboratories (Stamford, Conneticut) used this method to produce an LP record with one side called "Earthquakes of this World." Many seismologists have followed in Benioff's footsteps and this has become particularly easy now that digital seismograms can be converted to audio files using common software such as MatLab.

The human ear is an excellent spectral analyzer and most people are more used to listening to music than looking at seismograms. Therefore, sonification of seismograms can be an effective educational tool. Michael (1997) (http://earthquake.usgs.gov/learn/listen/index.php) developed a set of audio playbacks of seismograms that explores how the frequency of shaking varies with source dimension, epicentral distance, and site geology. This Web site includes a recording of the 1992 M7.3 Landers earthquake recorded at Mammoth Lakes where local earthquakes were triggered by the passing seismic waves. The local earthquakes are easily heard in the raw sonified seismogram but required filtering to be detected visually in the plotted seismograms (Hill et al., 1993). This demonstrates that audio can be a powerful educational tool.

Sound is a common part of experiencing earthquakes and so naturally it is part of artistic explorations of seismic events. The seismograms from the above site were used as the basis for my Earthquake Quartet #1 for Trombone, Cello, Voice, and Seismograms (Michael, 2000) which is available online at http://earthquake.usgs.gov/learn/music/. Earthquake Quartet #1 starts by describing the earthquake cycle and then explores the idea that society and culture, including music, take place with the earthquakes as an often-ignored backdrop. The earthquakes play the role of percussion because they have abrupt beginnings and broad, nonharmonic, spectra.

Another musical approach was taken by Loos and Scherbaum (1999) in a CD titled "Inner Earth, a seismosonic symphony" which used many different methods of shifting, converting, and transposing seismograms to make them audible. They produced music with harmonic properties that is purely based on seismograms but sounds nothing like the ones that have simply been sped up.

Overlaying a seismogram on a musical staff provides a form of graphical musical notation that can then be played back using synthesized instruments. Some seismograms from the Mt. Etna volcano can be heard as piano music at http://grid.ct.infn.it/etnasound/page4/page8/page8.html. The goal of that project was to provide a way to detect changes in the behavior of the volcano by hearing changes in the sonified seismograms (http://grid.ct.infn.it/etnasound/) and it also extended into the arts.

The "Great California ShakeOut" earthquake preparedness drills (Jones et al., 2008) use audio combining synthesized earthquake sounds with sounds of damage occurring and sirens to simulate an earthquake experience at the start of the drills. These files can be downloaded from http://www.shakeout.org/drill/broadcast/index.html.

"Memento Mori: an internet based earthwork" by Goldberg et al. (1999) used live seismic data to trigger nonseismic audio in a dark installation that also included a visual display of the seismic particle motion. This approach was later used for "Ballet Mori: a ballet conducted by the earth" performed by Muriel Maffre at the San Francisco Ballet to commemorate the centennial of the 1906 earthquake (Goldberg et al., 2006). Video of the dance can be seen at http://goldberg.berkeley.edu/art/Ballet-Mori/.

In 2009, the Parkfield Interventional Earthquake Fieldwork (PIEQF) (Rogers, 2010), and online at http://pieqf.allshookup.org/, consisted of a large shake table whose motion was triggered by earthquakes throughout the state of California. Both the table itself and a set of rods mounted on the table created noise when an earthquake occurred. This mechanical amplification of the shaking replaced the natural "roaring" discussed by Seneca with a clatter that allowed the viewers to hear distant seismic activity in a way that is not possible through natural shaking and sounds.

Summary

Earthquake sounds form an important part of the earthquake experience, and thus are a large part of artistic explorations of these events, but their study forms a small niche within seismology. The last few decades have produced the data necessary to finally understand this phenomenon along with its role in the study of earthquakes and explosions.

Acknowledgments

This article greatly benefited from reviews by Jeanne Hardebeck, David Hill, Carol Mark, Stephanie Ross, and Matthieu Sylvander.

Bibliography

Bolt, B. A., 1964. Seismic air waves from the great 1964 Alaskan earthquake. *Nature*, **202**, 1095–1096.

Clarke, M. A., and Geike, A., 1910. *Physical Science in the Time of Nero Being a Translation of the Quaestiones Naturales of Seneca*. London: MacMillian, p. 368.

Davis, W. M., 1897. Current notes on physiography. *Science*, **6**(153), 834–835.

Davison, C., 1938. Earthquake sounds. *Bulletin of the Seismological Society of America*, **28**(3), 147–161.

Donn, W. L., and Posmentier, E. S., 1964. Ground-coupled air waves from the great Alaskan earthquake. *Journal of Geophysical Research*, **69**(24), 5357–5361.

Ebel, J. E., 1982. The 1981 microearthquake swarm near Moodus, Conneticut. *Geophysical Research Letters*, **9**(4), 397–400.

Ewing, J. A., 1883. *Earthquake Measurement*. Tokyo: Tokio Daigaku, 125pp.

Goldberg, K., et al., 1999. *Memento Mori: an internet-based earthwork*. Berekely, California, http://goldberg.berkeley.edu/art/mori/.

Goldberg, K., et al., 2006. *Ballet Mori: a ballet conducted by the earth*. San Francisco, http://goldberg.berkeley.edu/art/BalletMori/.

Green, D. N., et al., 2009. Modelling ground-to-air coupling of the shallow Ml 4.3 Folkestone, United Kingdom, earthquake of 28 April 2007. *Bulletin of the Seismological Society of America*, **99**(4), 2541–2551.

Hill, D. P., 1976. *Physical constraints on sounds generated by very small earthquakes*. USGS Open File Report, 27 pp.

Hill, D. P., et al., 1976. Earthquake sounds generated by body-wave ground motion. *Bulletin of the Seismological Society of America*, **66**(4), 1159–1172.

Hill, D. P., et al., 1993. Seismicity remotely triggered by the magnitude 7.3 Landers, California, earthquake. *Science*, **260**(5114), 1617–1623.

Holzer, T. L., et al., 1996. Seismograms offer insight into Oklahoma City bombing. *EOS*, **77**(41), 393, 398–399.

Ichinose, G. A., et al., 1999. Seismic analysis of the 7 January 1998 chemical plant explosion at Kean Canyon, Nevada. *Bulletin of the Seismological Society of America*, **89**, 938–945.

Japan Meteorological Agency, 1968. Report on the Matsushiro earthquake swarm, August 1965–December 1967, Technical Report of the Japan Meteorological Agency, pp. 125–137.

Jones, L. M., et al., 2008. *The ShakeOut scenario*, USGS Open File Report, 312 pp.

Koper, K. D., et al., 1999. Seismic analysis of the 7 August 1998 truck-bomb blast and the American embassy in Nairobi, Kenya. *Seismological Research Letters*, **70**, 512–521.

Koper, K. D., et al., 2003. Seismic recordings of the Carlsbad, New Mexico, pipeline explosion of 19 August 2000. *Bulletin of the Seismological Society of America*, **93**(4), 1427–1432.

Le Pichon, A., et al., 2006. Multistation infrasonic observations of the Chilean earthquake of 2005 June 13. *Geophysical Journal International*, **167**, 838–844.

Loos, W., and Scherbaum, F., 1999. *Inner Earth, a Seismosonic Symphony*. Berlin: Traumton.

Mallet, C. E., 1862. *Great Neapolitan Earthquake of 1857*. London: Chapman and Hall, Vol. 2, 399 pp.

Mallet, C. E., and Mallet, J. W., 1858. *The Earthquake Catalogue of the British Association with the Discussion, Curves, and Maps, Etc*. London: British Association for the Advancement of Science, 768 pp.

Michael, A. J., 1997. *Listening to earthquakes*. USGS, http://earthquake.usgs.gov/learn/listen/index.php.

Michael, A. J., 2000. *Earthquake Quartet #1*. USGS, http://earthquake.usgs.gov/learn/music/.

Mutschlecner, J. P., and Whitaker, R. W., 2005. Infrasound from earthquakes. *Journal of Geophysical Research*, **110**, D01108.

Pulli, J. J., et al., 1980. Field investiation and fault plane solution of the Bath, Maine, earthquake of April 18, 1979. *Earthquake Notes*, **51**(4), 39–46.

Rogers, D. V., 2010. Parkfield Interventional EQ Fieldwork (PIEQF), Master's thesis. Sydney: University of New South Wales.

Souriau, A., 2006. Quantifying felt events: a joint analysis of intensities, accelerations and dominant frequencies. *Journal of Seismology*, **10**, 23–38.

Steinbrugge, K. V., 1974. A catalog of earthquake related sounds. *Bulletin of the Seismological Society of America*, **64**(5), 1409–1418.

Sylvander, M., and Mogos, D. G., 2005. The sounds of small earthquakes: quantitative results from a study of regional macroseismic bulletins. *Bulletin of the Seismological Society of America*, **95**(4), 1510–1515.

Sylvander, M., et al., 2007. Seismoacoustic recordings of small earthquakes in the Pyrenees: experimental results. *Bulletin of the Seismological Society of America*, **97**(1B), 294–304.

Thouvenot, F., et al., 2009. 200-m-deep earthquake swarm in Tricastin (lower Rhoˆ ne Valley, France) accounts for noisy seismicity over past centuries. *Terra Nova*, **21**(3), 203–210.

Tosi, P., et al., 2000. Spatial patterns of earthquake sounds and seismic source geometry. *Geophysical Research Letters*, **27**(17), 2749–2752.

Wald, D. J., et al., 1999. Utilization of the Internet for rapid community intensity maps. *Seismological Research Letters*, **70**, 680–697.

Cross-references

Earthquake, Focal Mechanism
Earthquakes, Intensity

Earthquakes, Source Theory
Propagation of Elastic Waves: Fundamentals
Seismic Instrumentation
Surface Waves

EARTHQUAKE, AFTERSHOCKS

Mian Liu[1], Seth Stein[2]
[1]Department of Geological Sciences, University of Missouri, Columbia, MO, USA
[2]Department of Earth and Planetary Sciences, Northwestern University, Evanston, IL, USA

Definition

Aftershocks. Smaller earthquakes following a large earthquake (the mainshock) in the same ruptured area.

Introduction

Earthquakes typically occur in sequences that may include foreshocks, the mainshock (the largest event or events), and aftershocks. Earthquake sequences without a clear mainshock are called swarms.

Aftershocks generally refer to the smaller earthquakes that follow a mainshock within certain spatial and temporal windows. However, the criteria for choosing these windows are somewhat arbitrary. Typically, aftershocks are defined within an area around the mainshock's source region (i.e., the ruptured fault segment, which is about 100 km long for a magnitude 7.0 earthquake). Most aftershocks occur on the main rupture surface; hence, they are often used to define the complex geometry of the rupture plane. However, in many cases, especially for earthquakes in subduction zones, the aftershock area increases significantly following the mainshock (Tajima and Kanamori, 1985). Some aftershocks also occur off the main rupture surface. Distant earthquakes triggered by the mainshock's seismic waves (see *Artificial Water Reservoir Triggered Earthquakes*) may also be regarded as aftershocks (Gomberg and Johnson, 2005). Temporally, aftershocks are defined as seismicity above the background activity following a main shock. This definition also leaves room for ambiguity, as the background seismic activity can be difficult to define in some places.

Aftershocks usually count for less than 5% of the total seismic energy release of the entire seismic sequence, so they are secondary products (Scholz, 2002). However, besides their spatial and temporal relations to the mainshock, aftershocks are fundamentally no different from other earthquakes and are recognized as aftershocks only retrospectively (Helmstetter et al., 2003; Utsu, 2002).

Scaling laws for aftershocks

A large earthquake is usually followed by many smaller events (aftershocks); their occurrence rate decreases with time, typically following a pattern known as the Omori's law, named for Fusakichi Omori's observation of the 1891 Nobi earthquake in Japan (Omori, 1894) and later modified by Utsu (1961):

$$n(t) = \frac{K}{(c+t)^p}$$

where $n(t)$ is the number of aftershocks by time t after the mainshock; K, p, and c are constants. The value of c is typically positive close to zero, and p close to 1. Hence, the number of aftershocks drops nearly hyperbolically with time (Figure 1). The values of these constants are obtained by fitting to the data for each aftershock sequence.

It has been observed that the largest aftershock is usually about 1 magnitude unit smaller than the mainshock, independent of the mainshock magnitude. This is known as Båth's law (Bath, 1965; Richter, 1958). However, because the data selection is retrospective and subjective, the size of aftershocks can vary substantially for different earthquake sequences.

Like all earthquakes, the size distribution of aftershocks follow the Gutenberg–Richter law (Gutenberg and Richter, 1954):

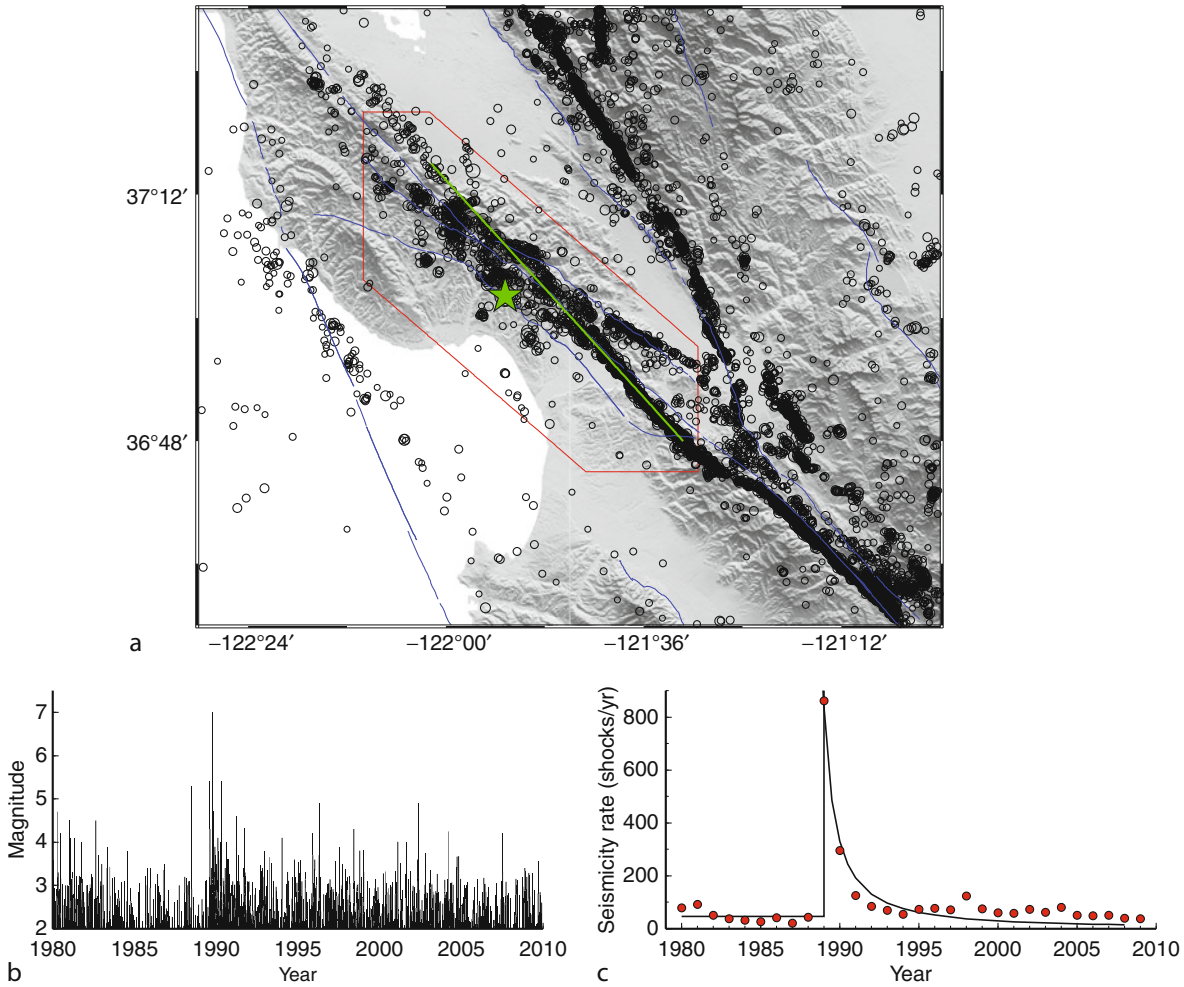
$$\log_{10} N(M) = a - bM$$

where $N(M)$ is the number of earthquakes of magnitude $\geq M$, a and b are constants. This relationship plots as a straight line with slope b, whose value, often referred to as the b-value, is typically in the range of 0.8–1.2. For $b = 1$, the number of earthquakes increases by a factor of 10 for every unit drop. The b-value varies in time and space even for the same fault. Some studies have suggested that the b-value can be used as a stress indicator, with lower b-values often associated with higher stresses (Schorlemmer et al., 2005).

Causes of aftershocks

Because most aftershocks occur on or near the rupture surface, they are thought to result from incomplete rupture and heterogeneous slip (e.g., Bullen and Bolt, 1947). This notion is consistent with some case studies that found aftershocks concentrated around the perimeter of the rupture zone or around asperities within the rupture zone where most of the coseismic slip occurred (Beroza, 1991; Scholz, 2002).

The main causes of aftershocks are thought to include mainshock-induced changes of frictional properties of the fault zone and stress perturbations. Laboratory experiments indicate that frictional properties change with the duration of stationary contact and the instantaneous sliding velocity (Scholz, 2002). A constitutive relationship based on such results, known as the rate-and-state law (Dieterich, 1979; Ruina, 1983), can explain the time-dependent changes of seismic rates that are consistent with the Omori's law (Dieterich, 1994). Stress perturbations may arise from creep recovery of rocks in the immediate areas of the fault (Benioff, 1951), viscous relaxation from the lower crustal and upper mantle (Lieber and Braslau,



Earthquake, Aftershocks, Figure 1 (**a**) Topography and seismicity (*circles*, size proportional to magnitude) around the epicenter (*star*) of the 1989 Loma Prieta earthquake (Mw 6.9) on the San Andreas Fault. *Green* line marks the fault rupture; *red* lines enclose the spatial window for the aftershocks. (**b**) Earthquake occurrence between 1980 and 2010 within the spatial window. (**c**) Plot of seismicity showing the t^{-1} decay of aftershocks.

1965; Stein and Liu, 2009), and pore-elastic effects (Scholz, 2002). In particular, the changes of the Coulomb static stress resulted from the mainshock have been shown to cause spatial migration of seismicity and trigger earthquakes (Parsons, 2002; Stein, 1999). Other causes may include dynamic triggering (Gomberg and Johnson, 2005) and pore fluid flow (Miller et al., 2004; Nur and Booker, 1972).

Aftershocks and earthquake hazard

Aftershocks can be large and damaging (e.g., Wiemer et al., 2002). The durations of aftershock sequences for plate-boundary earthquakes are typically around 10 years (Parsons, 2002) (Figure 1). However, aftershock sequences away from plate-boundary faults tend to last longer. In the broadly deforming western United States, for example, aftershocks are found to last 50–100 years. Within the stable continental interior, aftershock sequences may last hundreds of years or even longer (Stein and Liu, 2009). Such long aftershock sequences in continental interiors, where large earthquakes are infrequent and historic records are often incomplete, could bias hazard assessment.

Further complication may arise from the fact that earthquakes tend to cluster in time and space, which is best shown in aftershock sequences. Studies of large shallow earthquakes in the world show that these large events often occur in pairs or groups (called doublets or multiplets); the later events, which could be regarded as aftershocks, can be bigger than the triggering event (the mainshock) (Kagan and Jackson, 1999). Such clustering patterns are consistent with the emerging view that all earthquakes can trigger their own earthquakes that in turn trigger more quakes, and the triggered events may be bigger than the

triggering event (Helmstetter et al., 2003; Ogata, 1998). Because the triggered events (the aftershocks) statistically follow the Omori's law, statistical models can be developed to assess time-dependent probability of future damaging earthquakes after each event (Gerstenberger et al., 2005; Jones, 1985; Reasenberg and Jones, 1989).

Summary

Aftershocks are smaller earthquakes following the mainshock. They typically occur on or near the rupture plane of the mainshock, resulting from changes of stress and frictional properties of the fault zone caused by the mainshock. The duration of aftershock sequences is typically a few years for earthquakes at plate boundaries, but can last much longer within stable continental interiors.

Bibliography

Bath, M., 1965. Lateral inhomogeneities in the upper mantle. *Tectonophysics*, **2**, 483–514.

Benioff, H., 1951. Earthquakes and rock creep. *Bulletin of the Seismological Society of America*, **41**, 31–62.

Beroza, G. C., 1991. Near-source modeling of the Loma Prieta earthquake: evidence for heterogeneous slip and implications for earthquake hazard. *Bulletin of the Seismological Society of America*, **81**, 1603–1621.

Bullen, K. E., and Bolt, B. A., 1947. *An Introduction to the Theory of Seismology*. New York: Cambridge University Press. 499 pp.

Dieterich, J., 1979. Modeling of rock friction: 1. Experimental results and constitutive equations. *Journal of Geophysical Research*, **84**, 2161–2168.

Dieterich, J., 1994. A constitutive law for rate of earthquake production and its application to earthquake clustering. *Journal of Geophysical Research*, **99**, 2601–2618.

Gerstenberger, M. C., Wiemer, S., Jones, L. M., and Reasenberg, P. A., 2005. Real-time forecasts of tomorrow's earthquakes in California. *Nature*, **435**, 328–331.

Gomberg, J., and Johnson, P., 2005. Seismology: dynamic triggering of earthquakes. *Nature*, **437**, 830.

Gutenberg, B., and Richter, C. F., 1954. *Seismicity of the Earth and Associated Phenomena*. Princeton: Princeton University Press, Vol. ix, p. 310.

Helmstetter, A., Sornette, D., and Grasso, J.-R., 2003. Mainshocks are aftershocks of conditional foreshocks: how do foreshock statistical properties emerge from aftershock laws. *Journal of Geophysical Research*, **108**, 2046.

Jones, L. M., 1985. Foreshocks and time-dependent earthquake hazard assessment in southern California. *Bulletin of the Seismological Society of America*, **75**, 1667–1679.

Kagan, Y. Y., and Jackson, D. D., 1999. Worldwide doublets of large shallow earthquakes. *Bulletin of the Seismological Society of America*, **89**, 1147–1155.

Lieber, P., and Braslau, D., 1965. *On an earthquake and aftershock mechanism relating to a model of the crust and mantle*. Report Am-65-8. Office of Research Services, University of California, Berkeley, p. 141.

Miller, S. A., Collettini, C., Chiaraluce, L., Cocco, M., Barchi, M., and Kaus, B. J. P., 2004. Aftershocks driven by a high-pressure CO_2 source at depth. *Nature*, **427**, 724–727.

Nur, A., and Booker, J. R., 1972. Aftershocks caused by pore fluid flow? *Science*, **175**, 885–887.

Ogata, Y., 1998. Space–time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, **50**, 379–402.

Omori, F., 1894. On the aftershocks of earthquakes. *Journal of the College of Science, Imperial University of Tokyo*, **7**, 111–200.

Parsons, T., 2002. Global Omori law decay of triggered earthquakes: Large aftershocks outside the classical aftershock zone. *Journal of Geophysical Research*, **107**(B9), 2199, doi:10.1029/2001JB000646.

Reasenberg, P. A., and Jones, L. M., 1989. Earthquake hazard after a mainshock in California. *Science*, **243**, 1173–1176.

Richter, C. F., 1958. *Elementary Seismology*. San Francisco: W. H. Freeman. 768 pp.

Ruina, A., 1983. Slip instability and state variable friction laws. *Journal of Geophysical Research*, **88**, 10359–10370.

Scholz, C. H., 2002. *The Mechanics of Earthquakes and Faulting*. Cambridge/New York: Cambridge University Press. 471 pp.

Schorlemmer, D., Wiemer, S., and Wyss, M., 2005. Variations in earthquake-size distribution across different stress regimes. *Nature*, **437**, 539–542.

Stein, R. S., 1999. The role of stress transfer in earthquake occurrence. *Nature*, **402**, 605–609.

Stein, S., and Liu, M., 2009. Long aftershock sequences within continents and implications for earthquake hazard assessment. *Nature*, **462**, 87–89.

Tajima, F., and Kanamori, H., 1985. Global survey of aftershock area expansion patterns. *Physics of the Earth and Planetary Interiors*, **40**, 77–134.

Utsu, T., 1961. A statistical study of the occurrence of aftershocks. *Geophysical Magazine*, **30**, 521–605.

Utsu, T., 2002. Statistical features of seismicity. In Lee, W. H. K. (ed.), *International Handbook of Earthquake & Engineering Seismology, Part A*. San Diego: Academic, pp. 719–732.

Wiemer, S., Gerstenberger, M., and Hauksson, E., 2002. Properties of the aftershock sequence of the 1999 Mw 7.1 Hector Mine earthquake: implications for aftershock hazard. *Bulletin of the Seismological Society of America*, **92**, 1227–1240.

Cross-references

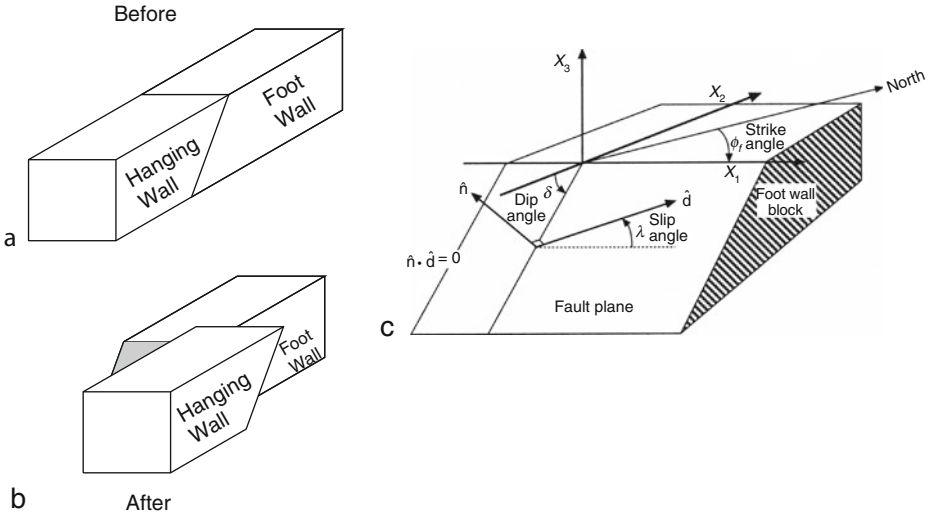
Earthquake, Foreshocks
Earthquakes, Source Theory

EARTHQUAKE, FOCAL MECHANISM

Emile A. Okal
Department of Earth & Planetary Sciences, Northwestern University, Evanston, IL, USA

Definition

Focal mechanisms are geometrical or mathematical representations of the faulting during an earthquake. In very simple terms, the latter consists of the relative motions of two blocks of Earth called *walls* along a planar surface called *fault*. Figure 1 shows that the description of an earthquake rupture needs three angles. The *strike* angle ϕ is the azimuth (with respect to North) of the trace of the fault on a horizontal plane such as the Earth's surface; the *dip* angle δ characterizes the steepness of the fault, and the *rake* or *slip* angle λ, the direction of motion, within the fault plane and relative to the horizontal, of the hanging wall relative to the foot wall. A full description of the earthquake rupture requires an additional scalar, related



Earthquake, Focal Mechanism, Figure 1 (**a**) and (**b**) Before and after sketches of the displacement of fault blocks during an earthquake. The fault plane is shaded in (**b**). (**c**) Definition of the strike, dip, and slip angles, ϕ, δ, λ in the geometry of the rupture represented at left. The unit vector $\hat{\mathbf{d}}$ represents the direction of motion of the hanging wall (not represented to prevent clutter) with respect to the foot wall. (**c**: After Stein and Wysession, 2003.)

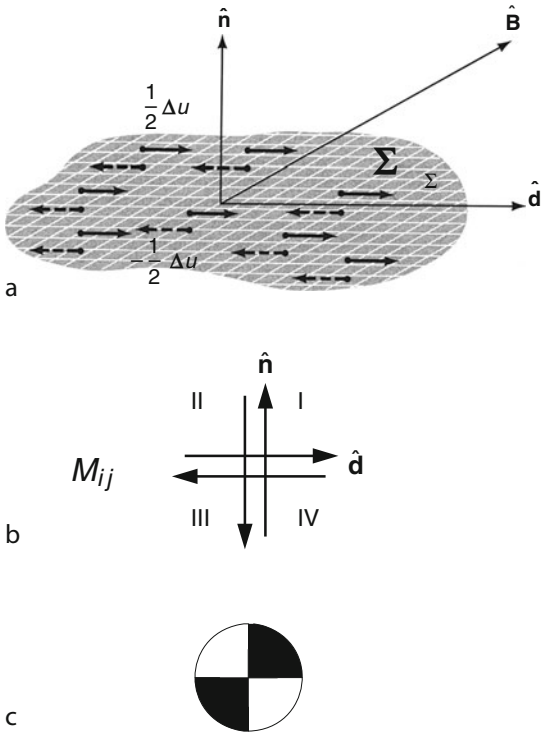
to the length of slip of the fault walls, and thus to the size or *magnitude* of the event.

Representation by a double-couple

As such, an earthquake source, described by three angles and a scalar, is a more complex mathematical entity than a simple vector (described by two angles and a scalar), and in particular cannot be represented physically by a single force. Vvedenskaya (1956), followed in the West by Knopoff and Gilbert (1959), introduced the concept of the double-couple, a system of forces shown on Figure 2 and consisting of two opposing torques of equal moment M_0; this system has no dynamic effect (translation or rotation) on a rigid body, but results in deformation of a medium of finite rigidity.

The representation theorem (e.g., Aki and Richards, 1980, pp. 38–41) considers an elastic medium of rigidity μ and states that a dislocation with slip Δu (in the direction $\hat{\mathbf{d}}$) along a fault of surface Σ cut into the medium (Figure 2a) will generate the same field of dynamic deformation (expressed as seismic waves) as a double-couple of moment $M_0 = \mu\Sigma\,\Delta u$, embedded in the medium in the absence of the cut, with the direction of forces and levers in Figure 2b along those of the slip, $\hat{\mathbf{d}}$, and of the normal to the fault plane, $\hat{\mathbf{n}}$.

The double-couple shown in Figure 2b separates the space around it into four quadrants containing respectively the heads and tails of the arrows. They are delineated by the fault plane and the plane normal to the slip $\hat{\mathbf{d}}$. In quadrants I and III, the seismic motion will be "*anaseismic*," or away from the source, while in II and IV, it will be "*kataseismic*" or toward the source. The quadrants are accordingly shaded black and white in Figure 2c. Note that in Figure 2b and c, the source is



Earthquake, Focal Mechanism, Figure 2 (**a**) Seismic slip along a fault (represented as the hatched area of surface $\boldsymbol{\Sigma}$). The total slip $\boldsymbol{\Delta u}$ is the difference between the displacements of the top wall (shown as *solid arrows*) and of the bottom wall (*dashed arrows*). (**b**) "Double-couple" system of forces, equivalent, in the unruptured medium, to the dislocation shown in (**a**). (**c**) "Beachball" representation of the quadrants defined by the orientation of the double-couple. (**a**: After Aki and Richards, 1980.)

approximated as a *point* in space, hence the concept of a "point-source double-couple."

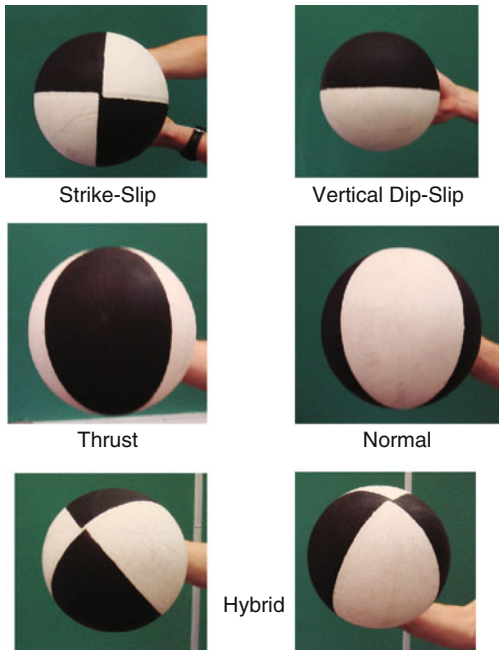
Focal mechanisms from P-wave first motions

Under the high-frequency approximation of geometrical optics, a most remarkable property of *P* waves is the conservation of the ana- or kataseismic character along the entire seismic ray, all the way to a receiving seismic station at the Earth's surface, where an anaseismic wave results in an initial *upward* motion of the ground, while a kataseismic one gives a *downward* first motion.

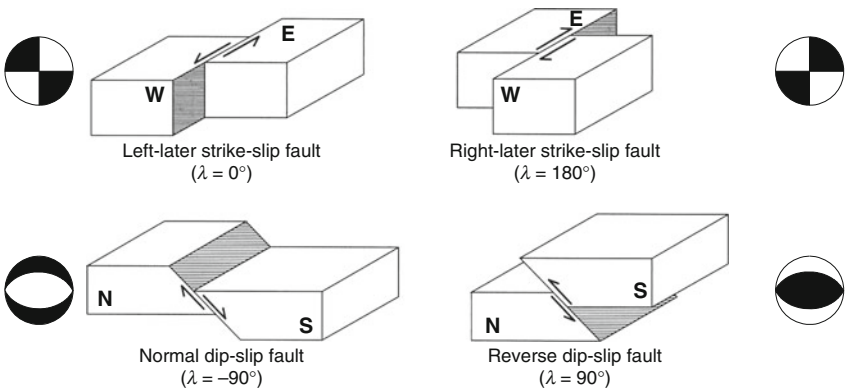
This property has been used extensively to obtain earthquake focal mechanisms by gathering first-motion *P*-wave data at large datasets of local and distant stations and backtracking this information along the relevant rays into a mapping of the focal sphere at the source. This amounts to resolving the orientation in space of the diagram on Figure 2c, known traditionally as a "beachball." Because of the difficulty of representing a full sphere on a planar figure (long known to world cartographers), and thanks to the symmetry of the beachball, it is convenient and sufficient to represent only a stereographic projection of the *lower focal hemisphere* separated at the horizontal plane and involving rays leaving down into the deep Earth, and eventually emerging at large distances. This constitutes the standard representation of earthquake focal mechanisms. Figure 3 illustrates examples of the most common types of mechanisms.

Note that a change of focal mechanism results in a mere *solid rotation* in space of the beachball pattern, but does not change its nature (Figure 4).

A common feature of the double-couple concept and of its beachball representation is their high level of symmetry. The two planes defining the quadrants on Figure 2b are conceptually interchangeable and as a result, there exists an inherent indeterminacy between two possible physical solutions obtained from any field of seismic data generated by a point source double-couple. For example, a left-lateral strike-slip earthquake on a vertical fault striking North–South is seismically indistinguishable from a right-lateral one on a fault striking East–West. This property, rooted in the symmetry of the double-couple, applies to all seismic waves, including *P* and *S* waves, surface waves, and the free oscillations of the Earth. The indeterminacy can be lifted only by considering a source extended in space, as opposed to a point source. In practice, this is done either by using field methods to recognize the orientation of the fault through its expression on the



Earthquake, Focal Mechanism, Figure 4 A basketball has been divided into four quadrants alternately painted black and white. By rotating the basketball in space, one can obtain a representation of all possible focal orientations of the double-couple shown in Figure 2b.This experiment demonstrates that earthquake focal geometries differ only by a solid rotation in space.



Earthquake, Focal Mechanism, Figure 3 Examples of focal geometries with corresponding beachball patterns. (Adapted from Stein and Wysession, 2003.)

ground, or by mapping the aftershocks of the event which are expected to occur along the direction of faulting, or by considering the so-called source-finiteness effects in the seismograms, which result from the interference of the elementary sources, distributed both in time and space, as the rupture propagates over the faulting area. In all cases, note again that the fault can no longer be represented as a point source.

Stress release

In another interpretation of focal mechanisms, one can expand the double-couple as a two-dimensional deviatoric and symmetric tensor

$$M_{ij} = M_0\left(\hat{d}_i\hat{n}_j + \hat{n}_i\hat{d}_j\right) \quad (1)$$

where $\hat{\mathbf{d}}$ and $\hat{\mathbf{n}}$ are the unit vectors along the direction of slip, and perpendicular to the fault, respectively. These components and the principal directions of the moment tensor are directly related to those of the stresses released during the earthquake.

In particular, the bisector $\mathbf{T} = \frac{1}{\sqrt{2}}\left(\hat{\mathbf{d}} + \hat{\mathbf{n}}\right)$ represents the axis of maximum tension and the direction of maximum positive amplitude of *P* waves; it plots at the center of the shaded quadrants on the beachball. The direction $\mathbf{P} = \frac{1}{\sqrt{2}}\left(\hat{\mathbf{d}} - \hat{\mathbf{n}}\right)$ is the axis of maximum compression (often and improperly called "pressure axis") and that of maximum negative amplitude of *P* waves; it plots at the center of the open quadrants. The direction $\mathbf{B} = \hat{\mathbf{d}} \times \hat{\mathbf{n}} = \mathbf{T} \times \mathbf{P}$ is the hinge of the fault planes; it is a direction where both *P* and *S* amplitudes vanish. For a given beachball, the **T, P, B** axes are independent of the choice between the two possible fault planes.

The interpretation of focal mechanisms in terms of stress release is particularly suited to the study of intraplate earthquakes, which express the release of stresses accumulated inside the plates by tectonic forces, while interplate events, which are controlled by the relative displacements of two of the Earth's plates at a common boundary, are more readily interpreted in terms of a slip vector on a fault plane.

Fundamental results

The most important application of focal mechanisms was the verification of the fundamental concepts of continental drift, seafloor spreading, and plate tectonics, in the mid-1960s. These ideas were derived using the spatial distribution of earthquake epicenters, but without the knowledge of the geometries of their ruptures. The latter could then be used as an independent verification of the concepts, following the development of the beachball model, and the deployment of the World-Wide Standardized Seismograph Network in the early 1960s. Sykes (1967) and Isacks et al. (1968) achieved a remarkable endorsement of the mechanisms predicted by plate motions (normal faulting expressing extension at the mid-oceanic ridges, underthrusting expressing convergence at the subduction zones), and most spectacularly, verified the polarity of strike-slip faulting at the oceanic transform faults, thus upholding the concept proposed by Wilson (1965).

Focal mechanisms inside subducting slabs were compiled by Isacks and Molnar (1971), who documented a universal mechanism of down-dip compression for events deeper than 450 km, expressing the difficulty of penetration by the slab of the lower mantle at the bottom of the transition zone. By contrast, a wide diversity of mechanisms is found in the upper parts of the slabs, where local parameters such as rate and steepness of subduction, as well as possible bending stresses, control the stress field accumulated inside the slab during the subduction process.

A most remarkable aspect of the focal mechanisms of deep earthquakes is the fact that, without exception, they can be modeled by a double-couple (e.g., using its beachball representation), even though the concept of brittle rupture should not apply at the relevant pressure and temperature conditions inside the slabs. In particular, modern analysis of major deep earthquakes (Kawakatsu, 1991; Okal, 1996) has ruled out the possibility of an isotropic (implosive) source even as a mere component of their mechanism, as initially envisioned by Bridgman (1945) and proposed by Gilbert and Dziewonski (1975). Deep events represent faulting on a planar dislocation in the framework of Figure 1, and differ from their shallow counterparts only in their usually higher level of stress drop (Antolik et al., 1999), but not in the geometrical or mechanical nature of their process. This observation supports the hypothesis that deep earthquakes may be induced by transformational faulting in volumes of metastable olivine persisting inside the slabs at greater depths than predicted by equilibrium thermodynamics (Kirby et al., 1996).

Modern focal solutions: inversion of the moment tensor

The double-couple used to represent a seismic source can be used, through its components given by Equation 1, as a forcing term in the wave equations expressing the fundamental law of dynamics (a variation of "$\mathbf{F} = m\,\mathbf{a}$"), to compute simulations of the Earth's response in space and time to the seismic disturbance. These are known as *synthetic seismograms*.

Conversely, and since the 1970s, inversion techniques have been used to derive the individual components of earthquake moment tensors directly from the waveforms of their recorded seismograms. Essentially, a best-fitting algorithm is used to express a large dataset of observed seismograms as a linear combination of synthetic seismograms computed for each individual component of the moment tensor. In addition, and for sufficiently large datasets, the inversion can resolve the best location in space of the source, and its result is known as a "centroid moment tensor" or CMT. These techniques, which can be applied on body waves, surface waves or

normal modes, as well as on regional phases in the near field, can be carried out automatically on digital data, and eliminate the occasional difficulty of picking a first motion polarity under noisy or emergent conditions. They have replaced the compilation of *P*-wave first motions as the primary source of earthquake focal mechanisms.

In particular, the Global CMT (*ex*–Harvard CMT) project, described initially by Dziewonski et al. (1981), has now inverted more than 30,000 earthquake sources since 1976, through a uniform, comprehensive procedure, into a catalog that has become the backbone of most modern seismotectonic studies.

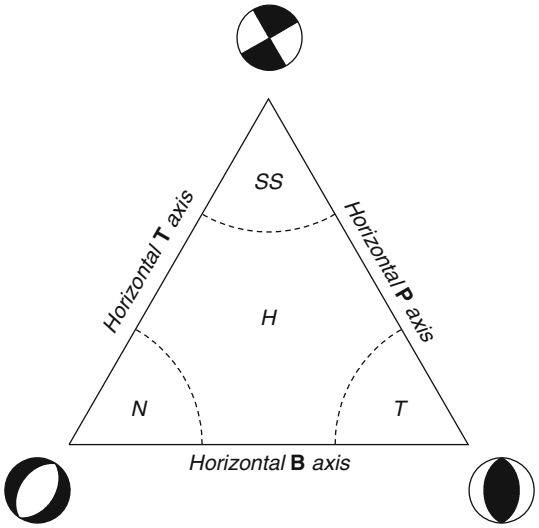
Evaluation tools for double-couple solutions

Because of the relative complexity of the double-couple source that cannot be represented by a vector, a number of applications have been developed to illustrate their properties in a user-friendly representation. Among them, Frohlich and Apperson (1992) have proposed to plot focal mechanisms in a ternary diagram whose apices are the "*pure*" mechanisms corresponding to vertical stress axes (**B**, strike-slip; **T**, thrust; or **P**, normal); using barycentric coordinates within the triangle, the position of the mechanism indicates at a glance the nature of faulting (Figure 5). Note that it remains independent of the azimuth ϕ of the fault strike.

In order to help compare different earthquakes, Kagan (1991) has offered an algorithm to quantify the proximity of two focal solutions given by their angles $(\phi_1, \delta_1, \lambda_1)$ and $(\phi_2, \delta_2, \lambda_2)$. It computes the angle ω of the minimum solid rotation bringing one of the beachballs onto the other. Because of their symmetry, the maximum value of ω is 120°.

Non-double-couple solutions

The double-couple is a four-dimensional mathematical entity which constitutes a subset (featuring a zero-eigenvalue) but not a sub-vector space, of the five-dimensional vector space of deviatoric (i.e., zero-trace), symmetric second-order tensors. Consequently, linear inversions are carried out in the full five-dimensional space. The quality of fit of the inverted solution to a double-couple is assessed through a so-called Compensated Linear Vector Dipole (CLVD) parameter ε, expressing the ratio of the intermediate eigenvalue of the moment tensor to its largest one (in absolute value). In principle, ε could range from -0.5 to 0.5, and should be zero for a perfect double-couple. The average value of $|\varepsilon|$ for the whole CMT catalog is 0.12, which constitutes a good verification of the concept of the double-couple, especially since it falls to 0.08 for earthquakes with moments $M_0 \geq 10^{26}$ dyn cm. Occasionally, larger values of $|\varepsilon|$ (which can reach 0.4) indicate a deviation from the model of a double-couple, which may result from a number of scenarios. For example, earthquakes with complex ruptures consisting of multiple events with differing geometries will feature an overall deviatoric moment tensor, which however cannot be described in terms of a lone double-couple. A recent example is the Samoa event of 29 September 2009 ($M_0 \approx 1.8 \times 10^{28}$ dyn cm; $\varepsilon = 0.37$), which was modeled as a main outer rise normal faulting event triggering a smaller underthrusting shock on the nearby interplate contact (Li et al., 2009).



Earthquake, Focal Mechanism, Figure 5 Ternary representation of focal mechanisms (Frohlich and Apperson, 1992). This approach uses barycentric coordinates with respect to end members with one vertical principal axis. Mechanisms close to apices are either predominantly strike-slip (*SS*), thrust (*T*) or normal (*N*). Hybrid mechanisms (*H*) plot in the center of the diagram. By contrast, the sides of the diagram represent geometries with one horizontal axis. This representation is useful when studying trends among large families of mechanisms, such as intraplate earthquakes, irrespective of their particular strike angles.

Other examples of non-double-couple solutions would include dyke injections during volcanic events (e.g., Tori-Shima Volcano, 13 June 1984; $\varepsilon = 0.33$; Kanamori et al., 1993), rupture on non-planar faults, again in a volcanic environment (Ekström, 1994), and certain cryoseismic events believed to involve calving at the front of glaciers (e.g., 15 August 2008, $\varepsilon = 0.33$; Nettles and Ekström, 2009).

Conclusion

Focal mechanisms provide a mathematical or illustrative representation of the geometry of the rupture process during an earthquake. Their compilation forms the basis of our interpretation of earthquakes in the framework of global tectonics.

Bibliography

Aki, K., and Richards, P. G., 1980. *Quantitative Seismology.* New York: W.H. Freeman, p. 932.

Antolik, M., Dreger, D. S., and Romanowicz, B., 1999. Rupture processes of large deep-focus earthquakes from inversion of

moment rate functions. *Journal of Geophysical Research*, **104**, 863–894.
Bridgman, P. W., 1945. Polymorphic transition and geological phenomenon. *American Journal of Science*, **243A**, 90–97.
Dziewonski, A. M., Chou, A.-T., and Woodhouse, J. H., 1981. Determination of earthquake source parameters from waveform data for studies of global and regional seismicity. *Journal of Geophysical Research*, **86**, 2825–2852.
Ekström, G., 1994. Anomalous earthquakes on volcano ring-fault structures. *Earth and Planetary Science Letters*, **128**, 707–712.
Frohlich, C., and Apperson, K. D., 1992. Earthquake focal mechanisms, moment tensors and the consistency of seismic activity near plate boundaries. *Tectonics*, **11**, 279–296.
Gilbert, J. F., and Dziewonski, A. M., 1975. An application of normal mode theory to the retrieval of structural parameters and source mechanisms from seismic spectra. *Philosophical Transaction of Royal Society London*, **278A**, 187–269.
Isacks, B. L., and Molnar, P., 1971. Distribution of stresses in the descending lithosphere from a global survey of focal-mechanism solutions of mantle earthquakes. *Reviews of Geophysics and Space Physics*, **9**, 103–174.
Isacks, B. L., Oliver, J., and Sykes, L. R., 1968. Seismology and the newglobal tectonics. *Journal of Geophysical Research*, **73**, 5855–5899.
Kagan, Y. Y., 1991. 3-D rotation of double-couple earthquake sources. *Geophysical Journal International*, **106**, 709–716.
Kanamori, H., Ekström, G., Dziewoński, A. M., Barker, J. S., and Sipkin, S. A., 1993. Seismic radiation by magma injection: an anomalous seismic event near Tori Shima, Japan. *Journal of Geophysical Research*, **98**, 6511–6522.
Kawakatsu, H., 1991. Insignificant isotropic components in the moment tensor of deep earthquakes. *Nature*, **351**, 50–53.
Kirby, S. H., Stein, S., Okal, E. A., and Rubie, D., 1996. Deep earthquakes and metastable mantle phase transformations in subducting oceanic lithosphere. *Reviews of Geophysics and Space Physics*, **34**, 261–306.
Knopoff, L., and Gilbert, J. F., 1959. Radiation from a strike-slip earthquake. *Bulletin of the Seismological Society of America*, **49**, 163–178.
Li, X., Shao, G., and Ji, C., 2009. Rupture process of the M_w = 8.1 Samoa earthquake constrained by joint inverting teleseismic body, surface waves and local strong motion. *Eos Transactions, American Geophysical Union*, **90**, (52), U21D–03, [abstract].
Nettles, M., and Ekström, G., 2009. Ice loss and glacial earthquakes in Greenland and Antarctica. *Eos Transactions, American Geophysical Union*, **90**(52), G52B–01 [abstract].
Okal, E. A., 1996. Radial modes from the great 1994 Bolivian earthquake: no evidence of an isotropic component to the source. *Geophysical Research Letters*, **23**, 431–434.
Stein, S., and Wysession, M. E., 2003. *An Introduction to Seismology, Earthquakes and Earth Structure*. Malden: Blackwell, 498 pp.
Sykes, L. R., 1967. Mechanisms of earthquakes and nature of faulting on the mid-oceanic ridges. *Journal of Geophysical Research*, **72**, 2131–2153.
Vvedenskaya, A. V., 1956. Opredelenie polej smeshchenii pri zemletryasenyakh s pomoshchyu teorii dislokatsii. *Izv Akad Nauk SSSR, Ser Geofiz*, **6**, 277–284 [in Russian].
Wilson, J. T., 1965. A new class of faults and their bearing on continental drift. *Nature*, **207**, 343–347.

Cross-references

Earthquakes, Source Theory
Earthquake Rupture: Inverse Problem

EARTHQUAKE, FORESHOCKS

Mian Liu
Department of Geological Sciences, University of Missouri, Columbia, MO, USA

Synonyms

Preshocks

Definition

Foreshocks. Smaller earthquakes preceding a large earthquake (the mainshock) in the same ruptured area.

Introduction

Earthquakes typically occur in sequences that may include foreshocks, the mainshock (the largest event or events), and *Earthquake, Aftershocks*. Earthquake sequences without a clear mainshock are called swarms.

Foreshocks are the most obvious premonitory phenomenon of earthquakes. They are thought to indicate earthquake nucleation and hence may have the potential for short-term earthquake prediction (Scholz, 2002). However, proof of the physical link between foreshores and earthquake nucleation is inconclusive, and earthquake prediction using foreshocks has not been reliable.

The problem begins with their recognition – foreshocks are earthquakes that are called foreshocks retrospectively, when a large event (the mainshock) followed. To be considered as foreshocks, these events need to occur within certain spatial and temporal windows, and these windows vary in different studies. Foreshocks are typically located within the mainshock's source area with a dimension roughly the length of the fault rupture, which is in the order of 100 km for a magnitude 7.0 earthquake. Temporally, most foreshocks occur a few days to hours before the mainshock (Jones and Molnar, 1979). In some cases, seismicity increases a few months or years before the mainshock. Some workers consider such increased seismicity foreshocks; others do not (Scholz, 2002). Depending on the choice of these spatial and temporal windows, a global survey shows that up to 70% of large earthquakes ($M \geq 7$) may be preceded by foreshocks (Jones and Molnar, 1979). While the percentage varies with different studies and in different regions, it is important to note that not all earthquakes are preceded by foreshocks.

Characteristics of foreshocks

Foreshocks, when they do occur, are typically few in comparison with aftershocks, sometimes consisting of only a few or a single event, hence are difficult for quantitative analysis. When considered collectively, they are shown to exhibit a temporal pattern similar to that of aftershocks (Jones and Molnar, 1979; Shaw, 1993). Most foreshocks occur within 10 days before the mainshock, and their numbers rapidly accelerate till the occurrence of the

mainshock. The collective time sequence of foreshocks seems to fit an inverse Omori's law:

$$n(t) = at^{-q}$$

where $n(t)$ is the number of foreshocks by time t before the origin of the mainshock, and a and q are constants, with q close to 1 (Kagan and Knopoff, 1978; Papazachos, 1975). Thus, the number of foreshocks tends to increase hyperbolically as the time approaches the origin of the mainshock. Some studies suggest that foreshocks may have lower b-values (see entry *Earthquake, Aftershocks*) relative to other earthquakes (Scholz, 2002).

It is not clear why some earthquakes have foreshocks while others do not. Studies of global and regional earthquake catalogs have found that neither the time sequences nor the largest magnitude of foreshocks correlate to the magnitude of the mainshock (Jones and Molnar, 1979), suggesting either the size of the mainshock is independent of the earthquake nucleation process, or foreshocks are not part of earthquake nucleation (Abercrombie and Mori, 1996). The incidence of foreshocks is found to decrease with increasing depth of the mainshock (Abercrombie and Mori, 1996). For earthquakes in California, more foreshocks are associated with strike-slip events than with thrust events (Abercrombie and Mori, 1996). This is opposite to the results from a study of global catalogs by Reasenberg (1999), who attribute the low rate of foreshocks for Californian thrust events to their relatively greater depths.

Causes of foreshocks

The causes of foreshocks remain uncertain. Some studies link foreshocks to earthquake nucleation (Scholz, 2002). Theoretically, fault ruptures only when slip has occurred over a fault patch of some critical radius. Ohnaka (1992) has proposed a theoretical model that attributes the generation of foreshock to failure of asperities loaded by accelerating premonitory creep. This model predicts numerous features found in foreshocks: their occurrence typically a few days before the mainshock, their accelerated rates as time approaches that of the mainshock, and their occurrence in the immediate vicinity of the hypocenter of the mainshock. However, not all foreshocks readily fit the nucleation model. Aside from the obvious problem that many earthquakes do not have foreshocks, the nucleation model has difficulties explaining foreshocks that occur away from the rupture plane or have significantly different focal mechanisms from that of the mainshock (Jones et al., 1982).

Others think that foreshocks are fundamentally the same as other earthquakes (Felzer et al., 2004; Helmstetter et al., 2003). In this view, foreshocks, the mainshock, and aftershocks are just different names for earthquakes within the same sequence of cascading ruptures triggered by a common mechanism.

Can foreshocks be used to predict earthquakes?

As the most obvious precursor of earthquakes, foreshocks are thought to be useful for short-term prediction of earthquakes, and successful predictions have been reported in numerous cases. The best example is perhaps the 1975 Haicheng earthquake (magnitude 7.3) in China, which was predicted largely based on foreshocks, more than 500 of which were recorded within 4 days before the mainshock (Jones et al., 1982). However, a recent investigation (Wang et al., 2006) has revealed that the role of foreshocks in this prediction was more psychological than scientific: the jolts and damages from increased seismicity in the preceding months stressed the minds of earthquake workers and the general public, and the intensified foreshocks in the last day before the mainshock prompted some local officials to order an evacuation. In other places where official orders were not issued, the increased seismicity caused many residents to evacuate voluntarily.

Aside from the problem that many earthquakes do not have foreshocks, short-term earthquake prediction using foreshocks is challenging in theory and in practice. For foreshocks to be a predictor, they must be an essential part of the physical process leading to the mainshock. Although this physical link between foreshocks and mainshock is suggested in the nucleation model, its validation in natural fault zones has not been conclusive. On the other hand, if foreshocks are fundamentally the same as other earthquakes (Felzer et al., 2004), then foreshocks cannot be predictors of mainshocks. In practice, we have yet to find any reliable ways to recognize foreshocks from background seismicity before the occurrence of the mainshock.

Nonetheless, there must be some symptoms associated with the stress buildup and physical property changes in and around the fault leading to a major rupture, and foreshocks are the most obvious symptoms known to us. Further studies of foreshocks will improve our understanding of the mechanics of earthquakes. A recent study of the foreshock sequences on East Pacific Rise transform faults (McGuire et al., 2005) suggests that foreshocks are linked to the mainshock through stress changes driven by aseismic slip transients or some fault preparation process, hence potentially can be useful for short-term earthquake prediction. On the other hand, by treating every earthquake as a potential foreshock that may be followed by a mainshock, statistical models based on the clustering properties of earthquake sequences can be developed to assess the real-time probability of damaging earthquakes, as has been done in California (Gerstenberger et al., 2005; Jones, 1985; Reasenberg and Jones, 1989).

Summary

Foreshocks are smaller earthquakes preceding a mainshock. They are thought to manifest the nucleation of the

mainshock, hence have the potential for short-term earthquake prediction. However, not all large earthquakes are preceded by foreshocks, and they are difficult to recognize from the background seismicity before the occurrence of the mainshock. Foreshocks may also be fundamentally the same as other earthquakes, if so they cannot be used as earthquake predictors.

Bibliography

Abercrombie, R. E., and Mori, J., 1996. Occurrence patterns of foreshocks to large earthquakes in the western United States. *Nature*, **381**, 303–307.

Dodge, D. A., Beroza, G. C., and Ellsworth, W. L., 1995. Foreshock sequence of the 1992 Landers, California, earthquake and its implications for earthquake nucleation. *Journal of Geophysical Research*, **100**, 9865–9880.

Felzer, K. R., Abercrombie, R. E., and Ekström, G., 2004. A common origin for aftershocks, foreshocks, and multiplets. *Bulletin of the Seismological Society of America*, **94**, 88–98.

Gerstenberger, M. C., Wiemer, S., Jones, L. M., and Reasenberg, P. A., 2005. Real-time forecasts of tomorrow's earthquakes in California. *Nature*, **435**, 328–331.

Helmstetter, A., Sornette, D., and Grasso, J.-R., 2003. Mainshocks are aftershocks of conditional foreshocks: how do foreshock statistical properties emerge from aftershock laws. *Journal of Geophysical Research*, **108**, 2046.

Jones, L. M., 1985. Foreshocks and time-dependent earthquake hazard assessment in southern California. *Bulletin of the Seismological Society of America*, **75**, 1667–1679.

Jones, L. M., and Molnar, P., 1979. Some characteristics of foreshocks and their possible relationship to earthquake prediction and premonitory slip on faults. *Journal of Geophysical Research*, **84**, 3596–3608.

Jones, L. M., Wang, B., Xu, S., and Fitch, T. J., 1982. The foreshock sequence of the February 4, 1975, Haicheng earthquake (M = 7.3). *Journal of Geophysical Research*, **87**, 4575–4584.

Kagan, Y. Y., and Knopoff, L., 1978. Statistical study of the occurrence of shallow earthquakes. *Geophysical Journal of the Royal Astronomical Society*, **55**, 67–86.

McGuire, J. J., Boettcher, M. S., and Jordan, T. H., 2005. Foreshock sequences and short-term earthquake predictability on East Pacific Rise transform faults. *Nature*, **434**, 457–461.

Ohnaka, M., 1992. Earthquake source nucleation: a physical model for short-term precursors. *Tectonophysics*, **211**, 149–178.

Papazachos, B. C., 1975. Foreshocks and earthquake prediction. *Tectonophysics*, **28**, 213–226.

Reasenberg, P. A., 1999. Foreshock occurrence before large earthquakes. *Journal of Geophysical Research*, **104**, 4755–4768.

Reasenberg, P. A., and Jones, L. M., 1989. Earthquake hazard after a mainshock in California. *Science*, **243**, 1173–1176.

Scholz, C. H., 2002. *The Mechanics of Earthquakes and Faulting*. Cambridge/New York: Cambridge University Press, Vol. xxiv. 471 pp.

Shaw, B. E., 1993. Generalized Omori law for aftershocks and foreshocks from a simple dynamics. *Geophysical Research Letters*, **20**, 907–910.

Wang, K., Chen, Q. F., Sun, S., and Wang, A., 2006. Predicting the 1975 Haicheng earthquake. *Bulletin of the Seismological Society of America*, **96**, 757–795.

Cross-references

Earthquake, Aftershocks

EARTHQUAKE, LOCATION TECHNIQUES

Clifford H. Thurber
Department of Geoscience, University of Wisconsin-Madison, Madison, WI, USA

Synonyms

Hypocenter determination

Definition

Earthquake location. Methods for determining the latitude, longitude, depth, and time of origin of a seismic event (earthquake, explosion, etc.) using the arrival times of seismic waves.

Introduction

Knowing the location of an earthquake (its latitude, longitude, depth, and origin time) is an essential starting point for the vast majority of quantitative seismological analyses. In the context of this article, earthquake location will be taken to mean determining the initiation point of fault rupture, that is, the hypocenter. After briefly tracing some of the early history of methods for the quantitative determination of earthquake locations, the focus of discussion turns to modern methods for single-event location and uncertainty estimation. The final section covers multiple-event location, with an emphasis on the substantial to – in some cases – profound improvement in relative location accuracy that can be achieved by combining waveform cross-correlation with relative location techniques.

Early history

John "Earthquake" Milne devised some of the first quantitatively based methods for earthquake location, known as the methods of circles, hyperbolas, and coordinates (Milne, 1886). For the method of circles, the difference in arrival time between later observing stations i ($i = 1, N$) and the first observing station 0, $t_i - t_0 = \mathrm{d}t_i$, is used to construct circles of radii $V\,\mathrm{d}t_i$, where V is the velocity of "shock" (i.e., seismic wave) propagation, and the center of the circle through station 0 that is roughly tangent to all the other circles defines the earthquake epicenter. The scale factor V is determined by trial and error. The commonly taught S minus P time location method (Bolt, 2006) is similar in concept. The method of hyperbolas is a variant on the method of circles, whereby differential times between pairs of stations are used to define hyperbolas of possible locations, and the intersection of hyperbolas defines the earthquake epicenter. The method of coordinates goes a step further by setting up a system of four equations (defined in Cartesian coordinates centered on the last-observing station) relating differential times to "shock" velocity and the (x,y,z) coordinates of the earthquake. With four equations and four unknowns, the solution is unique. These methods

were employed for many years, until slowly the method of Geiger (1910, 1912), discussed below, began to supplant it. Interestingly, Milne's methods have been revived recently for use in locating earthquakes in an early warning system (Satriano et al., 2008) and in tectonic (Zhou, 1994; Font et al., 2004) and volcanic (Sumiejski et al., 2009) settings.

Single-event location

The predominant method for earthquake location is based on what is commonly known as Geiger's method, which was developed early in the twentieth century (Geiger, 1910, 1912). This method involves deriving the least-squares (L2 norm) solution to a set of equations, for a trial hypocenter, of the form

$$r_i = \sum_{j=1}^{3} \frac{\partial T_i}{\partial x_j} \Delta x_j + \Delta t_o \tag{1}$$

where r_i is the residual at station i, defined as the observed arrival time of a seismic phase t_i minus the calculated arrival time given by the origin time t_o plus the calculated travel time T_i, the x_j's are the hypocenter coordinates, and Δ indicates the (unknown) parameter perturbations. The spatial derivatives $\partial T_i/\partial x_j$ have a simple relationship to the seismic velocity V and the ray direction at the source,

$$\frac{\partial T_i}{\partial x_j} = \frac{-1}{V} \frac{\mathrm{d}x_j}{\mathrm{d}s} \tag{2}$$

where s is the path length parameter. Equation 1 represents a nonlinear parameter estimation problem (Aster et al., 2005) so it is solved iteratively, and some convergence criteria must be defined. Mathematically, this problem can be expressed as

$$\mathbf{G}\Delta\mathbf{m} = \mathbf{r} \tag{3}$$

where the matrix $\mathbf{G}$ contains the partial derivatives, $\mathbf{\Delta m}$ is the vector of hypocenter perturbations (position and origin time), and $\mathbf{r}$ is the residual vector. Equation 3 is typically solved using iterative least squares, generally with some type of regularization (Aster et al., 2005). For a layered model, Geiger's method can be extended from this simple linearized form to a nonlinear form (Thurber, 1985), but that is of more theoretical than practical interest.

The seismic phases used for location vary depending in part on the epicentral distance domain – local (0–5°), regional (5–30°), and teleseismic (>30°) (note that these breakpoints are relatively standard ones, although that for the local–regional is somewhat more arbitrary). At local distances, first-arriving P and S phases are used almost to exclusion. At teleseismic distances, first-P and to a lesser degree first-S remain critically important, albeit with such complications as shadow zones and triplications, but a virtual cornucopia of other phases is also available. Among them, the depth phases (solid-surface reflections pP and sP and the water-surface reflections pwP and swP), which have travel-time derivatives for depth opposite in sign to the direct phases, have tremendous value in constraining event depth, although misidentification problems can be significant (Engdahl et al., 1998; Engdahl, 2006). The outer core reflection PcP has also proven quite useful (Schöffel and Das, 1999).

At regional distances, several factors combine to make the location situation complex. The mantle refractions Pn and Sn are the first-arriving body-wave phases, and their amplitudes can be small relative to crustal arrivals. Crustal thickness variations at a regional scale make accurate prediction of regional travel times difficult. These problems have an enormous impact on nuclear explosion monitoring efforts, where the goal is to be able to locate relatively small events often with only sparse regional data. The use of later arriving phases (crustal body-wave phases Pg, Sg, and PmP, and the surface-wave phase Lg) and also arrival azimuth is relatively common for this monitoring domain (Bratt and Bache, 1988). Arrival azimuth can be determined at high precision with a seismic array, or by a single three-component station with much lower precision.

There are some alternatives to Geiger's method, most of which fall into the general category of nonlinear optimization (Gill et al., 1981). One is the conceptually simple grid search approach (Rowlett and Forsyth, 1984; Sambridge and Kennett, 1986; Shearer, 1997; Lomax et al., 2000; Rodi, 2006). Others include the simplex method (Rabinowitz, 1988), genetic algorithms (Sambridge and Gallagher, 1993), and simulated annealing (Billings, 1994). Because these methods do not involve least squares and the calculation of derivatives (Equations 2 and 3), they are readily amenable to the use of the more robust L1 norm (Aster et al., 2005), and some can directly provide comprehensive uncertainty information, a topic we turn to next.

Location uncertainty

In some circumstances, knowing the uncertainty of a seismic event's location may be nearly as important as the location itself. The discussion of this topic can be separated into two parts: precision and accuracy. Precision is covered first, including a brief discussion of arrival-time estimation (phase picking). Accuracy is a far more subtle issue, with significant limitations on our ability to assess it thoroughly.

The standard approach for the evaluation of precision is the confidence ellipsoid (Flinn, 1965), which can be obtained from the model covariance matrix (Aster et al., 2005) for Equation 3,

$$\mathbf{C}_\mathrm{m} = \sigma_\mathrm{d}{}^2 (\mathbf{G}^\mathrm{T}\mathbf{G})^{-1} [(\mathbf{G}^\mathrm{T}\mathbf{G})^{-1}]^\mathrm{T} \tag{4}$$

where for now we assume that the data errors have a uniform standard deviation σ_d, often estimated from the data misfit, and are uncorrelated (independent). Given the desired confidence level p (e.g., 0.95), the corresponding confidence ellipsoid is obtained for $\mathbf{m}$ satisfying

$$(\mathbf{m}-\mathbf{m}^*)^{\mathrm{T}}\mathbf{C_m}^{-1}(\mathbf{m}-\mathbf{m}^*)=\kappa_p{}^2 \tag{5}$$

where $\mathbf{m}^*$ is the location solution and $\kappa_p{}^2$ is given by

$$\kappa_p{}^2=4\sigma_d{}^2\left[F_p(4,N-4)\right]/(N-4) \tag{6}$$

with F_p the F statistic at confidence level p and N the number of arrival-time data (Draper and Smith, 1981). Some limitations of this approach to uncertainty estimation are (1) that the misfit-based estimate of σ_d becomes unreliable for small numbers of observations (Evernden, 1969), (2) that it neglects the effects of velocity-model inaccuracy and correlated errors, and (3) that the linear approximation itself is inadequate in some circumstances.

Buland (1976) noted that data uncertainty estimated from either the misfit or a priori experience is generally too small due to the inaccuracy of the travel-time model. He suggested that the σ_d value should simply be increased to compensate for this problem. Jordan and Sverdrup (1981) developed an elegant statistical approach for making this compensation. Their philosophy is to replace the usual a posteriori estimate of the data variance, $s_d{}^2$, based on the data misfit for an individual earthquake,

$$s_d{}^2=||\mathbf{r}||^2/(N-4) \tag{7}$$

with an estimate based on a combination of a priori and a posteriori information:

$$s_d{}^2=\frac{Ks_0{}^2+\|\mathbf{r}\|^2}{(K+N-4)} \tag{8}$$

where the a priori variance estimate s_0 is given a relative weight of K. Thus, the a priori information is treated as contributing K observations to the estimate of $s_d{}^2$ and the data for the particular event contribute N observations. For self-consistency, given a large number of events, the mean of the estimated variances should approach s_0. Jordan and Sverdrup (1981) also show that the distribution of the estimated standard deviation needs to be consistent with the adopted value of K. In the limit of no prior information ($K = 0$), their estimate reverts to the standard formula (Flinn, 1965), whereas for infinite prior information ($K = \infty$), their estimate becomes equivalent to that of Evernden (1969).

One limitation of the Jordan and Sverdrup (1981) approach is that it assumes an uncorrelated, Gaussian error distribution with a constant data variance for all observations. Varying data uncertainty estimates can be incorporated simply by including a weighting matrix $\mathbf{W}$ in the standard linearized analysis (Aster et al., 2005):

$$\mathbf{WG}\Delta\mathbf{m}=\mathbf{Wr} \tag{9}$$

(e.g., Bolt, 1960) and then solving the weighted system as before. Weighting is generally done based on reported arrival-time reading precision (Engdahl et al., 1998), reported arrival quality (Buland, 1976), phase variance as a function of distance (Engdahl, 2006), residual size (Bolt, 1960), station distance (Klein, 1978), and/or azimuthal distribution of stations (Buland, 1976). However, the Gaussian data error assumption is generally inappropriate, as residual outliers are common in arrival-time data (Bolt, 1960; Buland, 1986). Discussions of correlated errors and their treatment can be found in Chang et al. (1983), Rodi and Myers (2007), and Bondár and McLaughlin (2009).

The "uniform reduction" method (Jeffreys, 1939) is effective for treating the situation of normally distributed errors plus a background of outliers. Bolt (1960) assumed a residual frequency distribution of the form

$$f(r)=as+\frac{(1\pm s)}{\sigma\sqrt{2\pi}}\exp\left(\pm\frac{(r\pm m)^2}{2\sigma^2}\right) \tag{10}$$

where a and s are constants and m and σ are the mean residual and its standard deviation. This corresponds to a residual weighting function of the form

$$w(r)=\left[1+\mu\exp\left(\pm\frac{(r\pm m)^2}{2\sigma^2}\right)\right]^{\pm1} \tag{11}$$

with $\mu=\sigma\,(2\pi)^{1/2}\,a\,s/(1-s)$.

It is important to point out that estimating the phase arrival time by itself is a difficult problem, particularly for S waves and other secondary phases. Many seismic networks use automated pickers, at least for preliminary locations. Simple auto-pickers use a threshold of the ratio of the short-term average (STA) to the long-term average (LTA) of the signal amplitude (Allen, 1978, 1982; Ruud and Husebye, 1992). There are a wide variety of other (generally more computationally demanding) approaches. Among these are energy analysis (Earle and Shearer, 1994), polarization analysis (Vidale, 1986; Cichowicz, 1993), autoregressive techniques (Maeda, 1985; Sleeman and Eck, 1999; Leonard and Kennett, 1999; Leonard, 2000), wavelet analysis (Zhang et al., 2003), or some combination (Wang and Teng, 1997; Diehl et al., 2009).

As noted above, constraints on event locations can be improved markedly via the inclusion of secondary arrivals. Gomberg et al. (1990) and Wilcock and Toomey (1991) examine the importance of even a single S arrival in reducing location uncertainty. A "chicken and egg" problem with the use of later phases is that their correct identification often requires knowledge of the event depth and distance. Recognizing this problem, Kennett et al. (1995) and Engdahl et al. (1998) included a phase re-association step in their location procedure. In their approach, catalog phase arrivals were reidentified after each relocation iteration using a statistically based association algorithm. This approach can be implemented by comparing probability density functions (PDF's) for relevant phases, centered on their theoretical relative travel times for a given hypocenter, to the observed phase arrivals. When PDF's overlap for a particular phase, the idea is to assign a phase identification in a probabilistic manner based on the relevant PDF values, making sure not to assign the same phase to two different arrivals.

The above discussion focuses almost exclusively on location precision. There are two general approaches for evaluating location accuracy: theoretical and empirical. Pavlis (1986) carried out a thorough theoretical analysis of both precision and accuracy aspects for local earthquakes. He identified and evaluated three sources of hypocenter error: (1) measurement errors, (2) velocity-model errors, and (3) nonlinearity. He treated the first via standard confidence ellipsoids, but the other two required the establishment of error bound estimates. Billings et al. (1994) explored aspects of (1) and (2) for the teleseismic case using a combination of Monte Carlo analysis and varying sets of observations, although their analyses involved real data, not synthetic data. One general conclusion of both these studies is that velocity-model-induced errors can be as significant as measurement errors.

"Ground truth" (GT) events provide a direct, empirical means to explore absolute location accuracy (also termed location bias). Having GT events allows one to evaluate location and uncertainty estimation strategies empirically that, for example, produce 95% confidence ellipses that overlap the true locations 95% of the time. This avenue of research has been motivated primarily by nuclear explosion monitoring needs (see *Seismic Monitoring of Nuclear Explosions*), but also is an important aspect of research on one-dimensional (1D) and three-dimensional (3D) earth models (see *Seismic Tomography*). See Bondár and McLaughlin (2009) and references therein for recent work on absolute location accuracy and Thurber and Engdahl (2000) for a review of the development of global 1D models.

The San Andreas Fault Observatory at Depth (SAFOD) provided a unique test of absolute location accuracy. One of SAFOD's goals (unfortunately not yet realized) was to penetrate the rupture zone of a magnitude ~2 repeating earthquake. This required determining a borehole trajectory that would pass within 100 m of the target event. Using data from a dense temporary array, permanent networks, deep borehole instruments, and active-source experiments, Thurber et al. (2004) carried out 3D tomography and used "virtual earthquakes" (receiver gathers of shots recorded on borehole instruments, effectively providing sources of known location at depth) to assess absolute location accuracy. Their work succeeded, as subsequent borehole recordings of target earthquake aftershocks showed the target event to lie within 100 m of the borehole.

Multiple-event location

There are two distinct but complementary aspects to this topic that will be discussed. One is the suite of methods used to infer the relative locations of events, including joint hypocenter determination (JHD), hypocentroidal decomposition (HDC), and double-difference (DD) methods. Note that local earthquake tomography takes multiple-event location a step further to determine seismic velocity structure simultaneously with the locations of many earthquakes (see *Seismic Tomography*). The other is the use of waveform cross-correlation (WCC) methods to determine differential times (DT's) and/or pick adjustments.

The use of catalog DT's for determining relative locations of earthquakes (or explosions) has been a fruitful field of work for many decades (e.g., Douglas, 1967). The precision of WCC DT's allows for spectacular relative location results to be achieved over a broad range of spatial scales when DT's from WCC are used in combination with joint location techniques. These and other studies have demonstrated the substantial improvement in the delineation of seismogenic features or in the accuracy of relative locations of GT events that is possible using multiple-event location methods in concert with high-precision arrival-time or differential-time data. WCC is covered first to streamline the subsequent discussion of the relative location techniques.

Waveform cross-correlation

For a pair of earthquakes that are located close to each other (relative to the predominant wavelength of the seismic signal from them) and have similar source mechanisms, the waveforms from them observed at a particular seismic station will be similar. The work by Poupinet et al. (1984) to use (frequency-domain) WCC on the coda of similar earthquakes ("doublets") to detect temporal changes in seismic velocity followed by the study of Fremont and Malone (1987) using the same WCC method to determine high-precision relative arrival times and earthquake locations at Mt. St. Helens ultimately led to an explosion in the use of WCC and location techniques in the 1990s and beyond.

The time-domain method for WCC is relatively simple, although computationally demanding for large sets of events. Windowed waveform segments spanning the time of the relevant arrival (P or S) for a pair of earthquakes at a particular station are cross-correlated over a range of lag times to find the time lag corresponding to the best alignment of the waveforms (i.e., the maximum CC value). Waveforms are typically filtered either with a simple bandpass filter or with coherency weighting (Rowe et al., 2002a). For very similar events (those with a high CC value), it can be worthwhile to apply a frequency-domain technique such as cross-spectral phase to determine the lag to a precision below the sample rate. This procedure takes advantage of the fact that, as an example, for two frequencies that are a factor of two different, the phase difference at the higher frequency should be exactly double that at the lower frequency. The lag is determined by calculating the phase difference as a function of frequency and fitting the result with a straight line through the origin at zero frequency.

Most of the studies using WCC adopt a threshold value (typically 0.7) below which DT's are discarded. This can result in good data being lost (correct lag value but low CC value – e.g., due to a noisy record) as well as bad data

being used (incorrect lag value but high CC value – e.g., due to a cycle slip). Du et al. (2004a) developed a procedure to provide independent lag estimates for verification of WCC results. They used the bispectrum (BS) method, which works in the third-order spectral domain, to check the reliability of the CC-determined time delay. They calculated two time delays with the BS method, one using the bandpass-filtered waveforms and the other with the raw data, and used them to verify (select or reject) the CC estimate computed with the filtered waveforms. This approach appears to be particularly effective in reducing cycle-slip errors.

Relative location

As noted above, the JHD method dates back to the work of Douglas (1967). In its basic form, the method takes the basic equation for earthquake location, Equation 1, adds a "station correction" term Δs_i, and sets up a large set of equations for a suite of events, $k = 1, K$:

$$r_{ik} = \sum_{j=1}^{3} \frac{\partial T_{ik}}{\partial x_{jk}} \Delta x_{jk} + \Delta t_{ok} + \Delta s_i \tag{12}$$

As noted by Wolfe (2002), JHD procedures assume that if the hypocentral separations among a set of earthquakes are small compared to the event-station distance and the scale length of heterogeneity, then the ray paths between the source region and a station are similar and the travel-time error introduced by an incorrect velocity model has a nearly constant value for a given station, thus justifying the linear term Δs_i. Pujol (2000) provides a review of JHD methods mostly aimed at local-scale studies. Pavlis (1992) presents a comprehensive analysis of relative location errors.

Wolfe (2002) thoroughly analyzed the HDC and DD methods. Her analysis shows that the solutions for both operators are similar, with the double-difference operator weighting fits to travel times on stations that record more earthquakes more heavily. The HDC method was introduced by Jordan and Sverdrup (1981). As the name indicates, it involves decomposing the system Equation 12 via projection operators into one involving just the locations of the events relative to the (unknown) average location (i.e., the hypocentroid). As Wolfe (2002) notes, the projection is in fact a travel-time differencing operation. The DD method of Waldhauser et al. (1999) and Waldhauser and Ellsworth (2000) constructs the event-pair difference equations directly, but it is distinguished from the original Jordan and Sverdrup (1981) method by evaluating the spatial derivatives at different points rather than assuming that they are constant and by including a separation-dependent weighting in order to reduce location-dependent path bias. Modern HDC methods also allow for spatially varying derivatives (Bergman and Engdahl, 2001), and the appropriateness of separation-dependent weighting has been subject to debate. The Markov-Chain Monte Carlo approach of Myers et al. (2007) takes a probabilistic approach that incorporates the probability of phase identification errors in the estimation of uncertainty.

Some examples of applications of WCC combined with relative location for microearthquakes at a local-scale include: Switzerland (Deichmann and Garcia-Fernandez, 1992), Kilauea volcano, Hawaii (Got et al., 1994; Rubin et al., 1998), California (Dodge et al., 1995; Shearer, 1997; Waldhauser et al., 1999; Waldhauser and Ellsworth, 2000; Schaff et al., 2002), Iceland (Slunga et al., 1995), the Soultz geothermal field, France (Rowe et al., 2002b), Mount Pinatubo, the Philippines (Battaglia et al., 2004), New Zealand (Du et al., 2004b), and Soufriere Hills volcano, Montserrat (Rowe et al., 2004). At a regional scale, Schaff and Richards (2004) determined relative locations for an earthquake sequence in China using WCC on Lg waves. For teleseismic data, WCC/relative location methods have been applied to body waves for nuclear explosions at the Balapan test site (Phillips et al., 2001; Thurber et al., 2001), to the aftershock sequences of the great Sumatra earthquakes (Pesicek et al., 2010), and to global seismicity (Waldhauser and Schaff, 2008).

Summary

Despite being one of the oldest quantitative problems in seismology, it is clear that there is still much to learn about earthquake location. The combination of WCC and high-precision relative location has brought new excitement to this rather old field of research. At the same time, the needs of nuclear explosion monitoring have motivated the careful study of absolute location capability and statistically consistent uncertainty estimation.

Bibliography

Allen, R. V., 1978. Automatic earthquake recognition and timing from single traces. *Bulletin of the Seismological Society of America*, **68**, 1521.

Allen, R. V., 1982. Automatic phase pickers: their present use and future prospects. *Bulletin of the Seismological Society of America*, **72**, S225.

Aster, R., Borchers, B., and Thurber, C., 2005. *Parameter Estimation and Inverse Problems*. London: Elsevier/Academic.

Battaglia, J., Thurber, C., Got, J.-L., Rowe, C., and White, R., 2004. Precise relocation of earthquakes following the June 15, 1991 explosion of Mount Pinatubo (Philippines). *Journal of Geophysical Research*, **109**, B07302.

Bergman, E. A., and Engdahl, E. R., 2001. Validation and generation of reference events by cluster analysis. *Seismological Research Letters*, **72**, 3.

Billings, S. D., 1994. Simulated annealing for earthquake location. *Geophysical Journal International*, **118**, 680.

Billings, S. D., Sambridge, M. S., and Kennett, B. L. N., 1994. Errors in hypocenter location: picking, model, and magnitude dependence. *Bulletin of the Seismological Society of America*, **84**, 1978.

Bolt, B. A., 1960. The revision of earthquake epicentres, focal depths and origin times using a high-speed computer. *Geophysical Journal of the Royal Astronomical Society*, **3**, 433.

Bolt, B. A., 2006. *Earthquakes*. New York: W.H Freeman.

Bondár, I., and McLaughlin, K., 2009. A new ground truth data set for seismic studies. *Seismological Research Letters*, **80**, 465.

Bratt, S. R., and Bache, T. C., 1988. Locating events with a sparse network of regional arrays. *Bulletin of the Seismological Society of America*, **78**, 780–798.

Buland, R., 1976. The mechanics of locating earthquakes. *Bulletin of the Seismological Society of America*, **66**, 173.

Buland, R., 1986. Uniform reduction error analysis. *Bulletin of the Seismological Society of America*, **76**, 217.

Chang, A. C., Shumway, R. H., Blandford, R. R., and Barker, B. W., 1983. Two methods to improve location estimates – preliminary results. *Bulletin of the Seismological Society of America*, **73**, 281.

Cichowicz, A., 1993. An automatic S-phase picker. *Bulletin of the Seismological Society of America*, **83**, 180.

Deichmann, N., and Garcia-Fernandez, M., 1992. Rupture geometry from high precision relative hypocentre locations of microearthquake clusters. *Geophysical Journal International*, **110**, 501–517.

Diehl, T., Deichmann, N., Kissling, E., and Husen, S., 2009. Automatic S-wave picker for local earthquake tomography. *Bulletin of the Seismological Society of America*, **99**, 1906.

Dodge, D. A., Beroza, G. C., and Ellsworth, W. L., 1995. Foreshock sequence of the 1992 Landers, California, earthquake and its implications for earthquake nucleation. *Journal of Geophysical Research*, **100**, 9865.

Douglas, A., 1967. Joint epicenter determination. *Nature*, **215**, 47.

Draper, N. R., and Smith, H., 1981. *Applied Regression Analysis*. New York: Wiley.

Du, W., Thurber, C. H., and Eberhart-Phillips, D., 2004a. Earthquake relocation using cross-correlation time delay estimates verified with the bispectrum method. *Bulletin of the Seismological Society of America*, **94**, 856.

Du, W., Thurber, C. H., Reyners, M., Eberhart-Phillips, D., and Zhang, H., 2004b. New constraints on seismicity in the Wellington region, New Zealand, from relocated earthquake hypocenters. *Geophysical Journal International*, **158**, 1088.

Earle, P., and Shearer, P. M., 1994. Characterization of global seismograms using an automatic-picking algorithm. *Bulletin of the Seismological Society of America*, **84**, 366.

Engdahl, E. R., 2006. Application of an improved algorithm to high precision relocation of ISC test events. *Physics of the Earth and Planetary Interiors*, **158**, 14.

Engdahl, E. R., van der Hilst, R., and Buland, R., 1998. Global teleseismic earthquake relocation with improved travel times and procedures for depth determination. *Bulletin of the Seismological Society of America*, **88**, 722.

Evernden, J. F., 1969. Precision of epicenters obtained by small numbers of world-wide stations. *Bulletin of the Seismological Society of America*, **59**, 1365.

Flinn, E. A., 1965. Confidence regions and error determinations for seismic event location. *Reviews of Geophysics*, **3**, 157.

Font, Y., Honn, K., Lallemand, S., Liu, C., and Chiao, L., 2004. Hypocentre determination offshore of eastern Taiwan using the Maximum Intersection method. *Geophysical Journal International*, **158**, 655.

Fremont, M.-J., and Malone, S. D., 1987. High-precision relative location of earthquakes at Mount St. Helens, Washington. *Journal of Geophysical Research*, **92**, 10223.

Geiger, L., 1910. Herdbestimmung bei erdbeden ans den ankunftzeiten. *Königlichen Gesellschaft der Wissenschaften Zu Göttingen*, **4**, 331.

Geiger, L., 1912. Probability method for the determination of earthquake epicenters from the arrival time only. *Bulletin of St. Louis University*, **8**, 60.

Gill, P. E., Murray, W., and Wright, M. H., 1981. *Practical Optimization*. New York: Academic.

Gomberg, J. S., Shedlock, K. M., and Roecker, S. W., 1990. The effect of S-wave arrival times on the accuracy of hypocenter estimation. *Bulletin of the Seismological Society of America*, **80**, 1605.

Got, J.-L., Frechet, J., and Klein, F., 1994. Deep fault plane geometry inferred from multiplet relative relocation beneath the south flank of Kilauea. *Journal of Geophysical Research*, **99**, 15375.

Jeffreys, H., 1939. *Theory of Probability.* London: Oxford University Press.

Jordan, T. H., and Sverdrup, K. A., 1981. Teleseismic location techniques and their application to earthquake clusters in the south-central Pacific. *Bulletin of the Seismological Society of America*, **71**, 1105.

Kennett, B. L. N., Engdahl, E. R., and Buland, R., 1995. Constraints on seismic velocities in the Earth from traveltimes. *Geophysical Journal International*, **122**, 108.

Klein, F. W., 1978. Hypocenter Location Program HYPOINVERSE. *U.S. Geological Survey Open-File Report 78–694,* 113 pp.

Leonard, M., 2000. Comparison of manual and automatic onset time picking. *Bulletin of the Seismological Society of America*, **90**, 1384.

Leonard, M., and Kennett, B. L. N., 1999. Multi-component autoregressive techniques for the analysis of seismograms. *Physics of the Earth and Planetary Interiors*, **113**, 247.

Lomax, A., Virieux, J., Volant, P., and Berge, C., 2000. Probabilistic earthquake location in 3D and layered models: introduction of a Metropolis–Gibbs method and comparison with linear locations. In Thurber, C. H., and Rabinowitz, N. (eds.), *Advances in Seismic Event Location*. Amsterdam: Kluwer, pp. 101–134.

Maeda, N., 1985. A method for reading and checking phase times in autoprocessing system of seismic wave data. *Zisin-Jishin*, **38**, 365.

Milne, J., 1886. *Earthquakes and Other Earth Movements*. New York: Appleton.

Myers, S. C., Johannesson, G., and Hanley, W., 2007. A Bayesian hierarchical method for multiple-event seismic location. *Geophysical Journal International*, **171**, 1049.

Pavlis, G. L., 1986. Appraising earthquake hypocenter location errors: a complete, practical approach for single event locations. *Bulletin of the Seismological Society of America*, **76**, 1699.

Pavlis, G. L., 1992. Appraising relative earthquake location errors. *Bulletin of the Seismological Society of America*, **82**, 836.

Pesicek, J. D., Thurber, C. H., Zhang, H., DeShon, H. R., Engdahl, E. R., and Widiyantoro, S., 2010. Teleseismic double-difference relocation of earthquakes along the Sumatra-Andaman subduction zone with a three-dimensional model. *Journal of Geophysical Research*, **115**, doi:10.1029/2010JB007443.

Phillips, W. S., Hartse, H. E., and Steck, L. K., 2001. Precise relative location of 25 ton chemical explosions at Balapan using IMS stations. *Pure and Applied Geophysics*, **158**, 173.

Poupinet, G., Ellsworth, W. L., and Fréchet, J., 1984. Monitoring velocity variations in the crust using earthquake doublets: an application to the Calaveras fault. California. *Journal of Geophysical Research*, **89**, 5719.

Pujol, J., 2000. Joint event location: the JHD technique and applications to data from local seismic networks. In *Advances in Seismic Event Location.* Dordrecht: Kluwer, pp. 71–99.

Rabinowitz, N., 1988. Microearthquake location by means of nonlinear simplex procedure. *Bulletin of the Seismological Society of America*, **78**, 380.

Rodi, W. L., 2006. Grid-search event location with non-Gaussian error models. *Physics of the Earth and Planetary Interiors*, **158**, 55.

Rodi, W. L., and Myers, S. C., 2007. Modeling travel-time correlations based on sensitivity kernels and correlated velocity anomalies. In *Proceedings of the 29th Monitoring Research Review:*

Ground-Based Nuclear Explosion Monitoring Technologies, Denver, Colorado, September 25–27.
Rowe, C. A., Aster, R. C., Borchers, B., and Young, C. J., 2002a. An automatic, adaptive algorithm for refining phase picks in large seismic data sets. *Bulletin of the Seismological Society of America*, **92**, 1660.
Rowe, C. A., Aster, R. C., Phillips, W. S., Jones, R. H., Borchers, B., and Fehler, M. C., 2002b. Using automated, high-precision repicking to improve delineation of microseismic structures at the Soultz geothermal reservoir. *Pure and Applied Geophysics*, **159**, 536.
Rowe, C. A., Thurber, C. H., and White, R. A., 2004. Dome growth behavior at Soufriere Hills volcano, Montserrat, revealed by relocation of volcanic event swarms, 1995–1996. *Journal of Volcanology and Geothermal Research*, **134**, 199.
Rowlett, H., and Forsyth, D. W., 1984. Microearthquakes and recent faulting at the intersection of the Vema Fracture Zone and the mid-Atlantic ridge. *Journal of Geophysical Research*, **89**, 6079.
Rubin, A. M., Gillard, D., and Got, J.-L., 1998. A reinterpretation of seismicity associated with the January 1983 dike intrusion at Kilauea volcano, Hawaii. *Journal of Geophysical Research*, **103**, 10003.
Rubin, A. M., Gillard, D., and Got, J., 1999. Streaks of microearthquakes along creeping faults. *Nature*, **400**, 635.
Ruud, B. O., and Husebye, E. S., 1992. A new three-component detector and automatic single station bulletin production. *Bulletin of the Seismological Society of America*, **82**, 221.
Sambridge, M., and Gallagher, K., 1993. Earthquake hypocenter location using genetic algorithms. *Bulletin of the Seismological Society of America*, **83**, 1467.
Sambridge, M. S., and Kennett, B. L. N., 1986. A novel method for hypocentre location. *Geophysical Journal of the Royal Astronomical Society*, **87**, 679.
Satriano, C., Lomax, A., and Zollo, A., 2008. Real-time evolutionary earthquake location for seismic early warning. *Bulletin of the Seismological Society of America*, **98**, 1482.
Schaff, D. P., and Richards, P. G., 2004. Repeating seismic events in China. *Science*, **303**, 1176.
Schaff, D. P., Bokelmann, G. H. R., Beroza, G. C., Waldhauser, F., and Ellsworth, W. L., 2002. High-resolution image of Calaveras Fault seismicity. *Journal of Geophysical Research*, **107**, 2186.
Schöffel, H.-J., and Das, S., 1999. Fine details of the Wadati–Benioff zone under Indonesia and its geodynamic implications. *Journal of Geophysical Research*, **104**, 13101.
Shearer, P. M., 1997. Improving local earthquake locations using the L1 norm and waveform cross correlation: application to the Whittier Narrows, California, aftershock sequence. *Journal of Geophysical Research*, **102**, 8269.
Sleeman, R., and van Eck, T., 1999. Robust automatic P-phase picking: an on-line implementation in the analysis of broadband seismogram recordings. *Physics of the Earth and Planetary Interiors*, **113**, 265.
Slunga, R., Rögnvaldsson, S. T., and Böðvarsson, R., 1995. Absolute and relative locations of similar events with application to microearthquakes in southern Iceland. *Geophysical Journal International*, **123**, 409.
Sumiejski, L., Thurber, C., and DeShon, H., 2009. Relocation of earthquake families associated with the 2006 eruption of Augustine volcano using the Equal Differential Time method. *Geophysical Journal International*, **176**, 1017.
Thurber, C. H., 1985. Nonlinear earthquake location: theory and examples. *Bulletin of the Seismological Society of America*, **75**, 779.
Thurber, C. H., and Engdahl, E. R., 2000. Advances in global seismic event location. In Thurber, C., and Rabinowitz, N. (eds.), *Advances in Seismic Event Location*. Dordrecht: Kluwer, pp. 3–22.
Thurber, C., Trabant, C., Haslinger, F., and Hartog, R., 2001. Nuclear explosion locations at the Balapan, Kazakhstan, nuclear test site: the effects of high-precision arrival times and three-dimensional structure. *Physics of the Earth and Planetary Interiors*, **123**, 283.
Thurber, C., Roecker, S., Zhang, H., Baher, S., and Ellsworth, W., 2004. Fine-scale structure of the San Andreas fault and location of the SAFOD target earthquakes. *Geophysical Research Letters*, **31**, L12S02.
Vidale, J. E., 1986. Complex polarization analysis of particle motion. *Bulletin of the Seismological Society of America*, **76**, 1393–1405.
Waldhauser, F., and Ellsworth, W. L., 2000. A double-difference earthquake location algorithm: method and application to the Northern Hayward Fault, California. *Bulletin of the Seismological Society of America*, **90**, 1353.
Waldhauser, F., and Schaff, D. P., 2008. Large-scale relocation of two decades of Northern California seismicity using cross-correlation and double-difference methods. *Journal of Geophysical Research*, **113**, B08311.
Waldhauser, F., Ellsworth, W. L., and Cole, A., 1999. Slip-parallel seismic lineations on the Hayward Fault and the San Andreas Fault near Parkfield, California. *Geophysical Research Letters*, **70**, 246.
Wang, J., and Teng, T., 1997. Identification and picking of S phase using an artificial neural network. *Bulletin of the Seismological Society of America*, **87**, 1140.
Wilcock, W. S., and Toomey, D. R., 1991. Estimating hypocentral uncertainties for marine microearthquake surveys: A comparison of the generalized inverse and grid search methods. *Marine Geophysical Researches*, **13**, 161.
Wolfe, C. J., 2002. On the mathematics of using difference operators to relocate earthquakes. *Bulletin of the Seismological Society of America*, **92**, 2879.
Zhang, H., Thurber, C., and Rowe, C., 2003. Automatic P-wave arrival detection and picking with multiscale wavelet analysis for single-component recordings. *Bulletin of the Seismological Society of America*, **93**, 1904.
Zhou, H.-W., 1994. Rapid three-dimensional hypocentral determination using a master station method. *Journal of Geophysical Research*, **99**, 15439.

Cross-references

Seismic Monitoring of Nuclear Explosions
Seismic Tomography

EARTHQUAKE, MAGNITUDE

Peter Bormann
Department 2: Physics of the Earth, GFZ German Research Centre for Geosciences, Potsdam, Germany

Synonyms

Richter scale

Definition

The *magnitude* of an earthquake is a number that characterizes the relative size or amount of elastic energy released by such an event (see *Earthquakes, Energy*). It is usually based on measurement of the maximum ground

motion recorded by a seismograph (sometimes for a particular wave type and frequency) and corrected for the decay of amplitudes with epicenter distance and source depth due to geometric spreading and attenuation during wave propagation (see *Propagation of Elastic Waves: Fundamentals*).

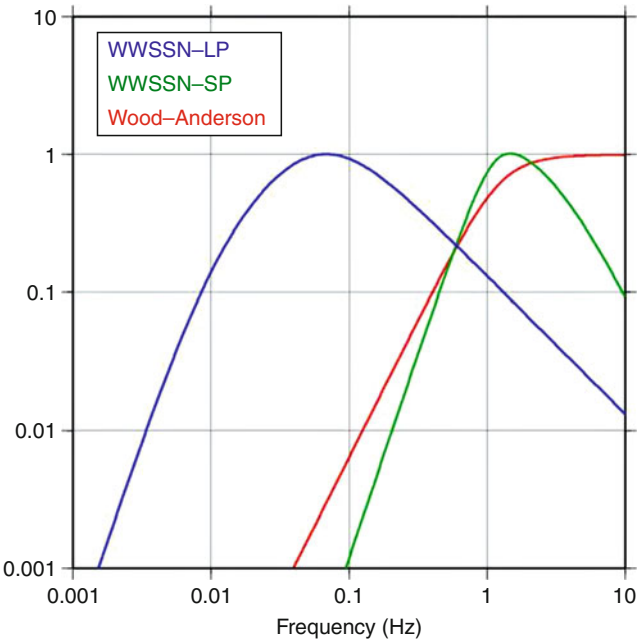
Original definition of earthquake magnitude

The first paper on earthquake magnitude was published by Charles F. Richter (1935), titled *An instrumental earthquake magnitude scale*. Therefore, often reference is made to the size or strength of earthquakes as measured on the *Richter scale*. Yet, this may be incorrect, especially for large earthquakes (magnitude saturation). Richter (1935, p. 1) wrote: "In the course of historical or statistical study of earthquakes in any given region it is frequently desirable to have a scale for rating these shocks in terms of their original energy, independently of the effects which may be produced at any particular point of observation." The term *magnitude* scale was recommended to Richter by H.O. Wood in distinction to the name *intensity* scale (see *Earthquakes, Intensity*), which classifies the severity of earthquakes mainly on the basis of felt shaking or damage manifestations at different localities.

Richter compared records of a standard type of short-period horizontal component Wood-Anderson (WA) torsion seismograph at different distances and azimuths from the earthquake source. All seismographs had an identical frequency response (red curve in Figure 1) and a static magnification of (said to be) 2800. Many stations in Southern California were equipped with WA seismometers. Figure 2 shows a typical record of a near earthquake. Plotting for each event the logarithm of the recorded maximum trace amplitudes A_{max} over the epicentral distance Δ (see *Earthquake, Location Techniques*) Richter derived an average attenuation law for a reference amplitude A_o up to $\Delta = 600$ km. He then corrected amplitudes measured at different stations by the $-\log A_o(\Delta)$ values. Thus measured amplitudes became comparable for a given event and differences of corrected amplitudes for different events indicative for difference in event size. In conclusion Richter (1935, p. 31) states:

> The magnitude of a shock is defined as the logarithm of the calculated maximum trace amplitude, expressed in microns, with which the standard short-period torsion seismometer ($T_o = 0.8$, $V = 2{,}800$, $h = 0.8$) would register that shock at an epicentral distance of 100 km.

Since then several other magnitude scales have been developed. They are based on different types of seismic waves (Bormann et al., 2002a; *Body Waves*; *Surface Waves*), recorded by seismographs with different frequency responses (e.g., those shown in Figure 1) and measured at different periods and source distances. Yet, all had to be mutually scaled so that at certain magnitudes they yield values in agreement with the above original definition.



Earthquake, Magnitude, Figure 1 Normalized amplitude–frequency responses of standard types of seismographs (or of their simulation filters) required for measuring several common types of magnitudes in agreement with recent IASPEI (2005) recommendations: Wood-Anderson seismograph and the short-period (SP) and long-period (LP) seismographs used in the former US World-Wide Standard Seismograph Network (WWSSN) (Seismic Instrumentation). (Reprint from IASPEI, 2005. With copyright granted by IASPEI.)

Local magnitude scales

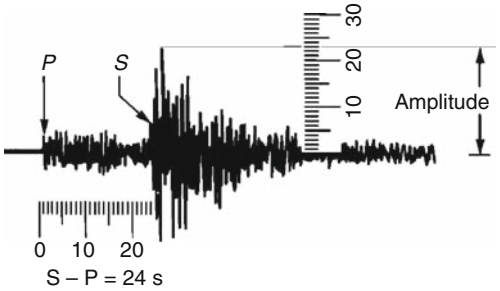
The *Richter scale* is termed M_L and applicable only to earthquakes recorded at local to near regional distances, typically less than 1,000 km distance. Its formula reads:

$$M_L = \log_{10} A_{max} - \log A_o(\Delta). \quad (1)$$

Figure 2 shows how the trace amplitude A_{max} on a WA record is measured and the distance Δ estimated via the arrival-time difference between the P- and S-waves. Figure 3 is a compilation of different $-\log A_o(\Delta)$ calibration curves for different seismotectonic regions of the world. The Richter values for Southern California form the dot-step curve. Tabulated Richter values and calibration formulas for the other curves are given by Bormann (2002).

Figure 3 illustrates that M_L calibration functions may strongly differ from region to region because of differences in crustal structure, thickness, age, heat flow and related wave propagation and attenuation properties. Therefore, in order to assure compatibility of M_L values measured in different regions one has to (a) investigate the regional amplitude- attenuation conditions, (b) derive regional calibration functions, and (c) scale them so as to yield for identical amplitudes measured at 100 km *epicentral* distance the same M_L as with the Richter $-\log A_o(\Delta)$

value at this distance. However, in order to assure a better comparison of earthquakes in regions having very different attenuation of waves within the first 100 km, Hutton and Boore (1987) recommend to define local scales such that M_L = 3 corresponds to 10 mm trace amplitude on a Wood-Anderson instrument at 17 km *hypocentral* distance rather than 1 mm at 100 km *epicentral* distance. Further, empirical measurements proved that the static magnification of the WA seismographs is not 2,800 but around 2,080 ± 60 (Uhrhammer and Collins, 1990).



Earthquake, Magnitude, Figure 2 Record of a short-period Wood-Anderson seismograph of a local earthquake. P marks the onset of the first arriving longitudinal wave, and S that of the shear wave. Note the long tail of scattered coda-waves following S. From the time difference S – P = 24 s follows a distance of about 190 km. The largest amplitude is A_{max} = 23 mm. With –logA_o(190 km) = 3.45 (Richter, 1958) Equation 1 yields M_L = 4.8. (Modified from Figure 1 in Bormann and Saul, 2009a. With copyright granted by Springer.)

Moreover, gain and response of modern seismographs may differ significantly from those of WA seismographs. Therefore, the Working Group on Magnitude Measurements of the International Association of Seismology and Physics of the Earth's Interior (IASPEI, 2005) recommends for crustal earthquakes in regions with attenuative properties similar to those of Southern California the following standard formula:

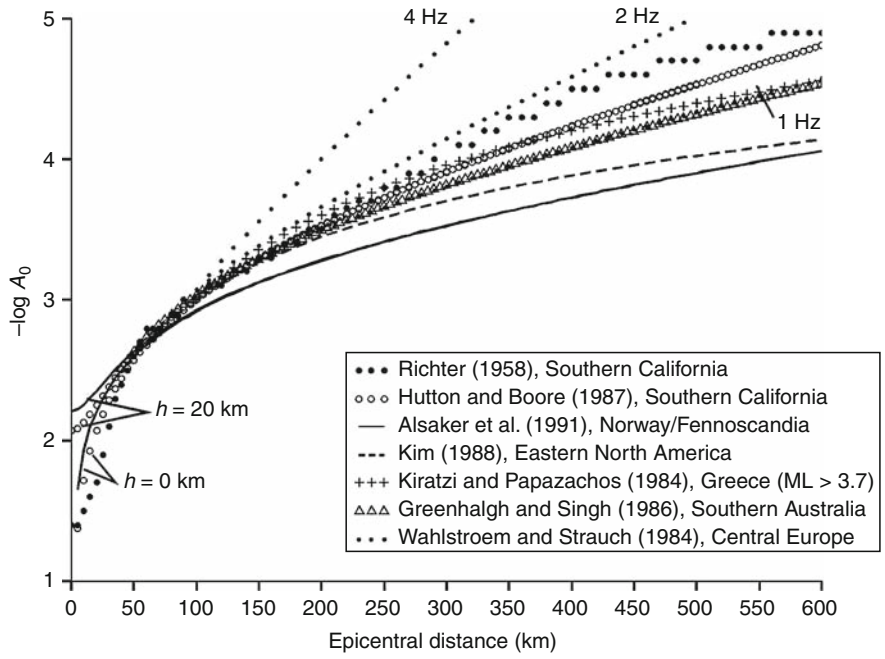
$$M_L = \log_{10} A_{max} + 1.11\log_{10} R + 0.00189 * R - 2.09 \quad (2)$$

where A_{max} = maximum trace amplitude in nm that is measured on the output from a horizontal-component instrument that is filtered so that the response of the seismograph/filter system replicates that of a Wood-Anderson standard seismograph but with a static magnification of 1, and R = hypocentral distance in km, typically less than 1,000 km.

For crustal earthquakes in regions with different attenuative properties and for measuring magnitudes with vertical-component seismographs, the standard equation is of the form:

$$M_L = \log_{10}(A_{max}) + C(R) + D \quad (3)$$

where A_{max} and R are defined as in Equation 2, except that A may be measured from a vertical-component instrument, and where $C(R)$ and D have been calibrated to adjust for the different regional attenuation and for any systematic differences between amplitudes measured on vertical



Earthquake, Magnitude, Figure 3 Calibration functions for M_L determination in different regions. (Reprint of Figure 3.12 in Bormann et al., 2002b. With copyright granted by IASPEI.)

instead of horizontal components. Besides M_L scales that measure A_{max}, typically of relatively weak motion, also M_L^{SM} scales have been developed, which are based on peak ground accelerations (e.g., Lee et al., 1990; Hatzidimitriou et al., 1993; *Earthquakes, Strong-Ground Motion*). Also the duration d of the event record or of its coda length c after the S onset can be used for estimating magnitude from locally recorded data (Bormann et al., 2002b). Further, there exist regression relationships which relate magnitude values to other observed parameter data such as (see *Earthquakes, Intensity*). But any such alternative local scale has to be developed individually for different seismic networks or regions because of regionally variable attenuation, scattering, signal-to-noise and other conditions. Also, for assuring the compatibility of their values these scales should be calibrated with amplitude-based M_L.

Teleseismic magnitude scales

Surface-wave magnitudes

Beno Gutenberg modified the Richter scale for application in the teleseismic range ($\Delta \geq 20°$; $1° = 111.22$ km) by using both body and surface waves. For shallow earthquakes the latter have by far the largest amplitudes in seismic records (Figure 4). The Gutenberg (1945a) formula for surface wave magnitude is

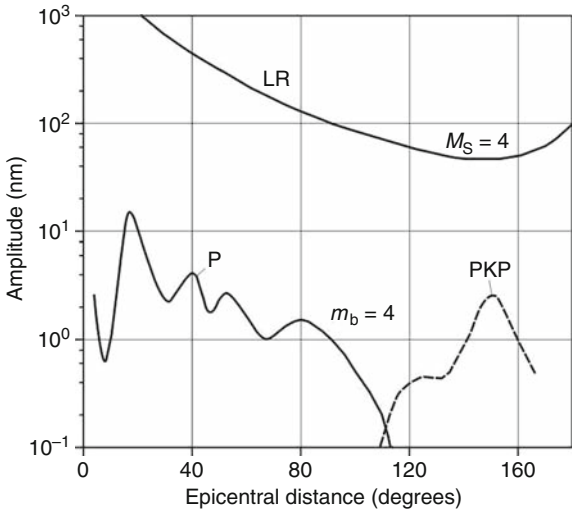
$$M_s = \log A_{Hmax} + 1.656 \log \Delta + 1.818 \quad (4)$$

with A_{Hmax} = vectorially combined maximum horizontal surface-waves displacement amplitude in micrometer measured at periods T around 20 s and the term 1.656 log Δ which compensates for the amplitude decay with distance in the range $20° \leq \Delta < 140°$. Equation 4 is best scaled for surface waves with oceanic travel paths.

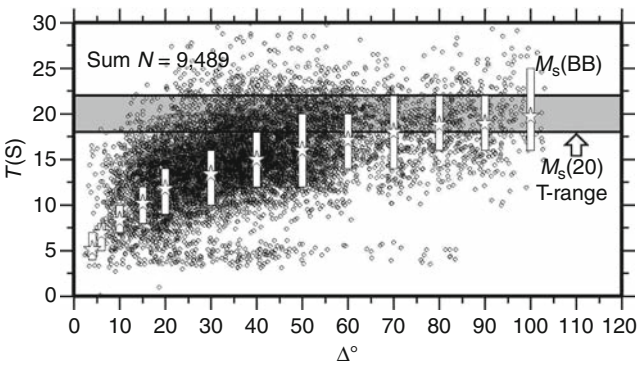
Later, a group of Czech and Russian seismologists investigated the propagation of surface waves with dominatingly continental travel paths, using records of seismographs of type Kirnos that are proportional to ground displacement in a wide range of periods. They found that surface waves may have their largest amplitudes in a much wider range of periods between about 3 and 30 s, as confirmed by Bormann et al. (2009); (see Figure 5). Moreover, estimates of M_s proved to be more stable in a wider range of epicentral distances between 2° and 160° if one measures instead of A_{max} the ratio $(A/T)_{max}$ (Vaněk et al.,1962). This ratio relates to ground velocity $V_{max} = 2\pi(A/T)_{max}$ and thus to the energy contained in the respective wave group. Therefore, these authors proposed a modified M_s formula, which is IASPEI standard for M_s determination since 1967:

$$M_s = \log (A/T)_{max} + 1.66 \log \Delta + 3.3. \quad (5)$$

National and global seismic networks and data centers run or initiated by the USA accepted Equation 5. However, with the introduction of the WWSSN they did not avail of medium- to long-period broadband displacement records that were standard equipment at first-rate stations over most of the Eurasian territory. Also, they continued



Earthquake, Magnitude, Figure 4 Amplitude decay with distance for short-period (1 Hz) P- and PKP-waves as well as long-period (≈20 s) Rayleigh surface waves LR radiated by an earthquake of magnitude 4. (Reprint of Figure 3.13 in Bormann et al., 2002a. With copyright granted by IASPEI.)



Earthquake, Magnitude, Figure 5 Variability with distance Δ of periods associated with $(A/T)_{max}$ in vertical component surface waves records. The gray shaded stripe marks the period range between 18 and 22 s, the light-gray columns mark the observed range of average periods as published by Vaněk et al. (1962), and the stars the average periods for different increments of the data cloud. (Reprint of Figure 8 by Bormann et al. 2009, p. 1874. With copyright granted by the Seismological Society of America.)

to measure A_{max} only at periods around 20 s rather than true $(A/T)_{max}$ in the whole surface-wave train. But then the use of Equation 5 results in significantly distance-dependent M_s estimates (see discussion in Bormann et al., 2009). Proposed new calibration functions for 20 s surface-waves, e.g., by von Seggern (1977), Herak and Herak (1993) and Rezapour and Pearce (1998) have not yet been accepted as global standards. In contrast, Kirnos-based M_s measurements in accordance with formula 5 have ever since been carried out in the former

Soviet Union and its allied countries, in the now Commonwealth of Independent States (CIS), and China.

Thus, two traditions of M_s measurements have been practiced for half a century, producing a host of valuable but not fully compatible data. Yet, the generic nomenclature M_s does not allow to discriminate between these two kinds of data. Moreover, M_s is measured by different agencies still in sometimes different ranges of source-depth, on different record components of different types of seismographs and sometimes even using different calibration functions, e.g., the International Data Centre (IDC) of the Comprehensive Test-Ban Treaty Organization (see *Seismology, Monitoring of CTBT*). This may result in procedure-dependent errors that have to be reduced by standardization of both nomenclature and procedures. This task is eased by the fact that modern digital broadband records allow to accurately synthesize records of classical seismographs, on which the original definition of magnitude scales often rests. Moreover, modern broadband records are mostly proportional to velocity and allow to measure $(A/T)_{max}$ directly and unambiguously on unfiltered traces. Therefore, IASPEI (2005) recommends for shallow earthquakes with source depth $h < 60$ km in future two surface-wave magnitude standards, namely, (a) a purely teleseismic one in the range $20° \leq \Delta \leq 160°$, termed $M_s(20)$, or when written in the International Seismological Format (ISF):

$$M_s\_20 = \log_{10}(A/T) + 1.66 \log_{10}\Delta + 0.3 \quad (6)$$

where A = vertical-component ground displacement in nanometer measured from the maximum surface-wave trace-amplitude having a period between 18 and 22 s on a waveform that has been filtered so as to replicate the record of the WWSSN-LP seismograph (Figure 1), and (b) another broadband (BB) surface-wave magnitude that is measured in a wider distance range of $2° \leq \Delta \leq 160°$, termed $M_s(BB)$, or in the ISF format:

$$Ms\_BB = \log_{10}(A/T)_{max} + 1.66 \log_{10}\Delta + 0.3 \quad (7)$$

where $(A/T)_{max} = (V_{max}/2\pi)$, with V_{max} = ground velocity in nm/s associated with the maximum trace-amplitude in the surface-wave train recorded on a vertical-component seismogram that is proportional to velocity, and where the period T is between 3 s $< T <$ 60 s. $M_s(BB)$ generally agrees well with $M_s(20)$, however, for weaker regional and local earthquakes it may yield up to about 0.5 magnitude units (m.u.) larger values than $M_s(20)$ (Bormann et al., 2009). Besides these two standard M_s there exist other surface-wave magnitude scales. E.g., Bonner et al. (2006) proposed a largely theory-based $M_s(V_{max})$ that requires multi-bandpass filtering and frequency-dependent attenuation corrections. Another scale, termed *Mm* (Okal and Talandier, 1989), uses very long-period surface-wave amplitudes with periods between 60 and 400 s, which penetrate into the earth mantle. *Mm* does not saturate (magnitude saturation) and thus is well suited for tsunami warning (e.g., Weinstein and Okal, 2005).

Body-wave magnitudes

Gutenberg (1945b, c) teleseismic magnitude scales for body-waves are applicable also to deep earthquakes down to source depths $h \approx 700$ km. Diagrams and tables with calibration values $Q(\Delta, h)$ for P, PP, and S waves (see *Seismogram Interpretation*), covering the distance range between about 5° and 170°, were published by Gutenberg and Richter (1956) (see also Bormann, 2002). Figure 6 shows the Q-values for vertical component P waves. It is the only classical body-wave scale which is still widely used. The general body-wave magnitude formula, applicable in the range $20° \leq \Delta \leq 100°$ reads, when A is measured in micrometer:

$$m_B = \log_{10}(A/T)_{max} + Q(\Delta, h). \quad (8)$$

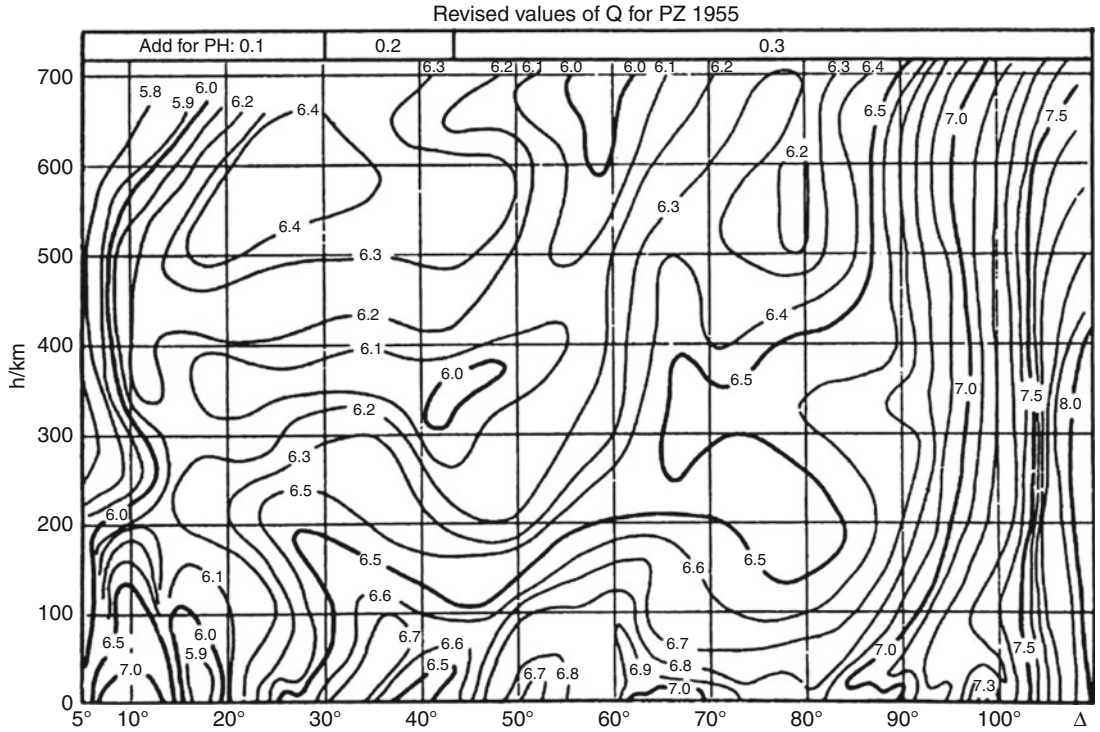
Gutenberg derived the Q-functions by analyzing mostly medium- to long-period body-wave amplitudes with periods 2 s $< T <$ 30 s (Abe, 1981). Therefore he did not account for frequency-dependent attenuation. Yet, the latter becomes significant for P-waves with $T <$ 4 s (Di Giacomo et al., 2008). Further, Gutenberg measured $(A/T)_{max}$ within the whole P-wave train and not within a fixed short time window after the P-wave onset. This was not taken into account when introducing with the WWSSN in the 1960s a short-period P-wave magnitude m_b. It measured A_{max} on narrow-band WWSSN-SP records (Figure 1) at periods < 3 s (mostly around 1 s) at fixed time-windows. The main advantage was the improved signal-to-noise ratio. It allowed teleseismic magnitudes to be determined down to $m_b \approx 4$. m_b became standard in all western countries, which discontinued to measure m_B. Yet, on average for values above 5.5, m_b tends to be smaller than other kinds of magnitudes (see Figure 7). Reasons are discussed in the section on *Magnitude saturation.*

Measurement procedures for m_b have been rather inconsistent, using different frequency responses, measuring the amplitudes within different time windows, applying different calibration functions (e.g., the IDC). The detrimental effect of such inconsistencies became strikingly evident after the great M_w9.3 Sumatra earthquake of 26 December 2004. The IDC reported $m_b = 5.7$, CENC 6.3, NEIC 7.0 and the German Regional Seismic Network (GRSN) 7.5. The IDC and CENC measured A_{max} within 5 s, the NEIC within 60 s, and the GRSN some 90 s after the first P onset. Such differences for the same event on a logarithmic scale are not tolerable and necessitate standardization of procedures. Broadband $m_B = 8.3$ was still larger than the largest m_b value.

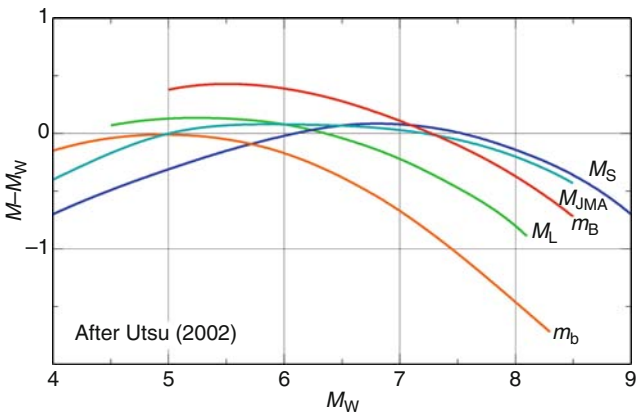
Based on such experience IASPEI (2005) proposes two complementary teleseismic body-wave magnitude standards:

$$m_b = \log_{10}(A/T) + Q(\Delta, h) - 3.0 \quad (9)$$

where: A = P-wave ground amplitude in nm, calculated from the maximum trace-amplitude in the entire P-phase train (time spanned by P, pP, sP, possibly PcP and



Earthquake, Magnitude, Figure 6 Calibration values $Q(\Delta, h)$ for vertical (Z) component P-wave amplitudes as a function of epicentral distance Δ in degree and source depth h in kilometer. They are used for the calculation of body-wave magnitudes m_B and m_b. (Figure redrawn from Gutenberg and Richter, 1956. Reprint of Figure 1a in Bormann, 2002. With copyright granted by IASPEI.)



Earthquake, Magnitude, Figure 7 Average relationships of different common types of magnitude scales with the moment magnitude M_w (non-saturating magnitude scales). M_{JMA} is determined by the Japanese Meteorological Agency. Modified from Figure 1 of Utsu (2002). Reproduction of Figure 4 in Bormann and Saul (2009a). With copyright granted by Springer.

their codas, and ending preferably before PP; *Seismogram Interpretation*); period $T < 3$ s, Δ in the range $20° \leq \Delta \leq 100°$ and h between 0 and 700 km. T and A are measured on the output of a vertical-component instrument that is filtered so that the frequency response of the seismograph/filter system replicates that of a WWSSN-SP seismograph (Figure 1). This new definition of m_b yields for great earthquakes significantly larger values than common procedures. None the less, short-period m_b should – at least for $m_b > 5.5$ – be complemented by broadband $m_B(BB)$, which reads in ISF format:

$$m_B\text{-}BB = \log_{10}(A/T)_{max} + Q(\Delta, h) - 3.0, \qquad (10)$$

where: $(A/T)_{max} = (V_{max}/2\pi)$, with V_{max} = ground velocity in nm/s associated with the maximum trace-amplitude in the entire P-phase train (see m_b) as recorded on a vertical-component seismogram that is proportional to velocity at least in the period-range 0.2 s $< T <$ 30 s; $Q(\Delta, h)$, Δ and h ranges are as for m_b.

Besides these two body-wave standards there exist several other broadband P-wave magnitudes. Most of them aim at fast estimates of the moment magnitude M_w (non-saturating magnitude scales).

Relationships between magnitude scales and released seismic energy E_s

There is no "true" single magnitude parameter which can sufficiently reliable and complete characterize both size and shaking strength of earthquakes and thus related *Seismic hazard*. Depending on the measured kinematic parameters and their frequency range the magnitudes

relate to different geometric, kinematic, or dynamic aspects of the (see *Earthquake Rupture: Inverse Problem*). Accordingly, the scales discussed so far do not scale 1:1, agree in some value ranges but differ in others. Although Figure 7 hints to sometimes nonlinear relationships, the most important ones are reasonably linear within the range of data scatter, e.g., that of Gutenberg and Richter (1956)

$$m_B = 0.63M_s + 2.5. \quad (11)$$

Equation 11 is well reproduced by the new standards $m_B(BB)$ and $M_s(BB)$ (Bormann et al., 2009) but differs strongly from the $m_b - M_s(20)$ relationship (Gordon, 1971):

$$m_b = 0.47M_s(20) + 2.79. \quad (12)$$

Despite the shortcomings of narrow-band m_b and $M_s(20)$ they are particularly useful for discriminating earthquakes and Nuclear Explosions.

Gutenberg and Richter (1956) published also a semi-empirical relationship between m_B and seismic energy E_s:

$$\log E_s = 2.4m_B - 1.2 (\text{with } E_s \text{ in units of Joule}). \quad (13)$$

When inserting Equation 11 into Equation 13 one gets

$$\log E_s = 1.5M_s + 4.8. \quad (14)$$

M_s in all these formulas means $M_s(20)$. But, 20 s displacement amplitudes are not very representative for the radiated seismic energy, especially not for smaller earthquakes. Another relationship has been derived by Kanamori et al. (1993) between M_L and $\log E_s$ for Southern California, valid in the range $1.5 < M_L < 6.0$:

$$\log E_s = 1.96M_L + 2.05. \quad (15)$$

But any such relationship allows only rough estimates of E_s within about an order of magnitude. This uncertainty can nowadays be reduced, down to a factor of about 3–5, by integrating squared velocity broadband records with good signal-to-noise ratio in the teleseismic range and thus measuring directly (see *Earthquakes, Energy*).

Magnitude "saturation"

Magnitude saturation is a widely but not uniquely used term in seismological literature. Originally, one assumed that classical magnitudes do not increase further beyond some value of a reference magnitude of earthquake size such as M_w (non-saturating magnitude scales) (e.g., Geller, 1976; Kanamori, 1983). However, such a complete saturation is not supported by Figures 7 and 8 and recently published data (e.g., Bormann et al., 2009). We therefore may speak of magnitude "saturation," or more correct underestimation of earthquake size, also if one magnitude Mx increases slowly with respect to increases in M_w (see Figure 7). Such saturation may occur when the earthquake rupture duration is much longer than the periods at which A_{max} or $(A/T)_{max}$ are measured.

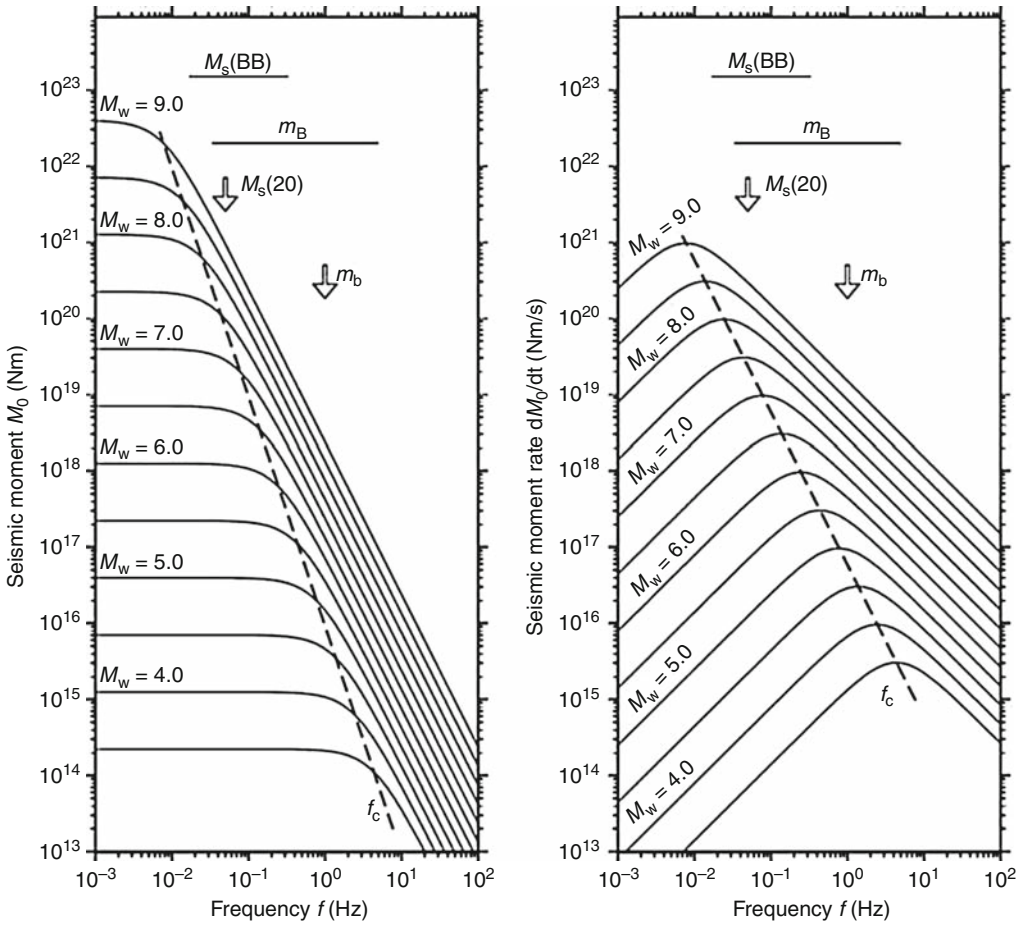
This is also illustrated by Figure 8, calculated for a common ω^{-2} source model (Aki, 1967), assuming an average rupture velocity, stress drop in the source volume and ratio E_S/M_0. For details see Bormann et al. (2009). The spectra have been plotted together with the periods, respectively period ranges, at which m_b, $m_B(BB)$, $M_s(20)$ and $M_s(BB)$ are measured. The left panel is scaled to $M_0 = \mu\ D\ F$ (where F = rupture area, D = average displacement over the fault and μ = average rigidity of the rock material in the source volume) and the right panel to the time derivative of M_0 (moment rate), which is proportional to far-field ground motion velocity. M_0 can only be estimated correctly when displacement amplitudes are measured at frequencies less than the corner frequency f_c (left panel) and seismic energy, which is proportional to the squared ground motion velocity, only when velocity spectral amplitudes are measured sufficiently wide around f_c. Accordingly, m_b measured around 1 Hz necessarily begins (on average) to underestimate both M_0 and E_s for $M_w > 5$ to 5.5, in contrast to the broadband magnitudes, which are more closely related to M_0 and E_s up to M_w 7.5–8. The spectral reasons for underestimating earthquake size or energy release may be aggravated by too short a measurement time window, which was rather common for m_b. Figure 9 illustrates both the spectral and time-window component of magnitude saturation.

Non-saturating magnitude scales

Kanamori (1977) , Purcaru and Berckhemer (1978) and Hanks and Kanamori (1979) proposed a non-saturating magnitude scale for very large earthquakes, based on M_0 data. Kanamori termed it M_w. For converting M_0 into M_w Kanamori (1977) replaced in Equation 14 M_s by M_w and assumed an average ratio $E_s/M_0 = 5 \times 10^{-5} = \text{constant}$. This led to a relationship, which reads in the IASPEI (2005) recommended standard form with M_0 in units of Nm:

$$M_w = (2/3)(\log_{10} M_0 - 9.1). \quad (16)$$

M_0 calculations are nowadays routine at major seismological centers, either by measuring the displacement amplitude u_0 on the long-period asymptote (plateau) of the wave spectrum, which is proportional to M_0 after correcting for the source radiation and wave propagation effects (as done in Figure 8), or by fitting long-period synthetic seismograms for different source models best to long-period filtered real seismic records, using wavelengths that are larger than the rupture length (Dziewonski et al., 1981). However, if this condition is not fulfilled, as sometimes in routine computations, then also M_w may be underestimated, e.g., for the great 2004 Sumatra earthquake values between 8.2 and 9.3 were reported. The last value resulted from using either extremely long-period free oscillation data (Stein and Okal, 2005) or from multiple CMT source analysis (Tsai et al., 2005). For all major earthquakes with $M_w > 5.5$ the Global Centroid Moment Tensor (GCMT) project publishes moment tensor and

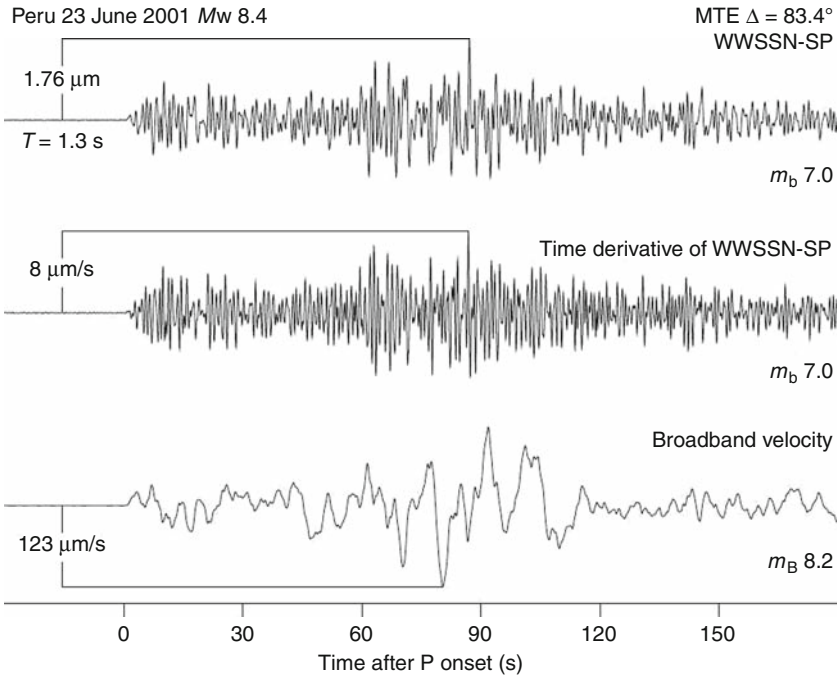


Earthquake, Magnitude, Figure 8 "Source spectra" of ground displacement amplitudes *A* for an "average" seismic shear source as a function of frequency *f*, scaled to M_0 and M_w. *Right*: The same as left, but for ground motion velocity amplitudes *V*, scaled to seismic moment rate and M_w. The maximum of seismic energy $Es \sim V^2$ is radiated around f_c. The open arrows point to the center frequencies on the abscissa at which m_b and $M_s(20)$ are measured. The horizontal bars mark the frequency range within which V_{max} of P- and Rayleigh-waves for *mB*(BB) and M_s(BB) should be measured. (Courtesy of Di Giacomo. Reproduced from Figure 1 of Bormann et al., 2009, p. 1870. With copyright granted by the Seismological Society of America.)

M_w solutions usually within a few hours after the event (http://www.globalcmt.org/CMTsearch.html). Regional centers, such as the Swiss Earthquake Sercice, may provide – with larger time delay – such data down to $M_w \approx 3.0$ when dense quality networks close to the source are available (Braunmiller et al., 2002).

Besides calculation of M_w via direct measurements of M_0 there exist several methods of fast M_w proxy estimates, e.g., M_{wp} in the context of tsunami early warning (Tsuboi et al., 1999). M_{wp} scales nearly 1:1 with broadband m_B (Bormann and Saul, 2009a) but also tends to underestimate very great and (see *Slow Earthquake*). More advanced fast and non-saturating versions of broadband P-wave magnitudes are M_{wpd} by Lomax and Michelini (2009), the (unspecified) *M* by Hara (2007), and m_{Bc}, a modified "cumulative" m_B procedure for great multiple-rupture earthquakes (Bormann and Saul, 2009b). All these procedures account additionally for rupture duration, estimated in real time. Another very perspective promising new method for rapid non-saturating moment tensor and M_w estimates uses very long-period W-phase observations between the P-wave and S-wave arrivals (Kanamori and Rivera, 2008).

However, both M_0 and M_w are only long-period or quasi-static measures of earthquake size and thus of the tectonic earthquake effect. More relevant for assessing seismic hazard in terms of the earthquake potential for causing shaking damage are estimates of the amount of released seismic energy E_s and of the corner frequency f_c. These two parameters largely control the relative amount of radiated high-frequency velocity amplitudes. This necessitates, complementary M_w, also an energy-scaled magnitude M_e, which better relates to the kinematics and dynamics of the earthquake rupture, such as rupture velocity and stress drop. Velocity-broadband records allow nowadays to estimate (see *Earthquakes, Energy*) directly by



Earthquake, Magnitude, Figure 9 Plot of vertical-component P-wave records of an M_w8.4 Peru earthquake at station MTE at $\Delta = 83.4°$. It illustrates the difference between broadband *mB* and short-period m_b for large earthquakes. *Top trace*: m_b based on measuring A_{max} and the related period T from the simulated WWSSN_SP record. *Second trace*: Time derivative of the top trace for making $V_{max} = 8$ µm/s for m_b comparable with $V_{max} = 123$ µm/s on the *bottom trace*. The latter yields $m_B = 8.2$, in contrast to IASPEI standard $m_b = 7.0$, or 6.4, when A_{max} is measured within the first 5 s. (Copy of Figure 1 in Bormann and Saul, 2008, p. 699. With copyright granted by the Seismological Society of America.)

integrating squared velocity amplitudes over a wide range of periods around the peak of the velocity spectrum (see Figure 8) and correcting them for wave propagation effects. Comparing so determined E_s values for hundreds of shallow earthquakes with their NEIC M_s magnitudes in the range $5.5 < M_s < 8.5$ Choy and Boatwright (1995) found

$$log\, E_s = 1.5M_s + 4.4. \quad (17)$$

Equation 17 differs from Equation 14 that was used for deriving the M_w relationship 16 only in the constant. Replacing in Equation 17 M_s by M_e and resolving it for M_e yields

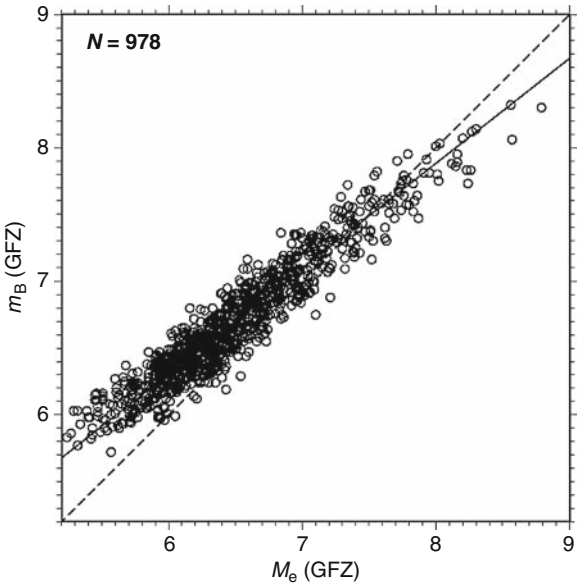
$$M_e = (2/3)(\log E_s - 4.4). \quad (18)$$

Di Giacomo et al. (2010b) showed that their $M_e(GFZ)$ and the related M_w(GCMT) agree on average in the range $5.5 < M_w < 9$ within 0.1 m.u., with a standard deviation SD = 0.25. The respective deviations between M_w(GCMT) and $M_e(NEIC)$ are larger. The reason is that $M_e(NEIC)$ is calculated from E_s values to which theoretically derived source mechanism corrections have been applied. In contrast, $M_e(GFZ)$ follows the tradition of classical magnitudes with no such corrections made to measured A_{max} or V_{max}. The pros and cons of such corrections and of their size under real Earth conditions of wave radiation and propagation are still a matter of debate and require further research.

M_e and M_w may differ for the same earthquakes up to about 1 m.u. This is mainly due to the fact that earthquake stress drop and the related ratio E_s/M_0 are not constant but may vary more than two to three orders. Examples of striking differences in M_e and M_w have been published, e.g., by Choy and Kirby (2004) and by Di Giacomo et al. (2010a). Suitable seismotectonic setting provided, $M_e << M_w$ may be indicative for high tsunamigenic potential, and vice versa, $M_e >> M_w$ for high shaking damage potential (see also Figure 11 and related discussion).

Relationship between M_w, M_e, and classical magnitudes

Scordilis (2006) published global regression relationships, based on many thousands of NEIC and ISC (International Seismological Centre) m_b and M_s values, with M_w values of the US Geological Survey and of Harward. First regression relationships between IASPEI standard magnitudes with non-saturating M_w, M_e, and m_{Bc} have been derived by Bormann and Saul (2008, 2009a, b) and by Bormann et al. (2009). From the classical magnitudes, M_s relates for strong earthquakes best with M_w, tends to be smaller, however, for $M_w < 6.2$–6.5 and > 7.8–8.5 (see Figure 7; Scordilis, 2006; Ekström and Dziewonski, 1988). Yet,



Earthquake, Magnitude, Figure 10 Relationship between automatically determined values of m_B and M_e according to procedures developed at the GFZ German Research Centre for Geosciences by Bormann and Saul (2008) and Di Giacomo et al. (2010). The *broken line* is the 1:1 and the solid line the orthogonal regression relation $m_B = 0.79M_e + 1.58$. (Courtesy of D. Di Giacomo and J. Saul.)

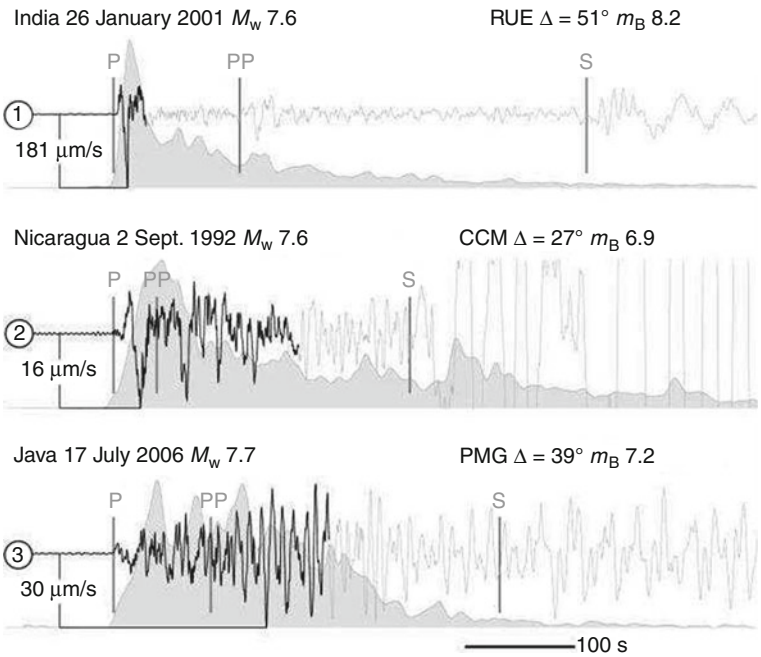
$M_s(BB)$ reduces to half the systematic underestimation of M_w by $M_s(20)$ towards smaller magnitudes (Bormann et al., 2009). The $M_s(20)$ bias had been predicted already theoretically by Kanamori and Anderson (1975).

The first magnitude-energy Equation 13 had been defined with m_B. Automatically determined $m_B(GFZ)$, which agrees with IASPEI standard $m_B(BB)$, correlates very well with $M_e(GFZ)$ (Figure 10), allowing estimates of M_e via simple m_B measurements with a SD = ±0.18 m.u. Both magnitudes are based on ground motion velocity and therefore reflect – when compared with M_w – in a similar way differences in source properties.

Figure 11 presents three records of equal M_w (within 0.1 m.u.), however very different rupture durations and m_B values. The Indian intra-plate Budj earthquake had the shortest rupture duration of about 30 s, the most impulsive waveform and largest m_B = 8.2, 0.6 m.u. larger than M_w. This event caused tremendous shaking damage. In contrast, the two very slow Nicaragua and Java earthquakes with rupture durations around 200 s were (see *Tsunami*) earthquakes. Their m_B values were 0.5 and 0.7 m.u. smaller than M_w. Both earthquakes were not or only weakly felt at the relatively nearby coasts.

Summary and conclusions

The aim and historical development of magnitude scales have been sketched, systematic inconsistencies between



Earthquake, Magnitude, Figure 11 Velocity broadband records of three earthquakes with similar M_w but exceptionally different rupture durations, m_B values and earthquake effects. The rupture duration D is estimated from the envelope of high-frequency P-wave radiation (*gray shaded area*). D corresponds approximately to the time window from the P-wave onset until the envelop amplitude drops to 40% of its maximum value (*black record trace*). (Copy of Figure 2 in Bormann and Saul, 2008, p. 700. With copyright granted by the Seismological Society of America.)

current measurement procedures and their sometimes detrimental effects been outlined, the need for international standardization highlighted, and the recently proposed IASPEI standard procedures introduced. Reduction of magnitude errors due to incompatible measurement practices is a precondition for improving the usefulness of magnitude data for a more realistic quick assessment of differences in source physics and associated actual earthquake risk. No single magnitude value can describe the peculiarity of individual earthquakes both in terms of their tectonic size and shaking strength sufficiently reliable. A multi-magnitude documentation in earthquake catalogs and a deeper understanding of the physics behind different magnitude data is indispensable also for upgrading procedures aimed at improved medium- to long-term seismic hazards assessment.

Bibliography

Abe, K., 1981. Magnitudes of large shallow earthquakes from 1904 to 1980. *Physics of the Earth and Planetary Interiors*, **27**, 72–92.

Aki, K., 1967. Scaling law of seismic spectrum. *Journal of Geophysical Research*, **72**, 1217–1231.

Bonner, J. L., Russell, D. R., Harkrider, D. G., Reiter, D. T., and Herrmann, R. B., 2006. Development of a time-domain, variable-period surface-wave magnitude measurement procedure for application at regional and teleseismic distances, part II: Application and M_s-m_b performance. *Bulletin of the Seismological Scoiety of America*, **96**, 678–696.

Bormann, P., 2002. Magnitude calibration functions and complementary data. In Bormann, P. (ed.), *IASPEI New Manual of Seismological Observatory Practice*. Potsdam: GeoForschungs Zentrum Potsdam, Vol. 2, DS3.1, pp. 1–7.

Bormann, P., and Saul, J., 2008. The new IASPEI standard broadband magnitude m_B. *Seismological Research Letters.*, **79**, 698–705.

Bormann, P., and Saul, J., 2009a. Earthquake magnitude. In Meyers, A. (ed.), *Encyclopedia of Complexity and Systems Science*. New York: Springer, Vol. 3, pp. 2473–2496.

Bormann, P., and Saul, J., 2009b. A fast, non-saturating magnitude estimator for great earthquakes. *Seismological Research Letters*, **80**, 808–816.

Bormann, P., Engdahl, E. R., and Kind, R., 2002a. Chapter 2: Seismic wave propagation and Earth models. In Bormann, P. (ed.), *IASPEI New Manual of Seismological Observatory Practice*. Potsdam: GeoForschungs Zentrum Potsdam, Vol. 1, Chapter 2, pp. 1–70.

Bormann, P., Baumbach, M., Bock, G., Grosser, H., Choy, G. L., and Boatwright, J., 2002b. Seismic sources and source parameters. In Bormann, P. (ed.), *IASPEI New Manual of Seismological Observatory Practice*. Potsdam: GeoForschungs Zentrum Potsdam, Vol. 1, Chapter 3, pp. 1–94.

Bormann, P., Liu, R., Xu, Z., Ren, K., Zhang, L., and Wendt, S., 2009. First application of the new IASPEI teleseismic magnitude standards to data of the China National Seismographic Network. *Bulletin of the Seismological Society of America*, **99**, 1868–1891.

Braunmiller, J., Kradolfer, U., Baer, M., and Giardini, D., 2002. Regional moment tensor determinations in the European-Mediterranian area – initial results. *Tectonophysics*, **356**, 5–22.

Choy, G. L., and Boatwright, J., 1995. Global patterns of radiated seismic energy and apparent stress. *Journal of Geophysical Research*, **100**, 18205–18228.

Choy, G. L., and Kirby, S., 2004. Apparent stress, fault maturity and seismic hazard for normal-fault earthquakes at subduction zones. *Geophysical Journal International*, **159**, 991–1012.

Di Giacomo, D., Grosser, H., Parolai, S., Bormann, P., and Wang, R., 2008. Rapid determination of Me for strong to great shallow earthquakes. *Geophysical Research Letters*, **35**, L10308, doi:10.1029/2008GL033505.

Di Giacomo, D., Parolai, S., Bormann, P., Grosser, H., Saul, J., Wang, R., and Zschau, J., 2010a. Suitability of rapid energy magnitude determinations for rapid response purposes. *Geophysical Journal International*, **180**, 361–374.

Di Giacomo, D., Parolai, S., Bormann, P., Grosser, H., Saul, J., Wang, R., and Zschau, J., 2010b. Erratum. *Geophysical Journal International*, **181**, 1725–1726.

Dziewonski, A. M., Chou, T. A., and Woodhouse, J. H., 1981. Determination of earthquake source parameters from waveform data for studies of global and regional seismicity. *Journal of Geophysical Research*, **86**, 2825–2852.

Ekström, G., and Dziewonski, A. M., 1988. Evidence of bias in estimations of earthquake size. *Nature*, **332**, 319–323.

Geller, R. J., 1976. Scaling relations for earthquake source parameters and magnitudes. *Bulletin of the Seismological Society of America*, **66**, 1501–1523.

Gordon, D. W., 1971. Surface-wave versus body-wave magnitude. *Earthquake Notes*, **42**, 20–28.

Gutenberg, B., 1945a. Amplitude of surface waves and magnitude of shallow earthquakes. *Bulletin of the Seismological Society of America*, **35**, 3–12.

Gutenberg, B., 1945b. Amplitudes of P, PP, and S and magnitude of shallow earthquakes. *Bulletin of the Seismological Society of America*, **35**, 57–69.

Gutenberg, B., 1945c. Magnitude determination of deep-focus earthquakes. *Bulletin of the Seismological Society of America*, **35**, 117–130.

Gutenberg, B., and Richter, C. F., 1956. Magnitude and energy of earthquakes. *Annali di Geofisica*, **9**, 1–15.

Hatzidimitriou, P., Papazachos, C., Kiratzi, A., and Theodulidis, N., 1993. Estimation of attenuation structure and local earthquake magnitude based on acceleration records in Greece. *Tectonophysics*, **217**, 243–253.

Hanks, C., and Kanamori, H., 1979. A moment magnitude scale. *Journal of Geophysical Research*, **84**, 2348–2350.

Hara, T., 2007. Measurement of duration of high-frequency energy radiation and its application to determination of magnitudes of large shallow earthquakes. *Earth, Planets, and Space*, **59**, 227–231.

Herak, M., and Herak, D., 1993. Distance dependence of M_S and calibrating function for 20 second Rayleigh waves. *Bulletin of the Seismological Society of America*, **83**, 1881–1892.

Hutton, L. K., and Boore, D. M., 1987. The M_L scale in Southern California. *Bulletin of the Seismological Society of America*, **77**, 2074–2094.

IASPEI, 2005. Summary of Magnitude Working Group recommendations on standard procedures for determining earthquake magnitudes from digital data (available online at http://www.iaspei.org/commissions/CSOI.html).

Kanamori, H., 1977. The energy release in great earthquakes. *Journal of Geophysical Research*, **82**, 2981–2987.

Kanamori, H., 1983. Magnitude scale and quantification of earthquakes. *Tectonophysics*, **93**, 185–199.

Kanamori, H., and Anderson, D. L., 1975. Theoretical basis of some empirical relations in seismology. *Bulletin of the Seismological Society of America*, **65**, 1073–1095.

Kanamori, H., and Rivera, L., 2008. Inversion of *W* phase: speeding up seismic tsunami warning. *Geophysical Journal International*, **175**, 222–238.

Kanamori, H., Mori, J., Hauksson, E., Heaton, Th. H., Hutton, L. K., and Jones, L. M., 1993. Determination of earthquake energy release and M_L using TERRASCOPE. *Bulletin of the Seismological Society of America*, **83**, 330–346.

Lee, V., Trifunac, M., Herak, M., Živčić, M., and Herak, D., 1990. M_L^{SM} computed from strong motion accelerograms

recorded in Yugoslavia. *Earthquake Engineering and structural dynamics*, **19**, 1167–1179.
Lomax, A., and Michelini, A., 2009. M_{wpd}: A duration-amplitude procedure for rapid determination of earthquake magnitude and tsunamigenic potential from *P* waveforms. *Geophysical Journal International*, **176**, 200–214.
Okal, E. A., and Talandier, J., 1989. M_m: A variable-period mantle magnitude. *Journal of Geophysical Research*, **94**, 4169–4193.
Purcaru, G., and Berckhemer, H., 1978. A magnitude scale for very large earthquakes. *Tectonophysics*, **49**, 189–198.
Rezapour, M., and Pearce, R. G., 1998. Bias in surface-wave magnitude M_S due to inadequate distance corrections. *Bulletin of the Seismological Society of America*, **88**, 43–61.
Richter, C. F., 1935. An instrumental earthquake magnitude scale. *Bulletin of the Seismological Society of America*, **25**, 1–32.
Richter, C., 1958. *Elementary Seismology.* San Francisco/London: W.H. Freeman.
Scordilis, E. M., 2006. Empirical global relations converting Ms and mb to mement magnitude. *Journal of Seismology*, **10**, 225–236.
Stein, S., and Okal, E. A., 2005. Speed and size of the Sumatra earthquake. *Nature*, **434**, 581–582, doi:10.1038/434581a.
Tsai, V. C., Nettles, M., Ekström, G., and Dziewonski, A., 2005. Multiple CMT source analysis of the 2004 Sumatra earthquake. *Geophysical Research Letters*, **32**, L17304, doi:10.1029/2005GL023813.
Tsuboi, S., Whitmore, P. H., and Sokolowski, T. J., 1999. Application of M_{WP} to deep and teleseismic earthquakes. *Bulletin of the Seismological Society of America*, **89**, 1345–1351.
Uhrhammer, R. A., and Collins, E. R., 1990. Synthesis of Wood-Anderson seismograms from broadband digital records. *Bulletin of the Seismological Society of America*, **80**, 702–716.
Utsu, T., 2002. Relationships between magnitude scales. In Lee, A., Kanamori, H., Jennings, P. C., and Kisslinger, C. (eds.), *International Handbook of Earthquake and Engineering Seismology.* Amsterdam: Academic, Part A, pp. 733–746.
Vaněk, J., Zátopek, A., Kárnik, V., Kondorskaya, N. V., Riznichenko, Y. V., Savarenski, E. F., Solov´ev, S. L., and Shebalin, N. V., 1962. Standardization of magnitude scales. *Izvestiya Akademii Nauk SSSR, Seriya Geofisičeskaya*, **2**, 153–157 (with English translation).
von Seggern, D., 1977. Amplitude-distance relation for 20-second Rayleigh waves. *Bulletin of the Seismological Society of America*, **67**, 405–411.
Weinstein, S. A., and Okal, E. A., 2005. The mantle wave magnitude Mm and the slowness parameter Θ: Five years of real-time use in the context of tsunami warning. *Bulletin of the Seismological Society of America*, **95**, 779–799.

Cross-references

Body Waves
Earthquake Rupture: Inverse Problem
Earthquake, Focal Mechanism
Earthquake, Location Techniques
Earthquakes, Energy
Earthquakes, Intensity
Earthquakes, Strong-Ground Motion
Great Earthquakes
Propagation of Elastic Waves: Fundamentals
Seismic Hazard
Seismic Instrumentation
Seismic Monitoring of Nuclear Explosions
Seismogram Interpretation
Slow Earthquake
Surface Waves
Tsunami

EARTHQUAKES AND CRUSTAL DEFORMATION

Robert McCaffrey
Department of Geology, Portland State University, Portland, OR, USA

Definition

The term earthquake is commonly used only to describe sudden slip on a fault within the Earth that produces seismic waves in the frequency bands that we can either feel or observe with seismometers. However, other sources that produce deformation of the Earth's surface do not generate seismic waves but are instead observable with sensitive geodetic instruments that measure the crustal deformation directly. Because of the great advances over the past few decades in observing and understanding them, in this chapter we broaden the definition of earthquake to include these "quiet" sources of Earth deformation. Other chapters deal with some of these sources in more detail.

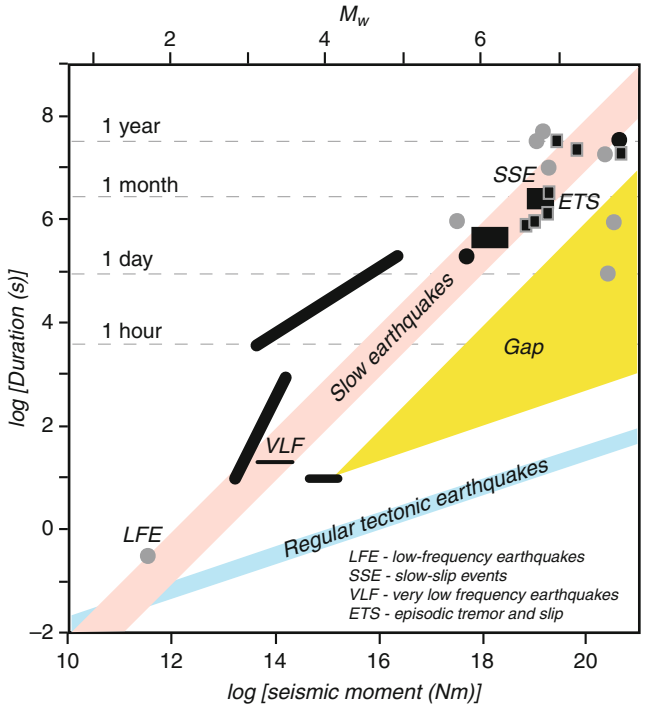
With this enhanced definition, earthquakes encompass a wide range of phenomena. The most common type we often think of is rapid slip within the Earth. The motion across a fault is typically of a shearing type but seismic disturbances can include volumetric or planar expansion in some settings, most notably near volcanoes. The classes of "earthquakes" now observed can be categorized by a rate of slip from very fast, kilometers per second, to slow, millimeters per year, covering ~12 orders of magnitude in slip rate. Figure 1, taken from Ide et al. (2007), shows the classes of earthquakes broken down by rate of the deformation. Because of the varying speeds of the motions, different types of instrumentation are required to cover specific bands.

Types of earthquakes and related phenomena defined

Here the types of earthquakes that produce crustal deformation are briefly defined. Many of them are described in detail in other chapters.

Tectonic earthquakes – Tectonic earthquakes are those that occur in response to plate motions or other predominantly shearing sources. They result primarily from the stick-slip behavior of faults. (Deep earthquakes occur by other mechanisms but are not discussed here as they do not visibly deform the crust.) As stress builds up across a fault surface, friction on the fault prevents the surface from sliding until the strength (some maximum shear stress) of the fault is exceeded. Then the two sides of the fault snap back (see elastic rebound below), releasing some or all of the elastic strain built up. This slip occurs at several kilometers per second, releasing high-frequency seismic waves. These are by far the most common type of earthquakes and the most destructive.

Tectonic tremors – These are a new type of seismic disturbance first recognized in Japan in 2002 (Obara, 2002). Originally they were called non-volcanic tremors



Earthquakes and Crustal Deformation, Figure 1 Durations and moments of various types of crustal deformation events, modified from Ide et al. (2007). *Black and gray circles* and *elongated bars* show silent earthquakes, afterslip and slowslip events from Italy, Nankai, Kilauea, San Andreas, and Sanriku (Ide et al., 2007). *Squares* are slowslip events from New Zealand (McCaffrey et al., 2008) and Guererro, Mexico (Larson et al., 2007). These follow a scaling relation of $\sim M_o/t$. Typical scaling relation for regular shallow interplate earthquakes is shown by the blue line. See Ide et al. (2007) for additional information.

due to some similarities to volcanic tremors. Whereas tectonic earthquakes are due to finite duration slip (on the order of seconds to minutes) on the fault, tectonic tremors appear as low amplitude, non-harmonic background disturbances lasting days to weeks. They are comprised in part of very small earthquakes but most of the signals are incoherent. Often, but not always, they accompany slow-slip events and are thought to be some aspect of the same process (Rubenstein et al., 2010).

Volcanic earthquakes – Volcanic earthquakes are similar to tectonic earthquakes in that they are due to the failure of rock along faults. However, their size is limited due to the relatively small fault surface area available near volcanoes.

Volcanic tremors – These earth disturbances are likely caused by fluid (water or magma) moving inside the earth. The signals are harmonic and can last a long time (relative to tectonic earthquakes) but are generally of low amplitude and generate little crustal deformation.

Slow earthquakes – Some earthquakes occur at slip speeds that are slower than typical tectonic earthquakes but fast enough to excite seismic waves. They occur when slip is in materials of low rigidity such as accretionary prisms of subduction zones. In such cases, they can produce tsunamis that are larger than expected for the earthquake's magnitude. Examples are earthquakes in Nicaragua (Kanamori and Kikuchi, 1993) and New Zealand (Doser and Webb, 2003).

Low-frequency earthquakes – Such earthquakes, called LFEs, have been observed in volcanic settings and more recently associated with tectonic tremors at subduction zones. As the name indicates they generate much lower frequencies than normal tectonic earthquakes. In both settings the low frequencies may be attributable to fluid interactions (Shelly et al., 2006).

Slow-slip events – In this class of events, called SSEs, the rupture speed is on the order of a few millimeters per day. Thus the rupture propagates at 11 orders of magnitude slower than a tectonic earthquake but only about 10 times faster than tectonic creep at plate motion rates. At these slow-slip speeds, seismic waves are not generated and slow-slip events are not detected by seismometers. To date, SSEs have been detected by geodetic instruments at subduction zones, Kilauea and on the San Andreas fault.

Post-seismic slip events – Another type of crustal motion that is often observed following large tectonic earthquakes is due to afterslip, thought to arise from continued slip on the fault at slow speeds. Afterslip can in small part be explained by aftershocks but the majority of the slip occurs without producing seismic waves. Such slip can last years and produce as much total slip as the main earthquake itself. The afterslip events often decay in accordance with Omori's Law.

Mantle relaxation – Following large earthquakes, the redistribution of the stress in the crust and mantle can cause the mantle to flow in a viscous manner, which then produces a long-period deformation of the crust above. This signal can often be confused with post-seismic slip.

Induced earthquakes – Earthquakes are induced by the activities of humans that modify the stress at shallow depths in the Earth. Examples of such activities are mining, drilling, and water impoundment behind dams. Induced earthquakes tend to be of significant hazard because they are very shallow and close to infrastructures (see *Artificial Water Reservoir Triggered Earthquakes*).

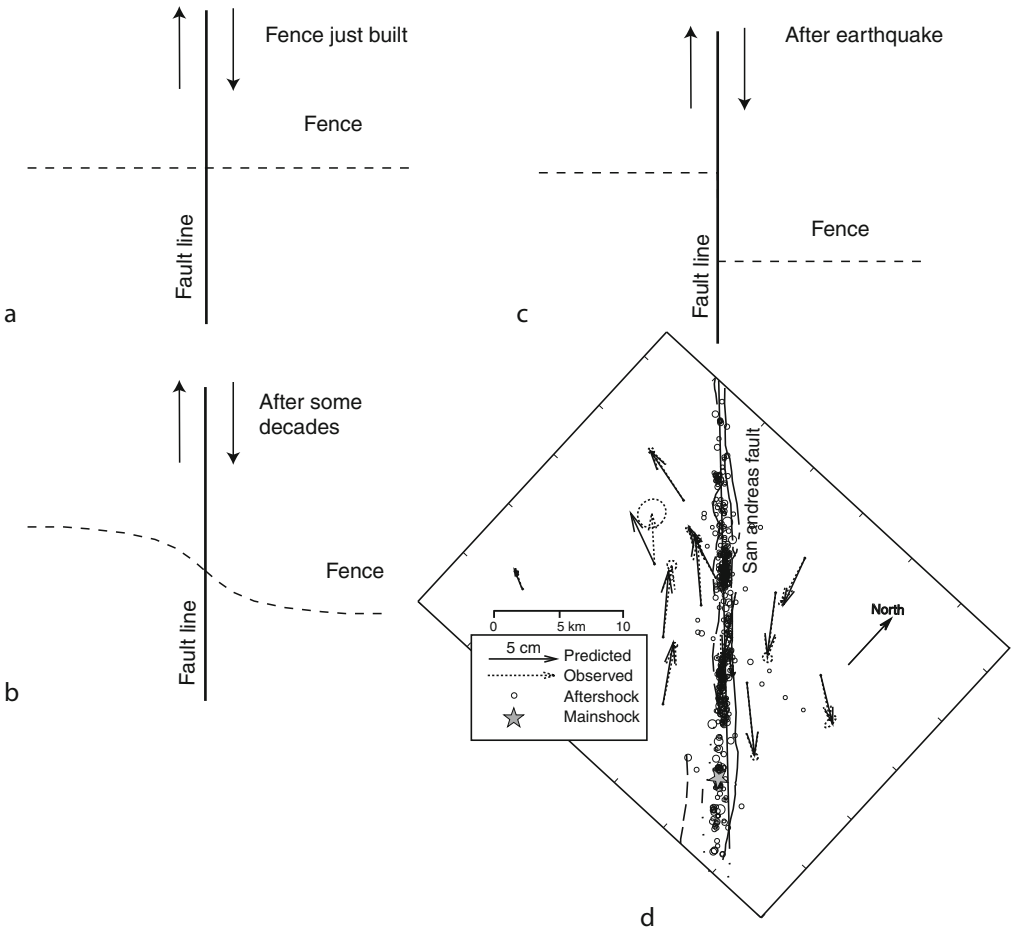
Fault morphology – Most deformation of the crust is accommodated by faulting. Crustal faults are the lasting expressions of past earthquakes and provide a means via geologic investigations to characterize the earthquakes. Compilations of estimates of slip rates and fault lengths are used to calculate long-term (Holocene, for example) deformation rates in continental regions.

Elastic rebound theory

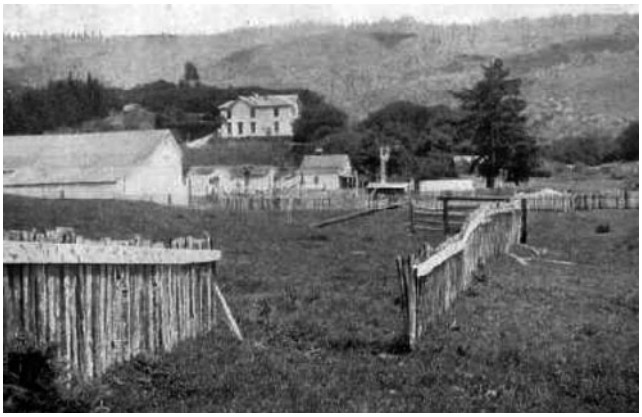
Large earthquakes, as noted above, are caused by slip on crustal faults and exhibit stages of strain build-up and release known as the earthquake cycle. A simple form of this cycle is elastic rebound. The concept of elastic

rebound as relating to the earthquake cycle was first introduced by Harry Reid (Reid, 1910) following earthquake observations in Sumatra and California. (See the interesting, in-depth description of the history of elastic rebound by Segall, 2010.) In this description, the crustal blocks adjacent to a fault are stuck together by friction on the fault surface and their relative movement causes strain in the blocks (Figure 2a and b). The material comprising the upper crust is thought to be linearly elastic, so the strain increases over time in a steady fashion as long as the blocks move at the same speed. At some point, the stress on the fault reaches its strength, and the fault slips suddenly in an earthquake (Figure 2c). The strain accumulated over decades to centuries is relieved in a matter of seconds. The slip on the fault causes the stored strain energy in the crust to be converted to kinetic energy which in turn is emitted as seismic waves and heat and the stress on the fault drops to a lower level. The use of the word "rebound" in this case refers to the idea that any feature on the surface, such as a fence crossing the fault, will return to its original shape (i.e., free of strain) after the earthquake (Figures 2 and 3). The earthquake in a broader sense produces permanent strain on a regional scale by displacing the crustal blocks along the fault (and offsetting the fence). Many observations of earthquake deformations have validated the elastic rebound model – the example of the 2004 Parkfield, California, earthquake rebound is shown in Figure 2d (Barbot et al., 2009).

We have discovered in the past few years that elastic rebound pertains not only to earthquakes of the type Reid observed, but also to earthquake afterslip (Heki et al., 1997) and so-called "slowslip events" (Dragert et al., 2001). In these phenomena, illustrated by GPS time series shown in Figure 4, slip on faults can occur at much slower rates than it does during earthquakes. In these time series showing the position of a point on the ground, the steady motion of the linear earthquake cycle is interrupted not only by earthquakes (abrupt offsets in Figure 4a) but also by afterslip that decays with time. In slowslip events, the linear site motion is interrupted for weeks to months



Earthquakes and Crustal Deformation, Figure 2 Illustration of the elastic rebound effect. The fence originally built in a *straight line* across the fault (**a**) is deformed by the strain around the fault in the period leading up to an earthquake (**b**). In (**c**) the earthquake has offset the fence but also straightened it again by the release of the strain. (**d**) Measured and modeled co-seismic offsets near Parkfield CA resulting from the 2004 M6 earthquake (Redrawn with permission from Barbot et al., 2009).



Earthquakes and Crustal Deformation, Figure 3 Picture taken in Marin County shows a fence that was offset about 10 ft along the trace of the San Andreas fault by the 1906 earthquake (Photo attributed to G.K. Gilbert; http://americahurrah.com/USGS/26.htm).

(Figure 4b) in Cascadia and New Zealand, for example. The spatial pattern of the slip and our knowledge of the rebound cycle guide our interpretation of these anomalous signals as slip on the fault surface.

The simple picture of elastic rebound theory is a good conceptual view of the earthquake cycle but in detail we see that it is much more complex. As more and better observations are made we see that the earthquake cycle includes time-dependent deformations on many time scales. These are discussed next.

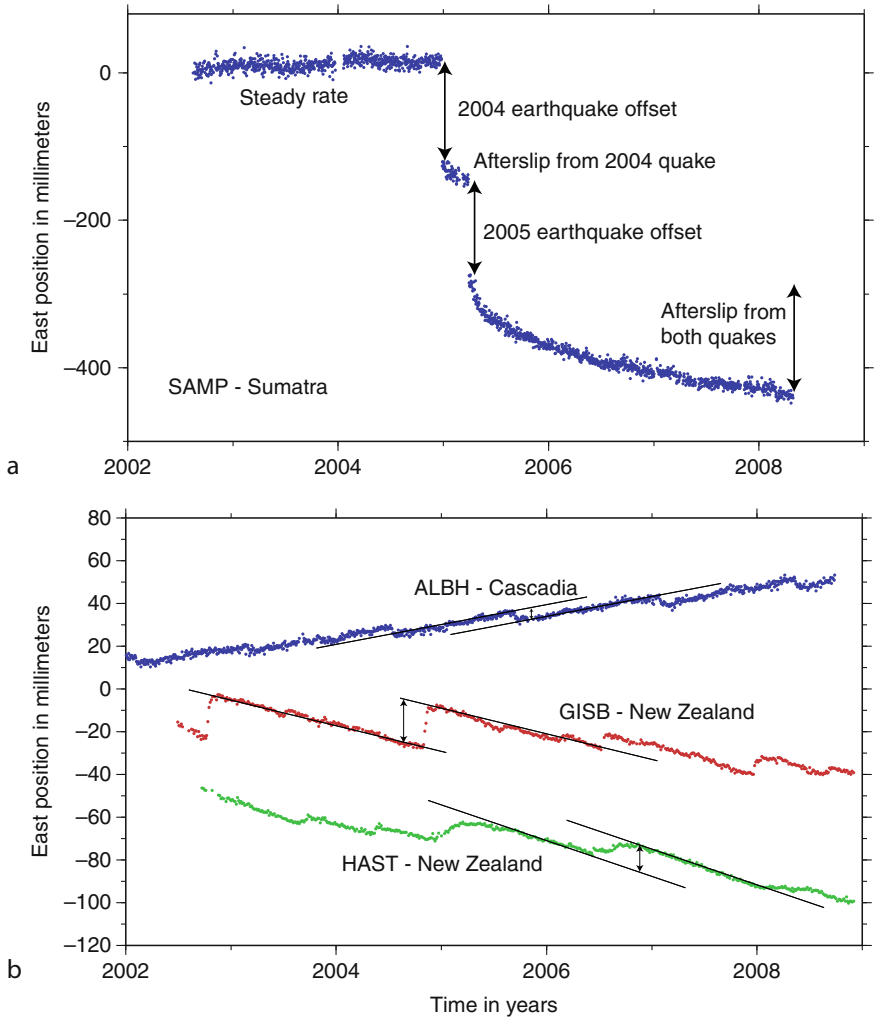
Pre-seismic crustal deformation

The deformation of the Earth's surface leading up to earthquakes has been of great interest to scientists for a long time due to the quest to identify precursors for prediction purposes. Such deformation can be grouped into long-term and short-term signals. For the long-term, the Earth deforms in the vicinity of faults largely in roughly the elastic rebound manner described above. Mapping out the temporal changes in crustal strain near faults is being used extensively to identify which parts of the fault are being loaded, or being subject to stress increases. In a simple sense, the parts of the fault being loaded may help delineate the future rupture region but the details are far from clear. For example, the current rate of loading of the Cascadia subduction thrust is inferred from the results of 15 years of GPS observations by McCaffrey et al. (2007) (Figure 5) and has the important implications for how close to populated regions the rupture may come during a large subduction earthquake. Such studies have been conducted for most subduction zones and continental transform faults globally and the relationship of such loading to seismic slip is just starting to be understood.

A now common approach to understanding the crustal deformation between earthquakes (also called the *interseismic* phase) is through the use of block models (Matsu'ura et al., 1986; McCaffrey, 2002; and many others). This method uses the idea that the crust or lithosphere is comprised of discrete entities (blocks) that are bounded on their lateral sides by other blocks and at their base by a stress-free region (asthenosphere). The lateral boundaries between the blocks are described by their frictional properties (i.e., a stuck fault). The blocks themselves rotate on the Earth's surface and so the motion can be described by an angular velocity (Euler pole) with its origin at the center of the Earth. Frictional interactions with adjacent blocks cause elastic strains to build in the blocks. Hence, new GPS observations contain information about both the angular velocities of the blocks (and slip rates of their bounding faults) and the frictional properties of the bounding faults. A recent example of this is the January 2010 M 7.0 Haiti earthquake – block modeling by Manaker et al. (2008) had shown that a M ~7.2 was possible for the Enriquillo Fault in Haiti (of course such modeling does not tell when such an event will occur).

Another approach to understanding crustal deformation is to assume that at some depth below the surface, the deformation field is continuous and that the surface faulting is only a vague reflection of that motion (England and McKenzie, 1982), in a similar fashion that floating ice blocks may show the general, but not specific, flow pattern of the water beneath. In this case, the surface faulting can be related to the continuous deformation field only in a statistical manner and the deformation field cannot be used to predict slip on unknown faults. On the other hand, the continuous deformation approach can be more readily related to the forces driving deformation through an assumption that the deformation is controlled by a viscous, rather than elastic, Earth. (The assumption of a viscous medium is required to relate our observations of deformation rates to forces.) For example, Jones et al. (1996) relate crustal deformation in the western USA to gravitational forces arising from spatial variations in crustal thickness acting on a viscous medium.

Short-term crustal deformation around faults leading up to earthquakes is much more elusive. To date there have been few or no convincing cases of isolated changes in deformation in the days to months prior to a large earthquake. Mogi (1985) shows evidence for accelerated uplift at tide gages in the 5 years preceding the 1983 Sea of Japan earthquake (see also Scholz, 2002). This evidence must be considered less certain than it was a few years ago before we knew of slowslip events. Although no SSEs have been observed in the past 10 years on modern GPS equipment in the vicinity of the 1983 earthquake, we cannot rule out that the pre-earthquake uplift was due to slowslip on the fault that subsequently ruptured (or some other fault) rather than a useful precursory change prior to the quake. (Because many SSEs have occurred without triggering tectonic earthquakes, we know that slowslip events are NOT short-term precursors to earthquakes.)

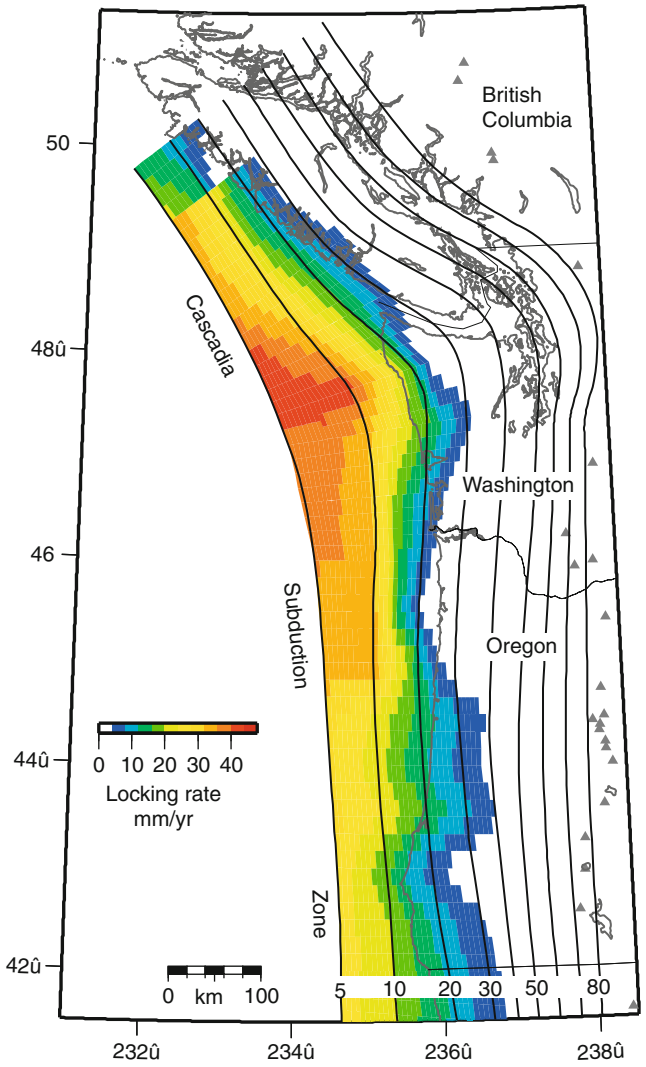


Earthquakes and Crustal Deformation, Figure 4 (**a**) The east position of continuous GPS site SAMP in northern Sumatra leading up to and following the December 2004 and March 2005 great earthquakes at the Sumatra subduction zone. The earthquakes each caused large sudden offsets of about 100 mm. The afterslip, which decays in a logarithmic fashion, produced over 100 mm of additional motion in the 3 years following the earthquakes. (**b**) GPS time series from Cascadia and New Zealand showing the geodetic signature of "slowslip" events. These events produce offsets of the time series but, unlike earthquakes, the offsets occur over weeks to months, rather than minutes. The *solid lines* show the linear trends (steady site velocities) between the slip events.

Co-seismic crustal deformation

The term co-seismic refers to the time right around the earthquake when the fault rupture is still propagating and seismic waves are being generated. Generally this amounts to a few seconds for a magnitude 6–7 earthquake and a minute or more for $M > 8$. The 2005 M9.3 Sumatra earthquake is thought to have ruptured for about 5 min – this is the time it takes to rupture such a long fault (∼1,300 km) at typical rupture speeds (∼ 3–4 km/s). Earthquakes generate both dynamic displacements, those that propagate but return to their original positions after the waves pass, and static displacements that are permanent. Seismometers can record the dynamic motions within their pass band but cannot measure the static displacement which requires geodetic instrumentation. Both the dynamic and static displacements induce stress changes in the crust and mantle that can lead to other forms of crustal deformation, such as triggered earthquakes, afterslip, slowslip and mantle relaxation. While GPS observations can measure the passing waves for very large earthquakes (as can seismographs), we are more concerned here with the static displacements.

The static displacements from a large earthquake provide much information on the faulting process and can be measured by comparing the position of ground points before and after the earthquake (Figure 2d). However, because the points are moving constantly due to tectonic plate motions and fault-induced elastic and anelastic strain, as discussed above, it is not always easy to know where the points were just prior to the earthquake.



Earthquakes and Crustal Deformation, Figure 5 Frictional loading of the Cascadia subduction zone inferred from crustal deformation of the coastal regions measured with 15 years of GPS observations (McCaffrey et al., 2007). The locking rate is the rate at which the subducting plate pushes the base of the overriding plate landward. This rate, integrated over time, provides an estimate, according to elastic rebound theory, of how much slip might occur in a large strain-relieving earthquake. Contours are of the depth to the plate boundary interface, in kilometers.

Consequently, in many active regions of the world, GPS data are collected continuously (called continuous GPS, or cGPS) which provide a direct measure of the pre-quake position. However, such observatories are expensive and generally not sufficiently dense to characterize the earthquake details when one happens. The geodetic characterizations of most earthquakes are now based on a combination of cGPS and survey-mode GPS, where measurements are made intermittently.

Simple physical representations of the Earth have been quite successful in characterizing the slip during large earthquakes based on the observed static displacements. (Seismic waves are also used to estimate the slip distributions of earthquakes – see *Earthquakes, Source Theory*). The most commonly used structure and source representation is the elastic, half-space dislocation (see Segall, 2010). The popularity of this method owes to the simplicity of the algorithm published by Okada (1992, 1995) for use in three dimensions and its success in reproducing the observations. More times than not, the geodetic and seismologic representations of an earthquake source agree within the levels of accuracy. In fact, when they do not agree, the blame is often placed on the complexity of the source rather than on the techniques.

Such studies of the distributions of slip on the fault during earthquakes have shown us just how complex the slip can be. It remains enigmatic why the earthquake slip distributions (fault unloading) are so much more heterogeneous that the loading distribution (as estimated from interseismic studies as discussed above). This is one of the challenges facing us if we want to use contemporary fault loading measurements to understand the real earthquake hazards.

Post-seismic crustal deformation

Unlike pre-seismic deformation, many striking examples of post-seismic deformation have been documented for large- and moderate-magnitude earthquakes. The mechanisms of this deformation are most likely either continued slowslip on the fault or adjustment of the viscous mantle or lower crust to the changes in stress brought about by the earthquake.

It has been known for a long time that large, shallow earthquakes produce aftershocks that decay with time (t) in a $1/t$ fashion. This aftershock decay was explained by the changes in friction on the fault as the slip rate decayed, known as Omori's Law. Continued slip on faults after large earthquakes sometimes shows a similar decay and this decay follows Omori's Law as well (though the total slip is many times greater than what can be explained by aftershocks alone). The first clear example of significant afterslip was for the 1994 Japan Trench earthquake observed by Heki et al. (1997) where the relatively aseismic afterslip exceeded the co-seismic slip magnitude in the year following the earthquake. The 2004 Parkfield earthquake was followed by aseismic afterslip that was about three times the co-seismic slip (Johanson et al., 2006).

We now describe the magnitude of an earthquake or any other slip event with the moment-magnitude scale. For a slip event on a fault, the moment M_o is defined as the average slip magnitude times the area of that slip times the rigidity modulus of the rocks around the fault, and the magnitude is $M_w = 2/3 \log(M_o) - 9.5$ (Hanks and Kanamori, 1979). Hence moment can be based on either seismic or geodetic measures of the slip and area and these

can be compared. By making such comparisons we now understand that aseismic afterslip (or slowslip) contributes as much or more to the total slip budget on some faults as do earthquakes. This discovery has positive implications for earthquake hazards in the sense that seismic events of a given size may be less frequent than previously estimated.

Observational advances and new discoveries

Our knowledge of the existence of the various types of earthquakes and their structures and our understanding of them rely on our ability to detect them and explain the observations. In the past 20 years, great advances in instrumentation have led directly to the detection of some new types of deformation sources. They, along with new analytical algorithms, have greatly advanced our understanding of earthquakes and crustal deformation. Here, we describe some of these new methods and the important discoveries that they have allowed.

Possibly the greatest advance in the past couple of decades has been in the use of remote sensing in earthquake and crustal deformation studies. These new methods include the Global Positioning System (GPS), Interferometric Synthetic Aperture Radar (InSAR), and Light Detection and Ranging (LIDAR). The first two of these allow us to measure displacements of the Earth's surface using information obtained from satellites orbiting the Earth. With repeated measurements we can watch the temporal variations in surface deformation with great accuracy. LIDAR is allowing unprecedented resolution of the topography of the Earth's surface beneath the biological layer and has delineated geologic structures (faults) in regions that are too remote or forest-covered to find otherwise.

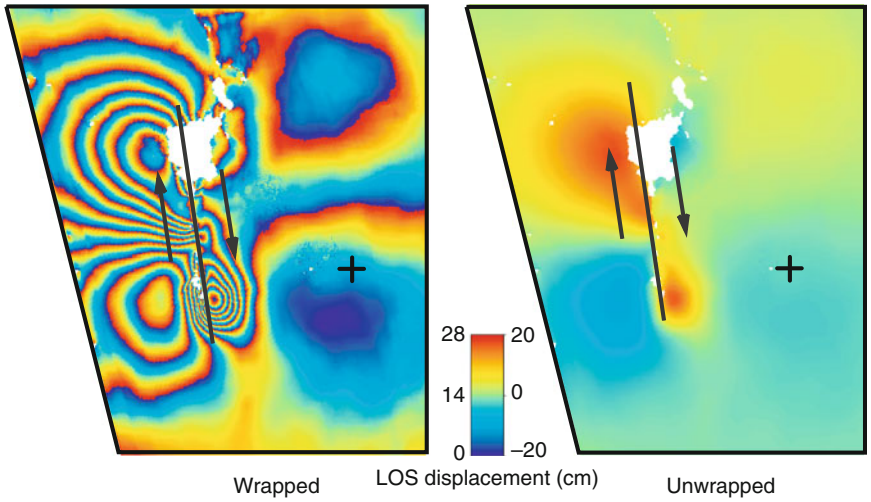
GPS, in particular when used in the differential mode, has resulted in the detection of the large, slow movements along faults – slowslip, post-seismic, afterslip, and mantle relaxation described above. Today, thousands of GPS instruments operate continuously at sites around the globe providing daily estimates of the site's three-dimensional position, which can be combined to form the equivalent of a displacement seismogram. The positional accuracy achieved for any given day is on the order of a few millimeters. The signal can also be sampled at a smaller time interval, such as seconds, to record the passing seismic waves in addition to the total static displacement produced by the earthquake. Such a resampling is possible only for very large earthquakes where the ground motions exceed a few millimeters. However, the greatest contribution of GPS has been in the discovery of crustal deformation sources that do not produce seismic waves.

The knowledge that fault movement could occur at rates somewhere between fast slip (km/s) and creep (mm/year) has been with us for years – such motions were observed on strain meters near the San Andreas fault in California (e.g., Linde et al., 1996). However, these slip events were very small compared to sizes of earthquakes on the fault. As described above, the observation by Heki et al. (1997) that post-seismic slip exceeded the slip during the earthquake itself was remarkable as it changed our ideas about how often earthquakes should occur on a given fault. In other words, we could no longer expect that the total slip on a fault in a given time period would be accompanied by earthquakes, an idea that held its roots in elastic rebound theory. Instead, some large portion of fault could slip without earthquake shaking.

As remarkable as that discovery was, in 1999 Dragert et al. (2001) recorded in the Western Canadia Deformation Array (continuous GPS network) a very slow deformation of the ground that looked just like a $M \sim 6.7$ earthquake, except that it occurred over 2 weeks time instead of just a few seconds. The maximum surface displacement was only 5 mm and at the time no seismic waves were detected. The slow surface deformation was interpreted by them to be due to slip on the Cascadia subduction fault directly beneath the GPS network. The slip on the fault was very slow, amounting to only 2 cm over the 2 weeks – this was 12 orders of magnitude slower than slip during a normal earthquake ($\sim$3 km/s).

Around the same time, Obara (2002) discovered the presence of low-level incoherent but intermittent seismic signals, called *tectonic tremor* above, coming from the subduction fault beneath Japan. Importantly, these signals appeared to originate on the fault just below the deepest slip extent from recent large earthquakes. Initial suggestions were that the tremor coincided in space and time with slowslip (Rogers and Dragert, 2003) but additional studies of both phenomena have not born that out. Further detailed studies of tectonic tremor requiring new seismological techniques revealed that it is likely in part due to numerous small thrust earthquakes of very low frequencies, called low-frequency earthquakes (LFE), on the subduction thrust boundary (Shelly et al., 2006). However, when they occur together, the total slip that can be accounted for with these small quakes is far less than what is seen in the slowslip events. Hence, it appears that the majority of the slowslip is indeed seismically silent.

While GPS provides a fine temporal sampling of the deformation, it is generally sparse in the spatial dimension because continuous GPS installations are expensive. The use of differential satellite interferometry (InSAR) has helped fill in the spatial gaps in pictures of the deformation of the Earth's surface by earthquakes and volcanoes. The InSAR method uses the interference patterns between two or more radar images of the ground that have been obtained from satellites. Comparing the images allows estimation of the displacement of the ground in the direction of the satellite (the LOS or line-of-sight component). The method relies on the ground surface changing at long wavelengths only, so that the images have strong correlation (coherence) and can be matched. InSAR provides a scalar observation of the deformation but at high spatial density (Figure 6) whereas GPS provides a vector of the deformation at high temporal sampling. Many studies use GPS and InSAR together to characterize the



Earthquakes and Crustal Deformation, Figure 6 Interferogram showing the ground deformation observed during the December 2003 M6.6 Bam, Iran, earthquake (http://neic.usgs.gov/neis/eq_depot/2003/eq_031226/). On *left*, the "wrapped" image shows line-of-sight (LOS) displacements by the phase of the difference in distance to the satellite; on the *right* the image has been "unwrapped," that is, converted to distance changes. In both cases the four-lobe displacement field of the strike-slip earthquake is evident. (Image with permission from Funning et al. [2005].)

deformation, taking advantage of the strengths of both types of observations.

The LIDAR method is extremely useful in detecting faults that intersect the ground surface. Because it uses shorter wavelengths than radar, it allows us to see finer features. Very often, faults produce only subtle lineations on the ground surface, which can now be detected at the sub-meter level. While LIDAR does not give details of the fault, like the slip rate, it can pinpoint targets for additional field study. An example of the utility of the LIDAR method is the delineation of the Seattle Fault in northwestern Washington state (Haugerud et al., 2003). This fault, which crosses the highly populated Puget Sound region, produced a large earthquake about 1,100 years ago. The fault scarp is large and obvious in some places but subtle and covered in forests in others.

The advance of the new geodetic techniques has had such a positive impact on geophysics that major organizational steps have been taken to utilize them. The consortium of universities called UNAVCO (University Navstar Consortium; unavco.org) was developed to support the global research community in the use of geodetic tools. In several countries, permanent dense geodetic networks and similar consortia have been established to monitor earthquakes and their associated phenomena.

Perspectives

Our ideas about earthquakes sensu strictu have undergone drastic changes in the past few years. Elastic rebound theory, while still a sensible conceptual model, does not cover the large range of observed phenomena that make up the earthquake cycle. A clear challenge that still faces us is to tie together the various deformation processes we observe throughout the cycle into a comprehensive earthquake theory. Some of the more important goals are to understand the observations we make in terms of future earthquake risk. For example, although we can measure current loading of faults quite well by its elastic strain near the fault, we do not know what it means for the slip distribution in future events. Geodetic methods of measuring crustal deformation have a short history relative to the times between large earthquakes, so geologic methods including paleoseismology (see *Paleoseismology*) and LIDAR that extend the observation history are crucial. Continued study and understanding of the aseismic slip events and tectonic tremor on faults may pay big dividends toward delineating the parts of the faults that are prone to seismic slip during large destructive earthquakes. The acquisition of deformation data at high resolution in both time and space will assist greatly in understanding processes operating at depth within the Earth.

Bibliography

Barbot, S., Fialko, Y., and Bock, Y., 2009. Postseismic deformation due to the Mw 6.0 2004 Parkfield earthquake: Stress-driven creep on a fault with spatially variable rate-and-state friction parameters. *Journal of Geophysical Research*, **114**, B07405, doi:10.1029/2008JB005748.

Doser, D. I., and Webb, T. H., 2003. Source parameters of large historical (1917–1961) earthquakes, North Island, New Zealand. *Geophysical Journal International*, **152**, 795–832.

Dragert, H., Wang, K., and James, T. S., 2001. A silent slip event on the deeper Cascadia subduction interface. *Science*, **292**, 1525–1528.

England, P., and McKenzie, D., 1982. A thin viscous sheet model for continental deformation. *Geophysical Journal of the Royal Astronomical Society*, **70**, 295–321.

Funning, G. J., Parsons, B., Wright, T. J., Jackson, J. A., and Fielding, E. J., 2005. Surface displacements and source parameters of the 2003 Bam (Iran) earthquake from Envisat advanced

synthetic aperture radar imagery. *Journal of Geophysical Research*, **110**, B09406, doi:10.1029/2004JB003338.
Hanks, T. C., and Kanamori, H., 1979. A moment magnitude scale. *Journal of Geophysical Research*, **84**, 2348–2350.
Haugerud, R. A., Harding, D. J., Johnson, S. Y., Harless, J. L., Weaver, C. S., and Sherrod, B. L., 2003. High-resolution Lidar topography of the Puget Lowland. Washington: GSA Today, pp. 4–10.
Heki, K., Miyazaki, S., and Tsuji, H., 1997. Silent fault slip following an interplate thrust earthquake at the Japan Trench. *Nature*, **386**, 595–598.
Ide, S., Beroza, G. C., Shelly, D. R., and Uchide, T., 2007. A scaling law for slow earthquakes. *Nature*, **447**, 76–79, doi:10.1038.
Johanson, I. A., Fielding, E. J., Rolandone, F., and Bürgmann, R., 2006. Coseismic and postseismic slip of the 2004 Parkfield earthquake from space-geodetic data. *Bulletin. Seismological Society of America*, **96**, 4b.
Jones, C. H., Unruh, J. R., and Sonder, L. J., 1996. The role of gravitational potential energy in active deformation in the southwestern US. *Nature*, **381**, 37–41.
Kanamori, H., and Kikuchi, M., 1993. The 1992 Nicaragua earthquake – a slow tsunami earthquake associated with subducted sediments. *Nature*, **361**, 714–716.
Larson, K. M., Kostoglodov, V., Miyazaki, S., and Santiago, J. A. S., 2007. The 2006 aseismic slow slip event in Guerrero, Mexico: New results from GPS. *Geophys Res Lett*, **34**, L13309, doi:10.1029/2007GL029912.
Linde, A. T., Gladwin, M. T., Johnston, M. J. S., and Gwyther, R. L., 1996. A slow earthquake sequence on the San Andreas fault. *Nature*, **383**, 65–68.
Manaker, D. M., Calais, E., Freed, A. M., Ali, S. T., Przybylski, P., Mattioli, G., Jansma, P., Prepetit, C., and de Chabalier, J. B., 2008. Interseismic Plate coupling and strain partitioning in the Northeastern Caribbean. *Geophysical Journal International*, **174**, 889–903, doi:10.1111/j.1365-246X.2008.03819.x.
Matsu'ura, M., Jackson, D. D., and Cheng, A., 1986. Dislocation model for aseismic crustal deformation at Hollister, California. *Journal of Geophysical Research*, **91**, 12661–12674.
McCaffrey, R., 2002. Crustal block rotations and plate coupling. In Stein, S., and Freymueller, J. (eds.), *Plate Boundary Zones*. AGU Geodynamics Series 30, pp. 101–122.
McCaffrey, R., Qamar, A. I., King, R. W., Wells, R., Ning, Z., Williams, C. A., Stevens, C. W., Vollick, J. J., and Zwick, P. C., 2007. Plate locking, block rotation and crustal deformation in the Pacific Northwest. *Geophysical Journal International*, doi:10.1111/j.1365-246X.2007.03371.x.
McCaffrey, R., Wallace, L. M., and Beavan, J., 2008. Slow slip and frictional transition at low temperature at the Hikurangi subduction zone. *Nature Geoscience*, **1**, 316–320.
Mogi, K., 1985. *Earthquake Prediction*. Tokyo: Academic.
Obara, K., 2002. Nonvolcanic deep tremor associated with subduction in southwest Japan. *Science*, **296**, 1679, doi:10.1126/science.1070378.
Okada, Y., 1992. Internal deformation due to shear and tensile faults in a half-space. *Bulletin. Seismological Society of America*, **82**, 1018–1040.
Reid, H. F., 1910. The Mechanics of the Earthquake, The California Earthquake of April 18, 1906, Report of the State Investigation Commission, Vol. 2, Washington, DC: Carnegie Institution of Washington.
Rogers, G., and Dragert, H., 2003. Episodic tremor and slip on the cascadia subduction zone: The chatter of silent slip. *Sciencexpress*, doi: 10.1126/science.1084783.
Rubenstein, J. L., Shelley, D. R., and Ellsworth, W. L., 2010. Nonvolcanic tremor: a window into the roots of fault zones. In Cloetingh, S., Negendank, J. (eds.), *New Frontiers in Integrated Solid Earth Sciences.* International Year of Planet Earth, doi: 10.1007/978-90-481-2737-5_8.
Scholz, C., 2002. *The Mechanics of Earthquakes and Faulting*, 2nd edn. Cambridge: Cambridge University Press, 471 pp.
Segall, P., 2010. *Earthquake and Volcano Deformation*. Princeton: Princeton University Press.
Shelly, D. R., Beroza, G. C., Ide, S., and Nakamula, S., 2006. Low frequency earthquakes in Shikoku, Japan, and their relationship to episodic tremor and slip. *Nature*, **442**, 188–191.

Cross-references

Earthquake Precursors and Prediction
Earthquake, Aftershocks
Earthquakes, Volcanogenic
GPS, Tectonic Geodesy
Paleoseismology
SAR Interferometry
Seismic Hazard
Seismicity, Intraplate
Seismicity, Subduction Zone
Slow Earthquake

EARTHQUAKES, EARLY AND STRONG MOTION WARNING

Richard M. Allen
Berkeley Seismological Laboratory, Department of Earth and Planetary Science, University of California, Berkeley, CA, USA

Synopsis

Strong motion earthquake warning (also known as earthquake early warning) is the rapid detection of an earthquake underway, the estimation of the ground shaking likely to result, and the issuance of an alert to people in the region where the ground shaking is likely to be hazardous. The last decade has seen rapid development of methodologies for early warning, which include the use of strong motion observations close to the epicenter to provide warnings at greater distances, the use of P-wave observations across a network to locate the event and map the distribution of likely strong ground motion, and the use of a P-wave detection at a single station to estimate the likely shaking intensity at the same station. These approaches have been listed in the order of increasing warning time, but the additional warning time comes at the expense of certainty. The earlier a warning is provided the greater the likelihood of false and missed alarms and the greater the uncertainty in the estimated shaking intensity. Typical warning times range from a few seconds to tens of seconds; the upper limit is about 1 min. Identified applications include personal protective measures where individuals move to a safe-zone within a few feet; automated mechanical response including stopping trains and isolating sensitive and hazardous machinery and chemicals; and situation awareness by large organizations that can help prevent cascading failures and be available before shaking disrupts communications. Large-scale public warning systems are currently operational in

Mexico and Japan, and smaller systems are used in Romania, Turkey, and Taiwan. Testing is underway by seismic networks in many other countries; however, implementation will require significant financial, political, and sociological hurdles to be crossed.

Introduction

Strong motion earthquake warning, more commonly referred to as earthquake early warning (EEW), is the rapid detection of an earthquake underway, estimation of the intensity of ground shaking likely to result, and issuance of a warning in the area likely to experience damaging strong motion. It provides high-probability warnings of shaking likely to occur within seconds to tens of seconds. The ability to provide warnings is relatively new and due to recent scientific development of methodologies to rapidly estimate the size of earthquakes underway, coupled with modern engineering of rapid communications to collect seismic data and distribute warnings, and a political will to operate state-of-the-art geophysical networks for the purpose of both scientific discovery and societal benefit.

EEW is one of three categories of earthquake warnings. Long-term warnings provide relatively low probabilities of earthquakes over long time periods, typically decades, and are usually referred to as earthquake forecasts. On these long timescales, earthquake likelihood can be estimated based on historic earthquake recurrence intervals and plate motion rates, linked together with deformation models incorporating fault geometries. For example, the probability of an $M > 6.7$ earthquake on the Hayward-Rodgers Creek fault that crosses the UC Berkeley campus is estimated to be 31% over 30 years (Working Group on California Earthquake Probabilities, 2007). These forecasts are also translated into estimates of ground shaking intensity that can be used to define building codes designed to prevent new buildings from collapse in the forecast of earthquakes.

The other end-member is earthquake prediction. The term prediction usually means a high probability of a large, damaging earthquake over a short time interval. For example, predicting an $M > 7$ earthquake on the Hayward Fault next week (Allen, 1976). Earthquake prediction is currently not possible, as we have not been able to identify a precursory signal that consistently occurs before earthquakes, and most seismologists agree that prediction will not be possible in the near future. EEW is intermediary in that it is possible to provide warning before shaking is felt, however, the warnings are only a few seconds to tens of seconds.

Uses of early warning

Uses of EEW that have been identified to date fall into three broad categories (Allen et al., 2009d): personal protection, automated mechanical actions, and situation awareness. The first use that people think about when introduced to early warning is personal protection. For most people, given that there will only be a few to a few tens of seconds of warning, the appropriate action is to move to a "safe-zone" within a few meters of your current location – under a sturdy table, away from falling hazards including windows, bookshelves, etc. These safe-zones need to be identified ahead of time in the locations that people spend most of their time, that is, in their homes and offices or workplaces. In most cases there is not sufficient time to evacuate buildings, and EEW is no substitute for building standards that prevent building collapse. It currently takes schoolchildren in Japan and Mexico ~5 s to get under their desks in response to audible warnings. Outdoors, people should move away from buildings, masonry walls, and other falling hazards including broken glass, street signs, etc. In hazardous work sites including construction sites, manufacturing and chemical facilities, the same principle applies: move to predefined safe-zones.

The second category is automated mechanical control. Within the transportation sector, implemented warning systems are currently used to slow and stop trains, stop airplanes from landing by issuing a go-around command, and to bring traffic to a halt. They could also be used to turn metering or tollbooth lights to red at the entrance to vulnerable road bridges, tunnels, and overpasses. Within buildings, warnings are now used to stop elevators at the next floor and open the doors, open emergency exit doors and window blinds, and turn on lights. Industrial facilities use the warning to place machinery or product into a safe mode. For example, a chip manufacturer in Japan places sensitive micro-robots into a hold mode to reduce damage, and isolates hazardous chemical systems to minimize any resulting leaks. In Japan, engineers are also experimenting with using warnings within automated structural control systems that change the mechanical properties of the building to better withstand the expected shaking. This is an area where there are a myriad of possible applications, most of which are very specific to individual users.

Finally, situation awareness is very important to organizations with responsibilities for large infrastructures and/or people. These include government agencies, utility, and transportation companies. While these groups may take early warning actions, they also have a need to rapidly understand why part or all of their system is failing. For example, an electrical power grid may start to fail in one area, and operators need to rapidly understand why it is failing in order to reduce cascading failures throughout the system. EEW can provide this information more rapidly than any of the existing post-earthquake information systems. EEW also has the advantage that much of the data needed to map out the likely shaking is collected before significant shaking has occurred. This means that the information can be transmitted across data-networks and out to users before strong shaking has the opportunity to disrupt communications.

One of the main concerns expressed in response to the early warning concept is panic: will people panic when they receive the warning and injure themselves or others as a result. Social scientists have long studied this issue

and find that it is not the source of concern that many seismologists think it is (Quarantelli, 1956; Goltz, 2002). The experience that we now have from public warnings in Mexico and Japan provides no evidence that warnings result in panic (Espinosa Aranda et al., 2009; Kamigaichi et al., 2009). There is no evidence for traffic accidents resulting from warnings and no evidence of public stampedes. Implementation of EEW does require an educational campaign to educate people about what EEW can do and help them identify the appropriate actions in their locations (Kamigaichi et al., 2009).

Approaches to early warning

All EEW methodologies must use just a few seconds of data to detect an earthquake underway and rapidly assess its intensity. There is generally a trade-off between the amount of warning time and the accuracy or certainty of the warnings. Greater certainty requires more data from one or more stations, but waiting for that data reduces the warning time. The challenge when building an early warning system is to determine whether additional accuracy is worth the increasing delay.

Strong-motion-based warning. The first warning methodologies used this approach. When strong ground shaking is observed at one location, close to the earthquake epicenter, the shaking intensity at greater distances can be estimated using ground motion prediction equations and a warning transmitted ahead of the ground shaking. This approach is called front detection and is used in Mexico City where the earthquake source zone is ~300 km from the city. Figure 1 shows the time at which peak ground shaking was observed at all available stations recording nine $M \geq 6$ earthquakes. A strong-motion-based warning would be possible at some point between 10 s and 20 s after the origin time for these events, depending on the proximity of the closest station(s). By 20 s, the onset of strong shaking is ~60 km from the epicenter meaning that warnings based on peak-shaking observations near the epicenter are only available at distances greater than ~60 km. The region within ~60 km of the epicenter would get no warning and is referred to as the "blind zone."

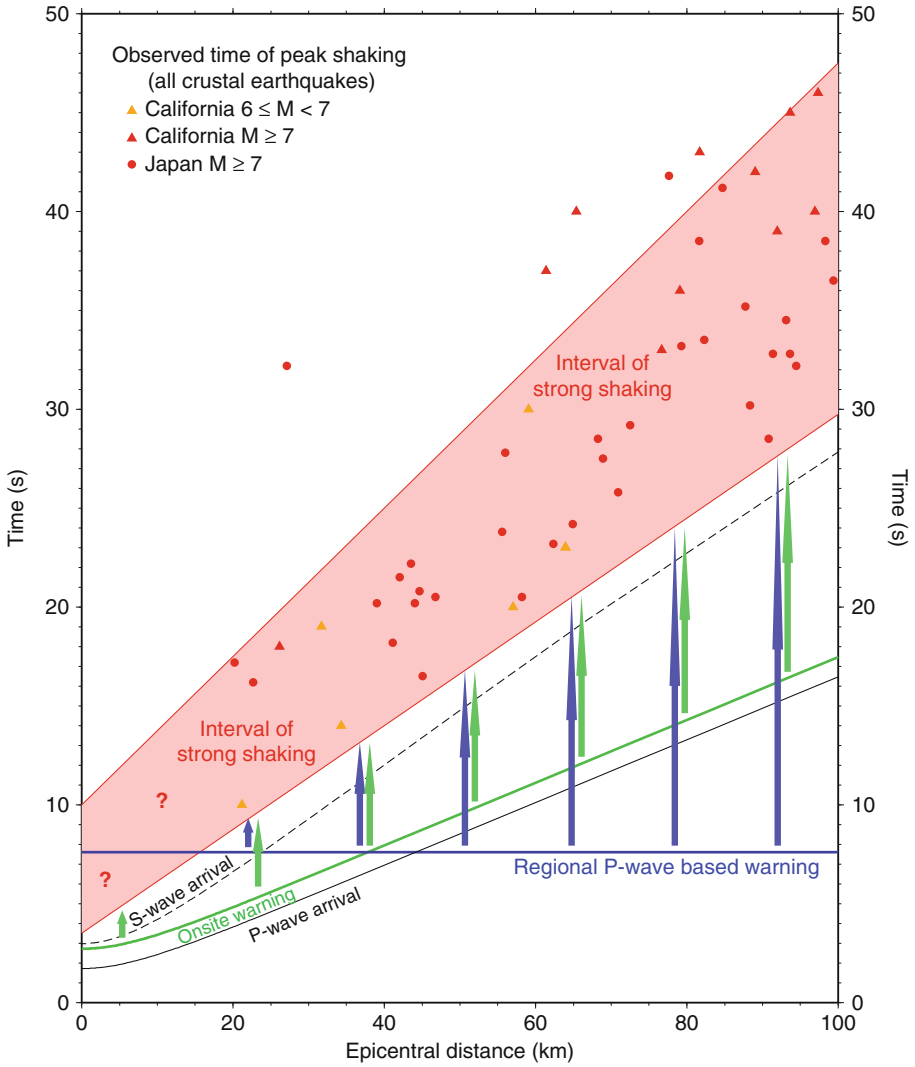
P-wave based warning. To reduce the size of the blind zone, P-wave based methodologies are used. These approaches use a few seconds of the P-wave to estimate either the magnitude or the peak-shaking amplitude. In order to estimate the magnitude, a variety of parameters have been developed that are measures of the frequency content (Nakamura, 1988; Allen and Kanamori, 2003; Kanamori, 2005) or the amplitude (Kamigaichi, 2004; Wu and Kanamori, 2005b; Zollo et al., 2006; Cua and Heaton, 2007; Wurman et al., 2007; Böse et al., 2008; Wu and Kanamori, 2008b; Köhler et al., 2009). In order to estimate magnitude, amplitude-based parameters need to be corrected for epicentral distance. However, the frequency-based parameters have been found to be insensitive to epicentral distance provided the observations are made within ~100 km of the epicenter. For all of these parameters empirical scaling relations between the observation and earthquake magnitude have been developed so that magnitude (plus or minus some uncertainty) can be estimated from an observation at one or more stations.

Regional warning. When a regional network is available, P-wave based estimates of magnitude can be combined with an estimate of the earthquake location to then map the distribution of strong ground motion using ground motion prediction equations. In their simplest form these maps are a "bulls-eye" centered on the epicenter. Corrections can also be made for site amplification effects. Examples of this approach include ElarmS (Allen, 2007; Wurman et al., 2007; Allen et al., 2009a; Brown et al., 2009), Presto (Zollo et al., 2009), the Virtual Seismologist (Cua and Heaton, 2007; Cua et al., 2009), and the NIED approach in Japan (Horiuchi et al., 2005).

When using a regional network, P-wave observations at multiple stations can be combined in order to verify that an earthquake is underway and improve the accuracy of the magnitude estimate. Generally, the more stations used, the better the accuracy of the warnings and the fewer false alarms. The time at which a warning is available therefore depends on the number of station detections required and the number of seconds of P-wave data required at each station. Figure 1 illustrates the amount of warning time available using this approach. The horizontal blue line at 7.6 s represents the time at which a warning is available for a particular scenario (Figure 1). In this case, the size of the blind zone is ~15 km and the amount of warning time increases linearly with distance beyond 15 km. There would be ~9 s warning at 50 km and ~22 s warning at 100 km.

Onsite warning. Onsite warning provides the most rapid approach to earthquake warning by removing the need for communications. The delay in issuing a warning is minimized by using a seismic station at the location where the warning is needed. Onsite methodologies are intended to estimate the intensity of ground shaking at the same location that the P-wave is detected (Odaka et al., 2003; Wu and Kanamori, 2005a, b; Nakamura and Saita, 2007a, b; Wu et al., 2007; Wu and Kanamori, 2008a, b; Böse et al., 2009a, b). The challenge with onsite warning is to provide enough accuracy that the warnings are useful. The character of P-wave arrivals is variable at different locations for the same earthquake due to variability in source radiation and also local site effects. Using a single station (or several stations within a few hundred meters) does not allow for averaging of this variability as in the case of regional methodologies. However, the onsite approach is faster.

Figure 1 illustrates the best-case scenario for onsite warning in which the warning is available using 1 s of P-wave data and the earthquake is at a sufficient depth that there is more than 1 s between the P-wave arrival and the onset of peak shaking at the epicenter. This means that there is no blind zone and a warning is available at the epicenter. The warning time increases with distance as the



Earthquakes, Early and Strong Motion Warning, Figure 1 Plot showing the time of observed peak shaking for large magnitude events and illustrating the time at which warnings could be available using various methods. The time of peak shaking was determined for all available stations in nine M $\geq$ 6 earthquakes from California (triangles, magnitudes 6.0, 6.5, 6.7, 7.1, 7.1, and 7.5) and Japan (circles, magnitudes 7.0, 7.1, and 7.3) and is plotted as a function of epicentral distance. All events are crustal earthquakes with epicentral depths of 21 km or less. The *red* shaded region illustrates the approximate interval of strong shaking when most of the peak-shaking observations are made. Note that there are few observations close to the epicenter due to the stations spacing of networks in California and Japan. The P-wave (*solid black line*) and S-wave (*dashed black line*) arrival times for a 10 km depth earthquake are also shown. The *blue line* and *arrows* illustrate available warning times for a regional P-wave based warning system. The horizontal line is the time at which a warning is available. It assumes that the required number of stations for a warning have detected a P-wave once the P-wave front reaches 25 km (at 4.6 s), that 1 s of P-wave is required to estimate magnitude (available at 5.6 s), and that an additional 2 s is needed for communications. The purpose-built seismic network in Japan provides data to the network centers within 1 s, and we assume that an additional 1 s is required to get the warning out to users. This means that the warning reaches users at 7.6 s after the origin time. The onset of strong shaking has reached ~15 km from the epicenter at this time. The available warning time increases linearly (shown by *blue arrows*) from 0 s at 15 km to ~22 s at 100 km. The *green line* and *arrows* illustrate the available warning time using the onsite P-wave based warning approach. It is assumes that 1 s of P-wave data is needed and the time of warning (*green line*) is therefore 1 s after the P-wave arrival. As shown, it is possible to provide a warning at the epicenter if the earthquake is more than ~10 km deep meaning there could be no blind zone. The warning time then increases with distance as the time between the P-wave arrival and peak shaking increases. Warning times (*green arrows*) range from ~2 s at 10 km to ~12 s at 100 km.

time-separation of the P-wave and peak shaking increases. Typical warning times are ~2 s at 10 km, ~7 s at 50 km, and ~12 s at 100 km. The onsite approach provides the most warning close to the epicenter. However, for locations at greater distances, regional warning provides more warning time that onsite warning as alerts are transmitted electronically ahead of ground motion. In the scenarios shown in Figure 1, the crossover when the regional method provides more warning than onsite is at ~35 km.

Implementations of early warning

There are currently two large-scale implementations of early warning in Mexico and Japan (Figure 2). The first public system started to provide warnings to Mexico City in 1993. It was built following the 1985 M8.1 earthquake that killed 10,000 people in the city. The system uses a line of seismometers deployed near the coastline to detect offshore earthquakes and transmit the warning 300 km to the city ahead of strong shaking (Espinosa Aranda et al., 1995). The offset between the source region and the city allows the use of a front detection approach and still provides more than 60 s of warning time. These warnings are broadcast by 58 radio stations and 6 TV channels in the city. There are also more than 250 users who receive the warning through dedicated radio links including schools and universities, emergency and civil protection agencies, government buildings, and the subway system.

Japan Railways first began to experiment with early warning in the 1960s. These systems became increasingly sophisticated including the first P-wave based application of UrEDAS in the 1980s (Nakamura, 1988). With the growth of computerized communication systems in the 1990s automated remote sensing of earthquakes expanded. Following the 1995 Kobe earthquake that killed 6,000 people, Japan installed dense seismic networks across the entire country (Okada et al., 2004). This network led to the development and implementation of a nationwide regional warning system that became operational in 2007 and is operated by the Japan Meteorological Agency (JMA). These warnings are now broadcast by the majority of TV and radio stations in Japan, through the government operated J-Alert system that issues warnings to municipalities (many of which relay the warning to loudspeaker systems in public spaces), by cell phone companies, and by a myriad of private providers and purpose-built consumer devices (Kamigaichi et al., 2009). The various Japanese systems use both onsite and regional warning systems. In the 2004 M_w7.2 Niigata earthquake the emergency brakes of one of the bullet trains were triggered by the onsite UrEDAS system (Nakamura and Saita, 2007a), and the regional warning system operated by JMA has issued warnings for multiple earthquakes including the 2008 M_w7.2 Iwate-Miyagi Nairiku earthquake (Kamigaichi et al., 2009).

Smaller scale and user-specific warning systems have also been implemented in various locations around the world. In Bucharest, warnings are provided to a nuclear research facility (Ionescu et al., 2007). Istanbul has warning systems for a power plant and a large banking building, and a more general use warning system has also been implemented in the last few years using a new strong motion network (Alcik et al., 2009). In Taipei, a test warning system provided warnings to the rail system, the university, and a hospital (Wu and Teng, 2002). Finally, in the US, a temporary warning system was implemented following the 1989 Loma Prieta earthquake (Bakun et al., 1994).

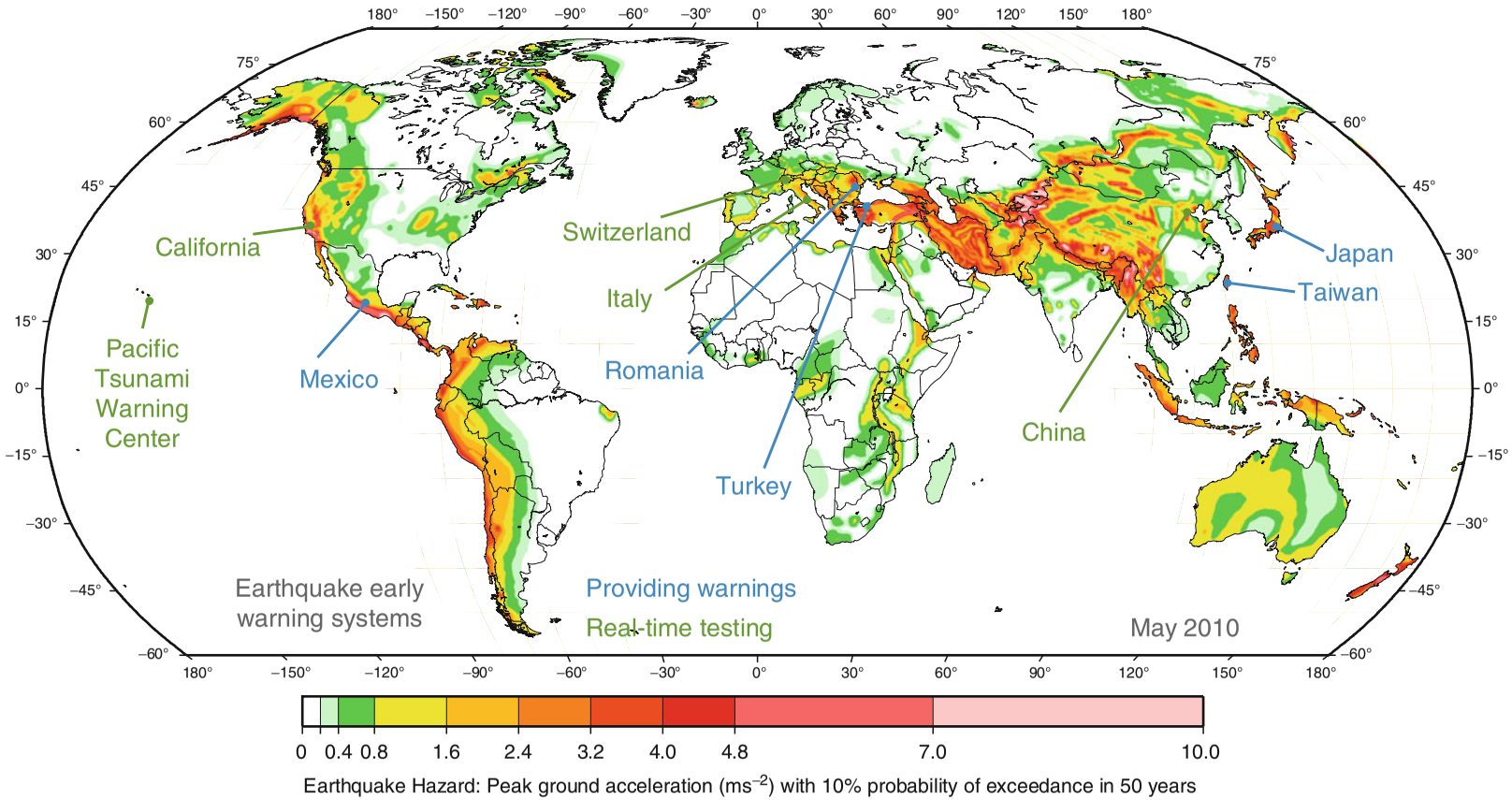
Testing of early warning methodologies is now underway on an increasing number of regional seismic networks around the world. This includes networks in California, China, Italy, at the Pacific Tsunami Warning Center, and in Switzerland (Figure 2). In many of these cases the technical feasibility of providing warnings has been demonstrated. Whether these test systems will transition to implementation of public warning systems will therefore be largely determined by financial, political, and sociological factors.

Summary

The last decade has seen rapid development of earthquake early warning methodologies (e.g., Gasparini et al., 2007; Allen et al., 2009b, c) by multiple research groups around the world. A range of methodologies has been developed that generate warnings based on observations of strong shaking near the epicenter, estimation of shaking hazard maps using P-wave detections across a network, and single-station detections of P-waves to estimate hazard at the same location. Warning times typically range from a few seconds to tens of seconds with an upper limit of around 1 min. With all these methodologies there is a trade-off between the accuracy of the warning and the amount of warning time. The challenge is to identify the optimal threshold at which to provide a warning.

There are now several implementations of public warning systems around the world, most notably in Mexico and Japan. These provide insight to the potential uses of earthquake alerts, which fall into three categories. Personal protection is about actions of individuals to protect themselves. For individuals it is important to identify "safe-zones" at home and work. These are spaces that are reachable within a few seconds where they will be protected from falling hazards. Automated mechanical control can be used to further reduce the impacts of earthquakes. This category includes slowing trains, opening elevators at the nearest floor, and isolating sensitive or hazardous machinery or chemicals. Finally, early warning can provide situational awareness information to large organizations that may be used to limit cascading failures and initiate response. This information could be available before much of the shaking, and therefore increases the likelihood that the information can be transmitted before communications are lost.

While 5 years ago there were many technical questions about the feasibility of earthquake early warning, today the primary hurdles to implementation are financial,



Earthquakes, Early and Strong Motion Warning, Figure 2 Map of seismic hazard around the world (*pink* represents the greatest hazard) showing regions where earthquake early warnings are currently being provided (*blue* labels) and the regions where real-time testing on existing seismic networks is underway (*green* labels) (after Giardini, 1999).

political, and sociological. The existing public warning systems in Mexico and Japan were built following major destructive earthquakes. The challenge for the hazard mitigation communities in other countries is to generate the necessary will to implement these systems before the next earthquake.

Bibliography

Alcik, H., Ozel, O., Apaydin, N., and Erdik, M., 2009. A study on warning algorithms for Istanbul earthquake early warning system. *Geophysical Research Letters*, **36**, L00B05.

Allen, C. R., 1976. Responsibilities in earthquake prediction. *Bulletin Seismological Society of America*, **66**, 2069–2074.

Allen, R. M., 2007. The ElarmS earthquake warning methodology and application across California. In Gasparini, P., and Zschau, J. (eds.), *Seismic Early Warning.* New York: Springer, pp. 21–44.

Allen, R. M., and Kanamori, H., 2003. The potential for earthquake early warning in southern California. *Science*, **300**, 786–789.

Allen, R. M., Brown, H., Hellweg, M., Khainovski, O., Lombard, P., and Neuhauser, D., 2009a. Real-time earthquake detection and hazard assessment by ElarmS across California. *Geophysical Research Letters*, **36**, L00B08, doi:10.1029/2008GL036766.

Allen, R. M., Gasparini, P., and Kamigaichi O. (eds.), 2009b. Earthquake early warning. *Seismological Research Letters*, **80**, 682–782.

Allen, R. M., Gasparini, P., and Kamigaichi, O. (eds.), 2009c. New methods and applications of earthquake early warning. *Geophysical Research Letters*, **36**, L00B05, L00B08, L00B03, L00B04, L00B02, L00B06, L00B01, L00B07.

Allen, R. M., Gasparini, P., Kamigaichi, O., and BöSe, M., 2009d. The status of earthquake early warning around the world: an introductory overview. *Seismological Research Letters*, **80**, 682–693, doi:610.1785/gssrl.1780.1785.1682.

Bakun, W. H., Fischer, F. G., Jensen, E. G., and Vanschaack, J., 1994. Early warning system for aftershocks. *Bulletin Seismological Society of America*, **84**, 359–365.

Böse, M., Wenzel, F., and Erdik, M., 2008. PreSEIS: a neural network-based approach to earthquake early warning for finite faults. *Bulletin Seismological Society of America*, **98**, 366–382.

Böse, M., Hauksson, E., Solanki, K., Kanamori, H., and Heaton, T. H., 2009a. Real-time testing of the on-site warning algorithm in southern California and its performance during the July 29 2008 Mw5.4 Chino Hills earthquake. *Geophysical Research Letters*, **36**, L00B03, doi:10.1029/2008GL036366.

Böse, M., Hauksson, E., Solanki, K., Kanamori, H., Wu, Y.-M., and Heaton, T., 2009b. A new trigger criterion for improved real-time performance of on-site early warning in southern California. *Bulletin Seismological Society of America*, **99**, 897–905.

Brown, H., Allen, R. M., and Grasso, V. F., 2009. Testing ElarmS in Japan. *Seismological Research Letters*, **80**, 727–739, doi:710.1785/gssrl.1780.1785.1727.

Cua, G., and Heaton, T., 2007. The Virtual Seismologist (VS) method: a Bayesian approach to earthquake early warning. In Gasparini, P., Manfredi, G., and Zschau, J. (eds.), *Earthquake Early Warning Systems*. Berlin: Springer, pp. 97–132.

Cua, G., Fischer, M., Heaton, T., Wiemer, S., and Giardini, D., 2009. Real-time performance of the virtual seismologist method in southern California. *Seismological Research Letters*, **80**, 740–748, doi:710.1785/gssrl.1780.1785.1740.

Espinosa Aranda, J. M., Jimenez, A., Ibarrola, G., Alcantar, F., Aguilar, A., Inostroza, M., and Maldonado, S., 1995. Mexico city seismic alert system. *Seismological Research Letters*, **66**, 42–52.

Espinosa Aranda, J. M., Cuellar, A., Garcia, A., Ibarrola, G., Islas, R., and Maldonado, S., 2009. Evolution of the Mexican Seismic Alert System (SASMEX). *Seismological Research Letters*, **80**, 682–693, doi:610.1785/gssrl.1780.1785.1682.

Gasparini, P., Manfredi, G., and Zschau, J., 2007. *Earthquake Early warning systems*. Berlin: Springer, p. 350.

Giardini, D., 1999. The global seismic hazard assessment program (GSHAP) - 1992/1999. *Annali Di Geofisica*, **42**, 957–974.

Goltz, J. D., 2002. Introducing earthquake early warning in California: a summary of social science and public policy issues. Report of California Office of Emergency Services by the Caltech Seismological Laboratory, Disaster Assistance Division. Pasadena, California.

Horiuchi, S., Negishi, H., Abe, K., Kamimura, A., and Fujinawa, Y., 2005. An automatic processing system for broadcasting earthquake alarms. *Bulletin Seismological Society of America*, **95**, 708–718.

Ionescu, C., Böes, M., Wenzel, F., Marmureanu, A., Grigore, A., and Marmureanu, G., 2007. An early warning system for deep Vrancea (Romania) earthquakes. In Gasparini, P., Manfredi, G., and Zschau, J. (eds.), *Earthquake Early Warning Systems*. Berlin: Springer, pp. 343–349.

Kamigaichi, O., 2004. JMA earthquake early warning. *Journal of the Japan Association for Earthquake Engineering*, **4**, 134–137.

Kamigaichi, O., Saito, M., Doi, K., Matsumori, T., Tsukada, S., Takeda, K., Shimoyama, T., Nakamura, K., Kiyomoto, M., and Watanabe, Y., 2009. Earthquake early warning in Japan – warning the general public and future prospects. *Seismological Research Letters*, **80**, 717–726, doi:710.1785/gssrl.1780.1785.1717.

Kanamori, H., 2005. Real-time seismology and earthquake damage mitigation. *Annual Review of Earth and Planetary Sciences*, **33**, 195–214.

Köhler, N., Cua, G., Wenzel, F., and Böse, M., 2009. Rapid source parameter estimations of southern California earthquakes using PreSEIS. *Seismological Research Letters*, **80**, 748–754, doi:710.1785/gssrl.1780.1785.1748.

Nakamura, Y., 1988. On the urgent earthquake detection and alarm system (UrEDAS). In *Proceedings of the 9th World Conference on Earthquake Engineering*, Vol. 7, pp. 673–678.

Nakamura, Y., and Saita, J., 2007a. UrEDAS, the earthquake warning system: today and tomorrow. In Gasparini, P., Manfredi, G., and Zschau, J. (eds.), *Earthquake Early Warning Systems*. Berlin/Heidelberg: Springer, pp. 249–282.

Nakamura, Y., and Saita, J., 2007b. FREQL and AcCo for a quick response to earthquakes. In Manfredi, G., Zschau, J., and Gasparini, P. (eds.), *Earthquake Early Warning Systems*. Berlin/Heidelberg: Springer, pp. 307–324.

Odaka, T., Ashiya, K., Tsukada, S., Sato, S., Ohtake, K., and Nozaka, D., 2003. A new method of quickly estimating epicentral distance and magnitude from a single seismic record. *Bulletin Seismological Society of America*, **93**, 526–532.

Okada, Y., Kasahara, K., Hori, S., Obara, K., Sekiguchi, S., Fujiwara, H., and Yamamoto, A., 2004. Recent progress of seismic observation networks in Japan – Hi-net, F-net, K-NET, KiK-net. *Earth, Planets and Space*, **56**, xv–xxviii.

Quarantelli, H., 1956. *The Condition of Panic*. Chicago: University of Chicago.

Working Group on California Earthquake Probabilities, 2007. The Uniform California Earthquake Rupture Forecast, Version 2 (UCERF 2), US Geological Survey, p.104.

Wu, Y.-M., and Kanamori, H., 2005a. Experiment on an onsite early warning method for the Taiwan early warning system. *Bulletin Seismological Society of America*, **95**, 347–353.

Wu, Y.-M., and Kanamori, H., 2005b. Rapid assessment of damage potential of earthquakes in Taiwan from the beginning of P waves. *Bulletin Seismological Society of America*, **95**, 1181–1185.

Wu, Y.-M., and Kanamori, H., 2008a. Exploring the feasibility of on-site earthquake early warning using close-in records of the 2007 Noto Hanto earthquake. *Earth, Planets and Space*, **60**, 155–160.

Wu, Y.-M., and Kanamori, H., 2008b. Development of an earthquake early warning system using real-time strong motion signals. *Sensors*, **8**, 1–9.
Wu, Y.-M., and Teng, T.-L., 2002. A virtual subnetwork approach to earthquake early warning. *Bulletin Seismological Society of America*, **92**, 2008–2018.
Wu, Y.-M., Kanamori, H., Allen, R. M., and Hauksson, E., 2007. Experiment using the tau-c and Pd method for earthquake early warning in southern California. *Geophysical Journal International*, **170**, 711–717.
Wurman, G., Allen, R. M., and Lombard, P., 2007. Toward earthquake early warning in northern California. *Journal of Geophysical Research*, **112**, B08311, doi:08310.01029/02006JB004830.
Zollo, A., Lancieri, M., and Nielsen, S., 2006. Earthquake magnitude estimation from peak amplitudes of very early seismic signals on strong motion records. *Geophysical Research Letters*, **33**, L23312.
Zollo, A., Iannaccone, G., Lancieri, M., Cantore, L., Convertito, V., Emolo, A., Festa, G., Gallovic, F., Vassallo, M., Martino, C., Satriano, C., and Gasparini, P., 2009. Earthquake early warning system in southern Italy: methodologies and performance evaluation. *Geophysical Research Letters*, **36**, L00B07.

Cross-references

Earthquake Precursors and Prediction
Earthquakes, Intensity
Earthquakes, Shake Map
Earthquakes, Strong-Ground Motion
Seismic Hazard

EARTHQUAKES, ENERGY

Domenico Di Giacomo, Peter Bormann
Department 2: Physics of the Earth, GFZ German Research Center for Geosciences, Potsdam, Germany

Definition and calculation

During the earthquake process, the strain energy W available for rupture is divided into fracture energy E_G, thermal energy E_H, and radiated seismic energy E_S. The stress conditions around the fault and the rheological and elastic characteristics of the Earth materials being ruptured determine the partitioning of the overall energy budget involved in the earthquake source process. In particular, E_G and E_H are the energies dissipated mechanically and as frictional heat during faulting, respectively, and E_S is the energy fraction that goes into elastic seismic waves energy. The latter being of great importance for seismologists in evaluating an earthquake's shaking potential, in the following the relationship between the strain energy and radiated seismic energy is outlined.

After Knopoff (1958), the change in strain energy before and after an earthquake can be obtained from:

$$\Delta W = S\bar{D}\bar{\sigma}, \tag{1}$$

where S = fault area, $\bar{D}$ = average displacement over the fault, and $\bar{\sigma} = (\sigma_0 + \sigma_1)/2$ is the mean stress (with σ_0 being the initial stress and σ_1 the final one). This energy represents the total work associated to the earthquake rupture process. The small energy fraction converted to E_S is (e.g., Wyss and Molnar, 1972):

$$E_S = \frac{1}{2}\left(\sigma_0 + \sigma_1 - 2\sigma_f\right) \cdot S\bar{D}, \tag{2}$$

where σ_f is the average frictional stress during faulting. It is also useful to relate E_S and ΔW via the radiation efficiency $\eta_R = E_S/(E_S + E_G)$ (see, e.g., Wyss, 1970; Husseini, 1977):

$$E_S = \eta_R \Delta W. \tag{3}$$

The radiation efficiency η_R, spanning between 0 and 1, expresses the amount of work dissipated mechanically, therefore if an earthquake is effective in radiating E_S then η_R approaches 1 (that is to say about all the energy budget is converted into the seismic waves energy), whereas if η_R is very small or = 0 then the energy is dissipated and no E_S is radiated.

Assuming that the final stress σ_1 and the frictional stress σ_f are identical (that is to say the stress drop $\Delta\sigma = \sigma_0 - \sigma_1$ is complete, Orowan, 1960), it has been shown by Kanamori (1977) that a measure of ΔW can be obtained considering another important physical parameter of the seismic source, that is the seismic moment M_0:

$$\Delta W = \left(\frac{\Delta\sigma}{2\mu}\right) M_0, \tag{4}$$

with $M_0 = \mu S\bar{D}$ (where μ = rigidity of the medium in the source area). M_0 is a measure of the overall static displacement caused by the fault rupture. It is a well determined seismological parameter (normally within a factor of 2) since it must be calculated from long and/or very long periods (usually >100 s for large earthquakes) of the recorded seismograms (Dziewonski et al., 1981), which are less affected than the short periods by the small-scale Earth heterogeneities along the path from the source to the receivers.

By assuming that the stress drop is complete and nearly constant (between some 20 and 60 bars = $2–6\cdot10^7$ dyn/cm^2 = 2–6 MPa) for very large earthquakes and that $\mu = 3–6\cdot10^{11}$ dyn/cm^2 = $3–6\cdot10^4$ MPa in the source area under average crust-upper mantle conditions, Kanamori (1977) showed that the ratio

$$E_S/M_0 = 5 \times 10^{-5} = \text{constant}. \tag{5}$$

Therefore, under these conditions (adopted also for introducing the moment magnitude *Mw*, see *Earthquake, Magnitude*), one could easily have an indirect estimation of E_S from known M_0.

However, M_0 and E_S are two distinct physical parameters. Indeed, M_0 is a static parameter that is not sensible to the details of the rupture history and does not bring any direct information about the frequency content radiated by the seismic source, whereas E_S depends on the dynamic characteristics of the seismic source and, being related to velocity, it is more important than M_0 for evaluating the earthquake damage potential over frequencies of

engineering interest (between about 0.1 and 10 Hz). Therefore, direct and independent measurements of E_S are fundamentals in seismic source studies, also considering that many authors pointed out how the assumption of constant stress drop may not be fulfilled (see, e.g., Purcaru and Berckhemer, 1978). Indeed, E_S is sensible to the stress conditions during the fault rupture, as shown by the following formulation (e.g., Rudnicki and Freund, 1981):

$$E_S = \int_{-\infty}^{\infty} \int_S \left[-\sigma_{ij}\ \gamma_j\ \dot{u}_i\right] dS dt, \tag{6}$$

where $\dot{u}_i$ = particle velocity, σ_{ij} = stress change associated with the fault displacement u_i, γ_j = unit vector normal to the surface S and pointing outward from the source, and t = time. This definition holds in the far field, that is, the radius of the surface S is much larger than the dimension of the earthquake source. It can be shown that Equation 6 is equivalent to another one often used in seismology and proposed by Haskell (1964):

$$E_S = \int_{-\infty}^{\infty} \int_S \rho\left[\alpha\, \dot{u}_\alpha{}^2 + \beta\, \dot{u}_\beta{}^2\right] dS dt, \tag{7}$$

where ρ = density, α and β = P- and S-wave velocities in the medium, $\dot{u}_\alpha$ and $\dot{u}_\beta$ = far-field velocity records of P- and S-waves, respectively. Thus, it would be straightforward to compute E_S by analyzing the recorded waveforms. However, with respect to the determination of M_0, for computing E_S it must be considered not only the low-frequency part but especially the high-frequency content of the seismograms, which is more sensible to the fine heterogeneities of the Earth structure. Since the seismic recordings must be corrected for the energy loss due to the geometrical spreading and anelastic attenuation of body waves during their travel from the seismic source to the receivers (see *Propagation of Elastic Waves: Fundamentals*), it follows that the computation of E_S is more difficult and less stable than M_0. However, thanks to the worldwide deployment of high-quality digital broadband seismic station networks in the past few decades, much progress has been done and the calculation of E_S has started on a routine basis. Here a brief description of some techniques to calculate E_S is reported.

Since the correction for the high frequencies (above 1–2 Hz) is difficult to perform, many approaches to compute E_S use teleseismic recordings (usually between 30° and 90° epicentral distance Δ, although some procedure for rapid response purposes uses also recordings down to 20° or 25°, see Newmann and Okal, 1998; Di Giacomo et al., 2010a) for global earthquakes (magnitude above ~5) so that the effects of the Earth filter can be reasonably modeled by assuming an average 1-D layered structure. Moreover, since the P-waves are less affected by attenuation than S-waves, for which the correction for propagation effects at frequencies $f > 0.5$ Hz may be difficult on a routine basis, the analysis is often performed (1) by considering the P-wave train of the seismograms and (2) in the frequency domain by exploiting the Parseval's identity:

$$\int_{-\infty}^{\infty} |\hat{s}(\omega)|^2 d\omega = 2\pi \int_{-\infty}^{\infty} |s(t)|^2 dt, \tag{8}$$

where $\hat{s}(\omega)$ is the Fourier transform of the time signal $s(t)$ and $\omega = 2\pi f$ the angular frequency.

Currently E_S determinations at the USGS are obtained from the approach proposed by Boatwright and Choy (1986) and are available at the SOPAR database (http://neic.usgs.gov/neis/sopar/). This method calculates the energy flux ε^* from the velocity spectrum of P-waves $|\dot{u}(\omega)|$ as:

$$\varepsilon^* = \frac{\rho\alpha}{\pi} \int_0^{\infty} |\dot{u}(\omega)|^2 e^{\left(\omega t_\alpha^*(\omega)\right)} d\omega, \tag{9}$$

where the symbol * indicates that the energy flux is corrected for anelastic attenuation by the function $t_\alpha^*(\omega)$ (Choy and Cormier, 1986). Then, the computation of the average E_S is obtained from:

$$E_S = 4\pi \langle F \rangle^2 \frac{\sum R_i^2 \varepsilon_i^*}{\sum F_i^2}, \tag{10}$$

where $\langle F \rangle$ = is the average radiation pattern coefficient for P-waves over the focal sphere, ε_i^* the energy flux at a given station i, R_i the geometrical spreading, and F_i the specific radiation pattern for the station i (Aki and Richards, 1980).

Other authors calculate E_S from integration of the moment rate spectral function (an average theoretical curve is shown in Figure 8 of the entry *Earthquake Magnitude*). The moment rate spectrum $\left|\hat{\dot{M}}(f)\right|$ can be obtained from the teleseismic P-waves spectrum $|\hat{u}(f)|$ as (e.g., Houston and Kanamori, 1986):

$$\left|\hat{\dot{M}}(f)\right| = \frac{4\pi\rho\alpha^3 R_E e^{(\pi f t^*)} |\hat{u}(f)|}{g(\Delta) R(\theta,\varphi) C \left|\hat{I}(f)\right|}, \tag{11}$$

where, R_E is the Earth radius, g(Δ) the geometrical spreading factor, $R(\theta,\phi)$ the radiation pattern, C the free surface coefficient, and $\left|\hat{I}(f)\right|$ the instrument response. Recalling that E_S is proportional to velocity, and, therefore, to the derivative of the function obtained from Equation 11, E_S is obtained from integration over frequency f as (see, e.g., Venkataraman and Kanamori, 2004):

$$\begin{aligned} E_S &= \left[\frac{8\pi}{15\rho\alpha^5} + \frac{8\pi}{10\rho\beta^5}\right] \int_0^{\infty} f^2 \left|\hat{\dot{M}}(f)\right|^2 df \\ &= \left[\frac{2}{15\pi\rho\alpha^5} + \frac{1}{10\pi\rho\beta^5}\right] \int_0^{\infty} \left|\hat{\ddot{M}}(f)\right|^2 df, \end{aligned} \tag{12}$$

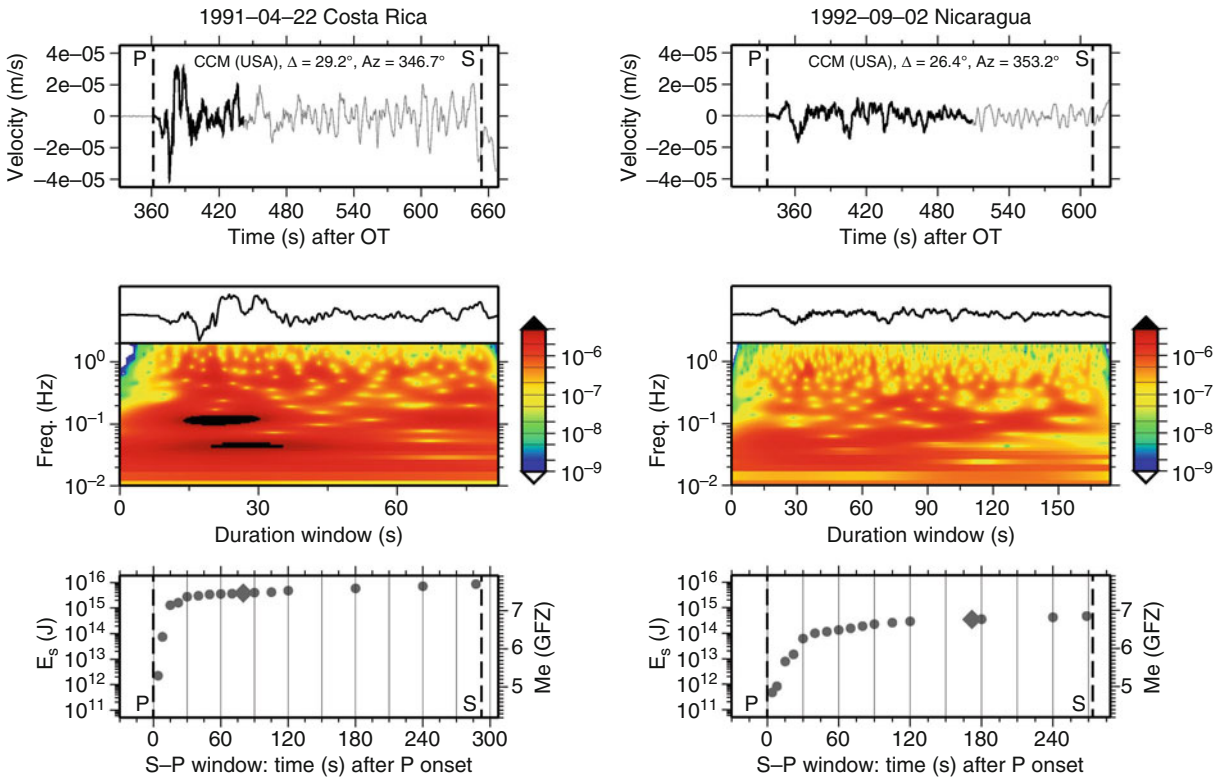
which holds for point sources, that is, sources smaller than the analyzed wavelengths (see *Earthquakes, Source Theory*) and neglecting directivity effects.

In sum, the calculation of E_S requires, after correction for propagation effects, to integrate the squared velocity spectrum recorded at seismic stations over a broad frequency range. Correction for the propagation effects is one of the most challenging tasks in seismological practice, since – in contrast to seismic moment determination (see *Earthquake, Magnitude*) – the high-frequency part of the source spectrum must also be considered. Correcting frequencies >1 Hz requires detailed knowledge of the Earth fine structure and attenuation properties, also below the seismic station, which is often not available. Consequently, E_S determinations from local and/or regional recordings are more problematic and are, anyway, more reliable only in areas with a high density of seismic stations (like California and Japan, see, e.g., the recent works by Izutani and Kanamori, 2001; Venkataraman et al., 2002; Boatwright et al., 2002; Oth et al., 2010) where a better knowledge of the local Earth structure is available.

Summary and outlook

The energy budget involved in the earthquake process is partitioned in different forms. In seismology it is of great importance to quantify two macroscopic parameters of the source, namely the seismic moment M_0 and the radiated seismic energy E_S. While the former (and the related magnitude *Mw*, see *Earthquake, Magnitude*) is a static measure of the seismic source and fundamental for the tsunami potential evaluation of an earthquake, the latter (and the related magnitude *Me*) represents the fraction of the strain energy converted in elastic seismic waves energy and fundamental for the evaluation of the shaking potential.

However, only since the early 1990s procedures to calculate E_S can be implemented on a routine basis at regional and global seismological centers. The approach of Boatwright and Choy (1986) allowed systematic and direct measurements of E_S from digital data, and later Choy and Boatwright (1995) demonstrated the importance of such measurements showing the significant variability of E_S for a given value of M_0. Therefore, the indirect



Earthquakes, Energy, Figure 1 *Upper panels*: *S-P* time windows (see *Body Waves*) of the vertical component velocity seismograms recorded at station CCM for the 1991 Costa Rica (*left*) and the 1992 Nicaragua earthquake (*right*). Vertical broken lines mark the theoretical arrival times of the *P* and *S* phases. *Middle panels*: time-frequency analysis via the *S-transform* (Stockwell et al., 1996) calculated over the *P*-wave duration window (black record traces in the upper panels that are also shown enlarged just above each middle panel); the color scales represent the amplitude of the *S-transform* in m/s. *Lower panels*: E_S and *Me* values as a function of time for different cumulative *P*-wave windows. The diamonds mark the end of the *P*-wave groups (see *black* record traces in the *upper panels*) that have been used for the single station E_S-*Me* estimates.

measurement of E_S, based on the assumptions leading to Equation 5, may be affected by large errors.

Of course, a better picture of the source characteristics can be obtained and help end users in the evaluation of the tsunami and shaking potential of an earthquake by complementing the information provided by direct measurements of E_S and M_0 and their related magnitudes. To this regard it is remarkable the case of two earthquakes that occurred about 500 km apart from each other in Central America, namely the 1991 Costa Rica earthquake and the 1992 Nicaragua earthquake. Both had the same moment magnitude $Mw = 7.6$ (i.e., about the same M_0). Because of their small difference in epicenter location, one can assume very similar propagation paths to teleseismic stations. However, the broadband waveforms of both earthquakes, recorded at identical stations, differ strongly. As an example, Figure 1 illustrates the *P*-wave windows, their time-frequency analysis performed with the *S-transform* (Stockwell et al., 1996), and the corresponding E_S-*Me* (see Di Giacomo et al., 2010a, b) values at station CCM for both earthquakes. The 1991 Costa Rica earthquake has much larger amplitudes, especially at higher frequencies, than the 1992 Nicaragua event. Such large differences can be only ascribed to significant differences in the medium-to-high-frequency parts of the source spectra of these two earthquakes with about the same M_0. Accordingly, E_S (and the corresponding *Me*) for the Costa Rica earthquake is much larger than for the Nicaragua earthquake. The latter caused a damaging *Tsunami* with 8 m run-up height, in contrast to severe damages caused by strong ground shaking of *intensity* I = X in the epicentral area of the Costa Rica earthquake (see *Earthquakes, Intensity*). This shows that M_0 alone (and a single type of magnitude as *Mw*) may not describe accurately enough the kind of seismic hazard and damage potential posed by a given earthquake and that a combined use of M_0 and E_S (and a multiple magnitude approach, see *Earthquake, Magnitude*) is preferable.

Thus, for improving seismic hazard assessment, future studies must focus on combining in a systematic way E_S and M_0 determinations in order to fully harness the information provided by these two important physical parameters of the seismic source.

Bibliography

Aki, K., and Richards, P., 1980. *Quantitative Seismology. Theory and Methods*. San Francisco: Freeman, p. 932.

Boatwright, J., and Choy, G. L., 1986. Teleseismic estimates of the radiated energy by shallow earthquakes. *Journal of Geophysical Research*, **91**(B2), 2095–2112.

Boatwright, J., Choy, G. L., and Seekins, L. C., 2002. Regional estimates of radiated seismic energy. *Bulletin of the Seismological Society of America*, **92**(4), 1241–1255.

Choy, G. L., and Cormier, V. F., 1986. Direct measurements of the mantle attenuation operator from broadband P and S waveforms. *Journal of Geophysical Research*, **91**(B7), 7326–7342.

Choy, G. L., and Boatwright, J., 1995. Global patterns of radiated seismic energy and apparent stress. *Journal of Geophysical Research*, **100**(B9), 18,205–18,228.

Di Giacomo, D., Parolai, S., Bormann, P., Grosser, H., Saul, J., Wang, R., and Zschau, J., 2010a. Suitability of rapid energy magnitude determinations for rapid response purposes. *Geophysical Journal International*, **180**, 361–374.

Di Giacomo, D., Parolai, S., Bormann, P., Grosser, H., Saul, J., Wang, R., and Zschau, J., 2010b. Erratum to "Suitability of rapid energy magnitude estimations for emergency response purposes". *Geophysical Journal International*, **181**, 1725–1726, doi:10.1111/j.1365-246X.2010.04610.x.

Dziewonski, A. M., Chou, T. A., and Woodhouse, J. H., 1981. Determination of earthquake source parameters from waveform data for studies of global and regional seismicity. *Journal of Geophysical Research*, **86**(B4), 2825–2852.

Haskell, N. A., 1964. Total energy and energy spectral density of elastic wave radiation from propagating faults. *Bulletin of the Seismological Society of America*, **54**(6), 1811–1841.

Houston, H., and Kanamori, H., 1986. Source spectra of great earthquakes: teleseismic constraints on rupture process and strong motion. *Bulletin of the Seismological Society of America*, **76**(1), 19–42.

Husseini, M. I., 1977. Energy balance for formation along a fault. *Geophysical Journal of the Royal Astronomical Society*, **49**, 699–714.

Izutani, Y., and Kanamori, H., 2001. Scale-dependence of seismic energy-to-moment ratio for strike-slip earthquakes in Japan. *Geophysical Research Letters*, **28**(20), 4007–4010.

Kanamori, H., 1977. The energy release in great earthquakes. *Journal of Geophysical Research*, **82**(20), 2981–2987.

Knopoff, L., 1958. Energy release in earthquakes. *Geophysical Journal*, **1**, 44–52.

Newman, A. V., and Okal, E. A., 1998. Teleseismic estimates of radiated seismic energy: The E/M_0 discriminant for tsunami earthquakes. *Journal of Geophysical Research*, **103**, 26,885–26,897.

Orowan, E., 1960. Mechanisms of seimic faulting in rock deformation: a symposium. *Geological Society of America Memoirs*, **79**, 323–345.

Oth, A., Bindi, D., Parolai, S., and Di Giacomo, D., 2010. Earthquake scaling characteristics and the scale-(in)dependence of seismic energy-to-moment ratio: insights from KiK-net data in Japan. *Geophysical Research Letters*, **37**, L19304, doi:10.1029/2020Gl044572.

Purcaru, G., and Berckhemer, H., 1978. A magnitude scale for very large earthquakes. *Tectonophysics*, **49**, 189–198.

Rudnicki, J. W., and Freund, L. B., 1981. On energy radiation from seismic sources. *Bulletin of the Seismological Society of America*, **71**(3), 583–595.

Stockwell, R. G., Mansinha, L., and Lowe, R. P., 1996. Localization of the complex spectrum: the S transform. *IEEE Transactions on Signal Processing*, **44**, 998–1001.

Venkataraman, A., Rivera, L., and Kanamori, H., 2002. Radiated energy from the 16 October 1999 Hector mine earthquake: regional and teleseismic estimates. *Bulletin of the Seismological Society of America*, **92**(4), 1256–1265.

Venkataraman, A., and Kanamori, H., 2004. Effect of directivity on estimates of radiated seismic energy. *Journal of Geophysical Research*, **109**, B04301, doi:10.1029/2003JB002548.

Wyss, M., 1970. Stress estimates for South American shallow and deep earthquakes. *Journal of Geophysical Research*, **74**(8), 1529–1544.

Wyss, M., and Molnar, P., 1972. Efficiency, stress drop, apparent stress, effective stress, and frictional stress of Denver, Colorado, earthquakes. *Journal of Geophysical Research*, **77**(8), 1433–1438.

Cross-references

Body Waves
Earthquake, Magnitude
Earthquakes, Intensity
Propagation of Elastic Waves: Fundamentals
Tsunami

EARTHQUAKES, INTENSITY

Gottfried Grünthal
Helmholtz Centre Potsdam, German Research Centre for Geosciences GFZ, Potsdam, Germany

Synonyms

Macroseismic intensity

Definition

The *intensity*, or *macroseismic intensity*, represents a classification of the severity of ground-motion shaking during an earthquake on the basis of observed effects at a given place. The word "macroseismic" refers to perceptible effects of earthquakes as opposed to instrumental observations. Intensity data are not only a surprisingly good measure of ground motion (Atkinson and Wald, 2007) but are also capable of constraining the magnitude of an event with the same order of uncertainty as individual instrumental magnitude readings (Johnston, 1996). Intensity data of an earthquake are usually presented in the form of a map showing for each intensity point either a corresponding color code or symbol. Contour lines in such maps separating intensities of different degrees are called isoseismals. The epicentral intensity I_0 is usually the largest observed intensity near to the epicenter or the barycenter of the highest intensity values. It can only be given for onshore events.

Macroseismic scales and intensity assignment

Earthquake intensities are defined in macroseismic scales. Since the early twentieth century, they usually contain 12 intensity degrees with the prominent exception of the 7° Japanese Meteorological Agency (JMA) scale (later upgraded to 10°). Earthquakes generate different intensities at different places, generally decreasing with distance from the epicenter. The epicentral intensity can be used in combination with the hypocentral depth as a classification of earthquake strength – with restrictions for large earthquakes ($M \geq 7$) with extensive fault planes.

Although not a physical parameter as such, intensity is a measure of strength of shaking, which can be related to physical ground-shaking parameters. It is a shorthand description of common observations from an earthquake, for example, instead of writing "the event was felt indoors by most, outdoors by few, and china and glasses clattered together", one summarizes "intensity 5".

Intensity is an integer quantity. Uncertain intensity assignments can be expressed as, for example, 8–9 (meaning either 8 or 9), being used in numerical procedures as 8.5, without claim of precision. A macroseismic scale defines each intensity degree with a series of diagnostics. Each diagnostic describes typical effects of ground vibrations related to different sensors. For the lower intensities, these main sensors are people and objects indoors, whereas for higher intensities they are buildings, which suffer progressively more severe damage with increased intensities. The type of building, for example, masonry or reinforced concrete, its condition and workmanship, as well as the level of earthquake-resistant design influence its strength considerably.

The most recent scale in general use, the European Macroseismic Scale (EMS-98), fully considers the varying strength of buildings in the form of six vulnerability classes, five damage grades for both masonry and reinforced concrete structures, as well as differentiating structural and nonstructural damage (Grünthal, 1998). Another element of the EMS-98 associated with the diagnostics is the relative frequency with quantitative definitions of the qualitative terms "few, many, most." It is the only intensity scale supplemented by comprehensive guidelines and background materials, and this contributes, together with the other mentioned elements, to a reduction of the subjectivity, which often is associated with intensity assignments.

The EMS-98 is easily adaptable for use to the building stock anywhere in the world. Table 1 gives the short form of the EMS-98 as a very simplified and generalized description. The long, full version should be used for intensity assignments.

The EMS-98 should be the basis for intensity evaluation in European countries and has also been applied in a number of countries outside Europe. Some South European countries, like Italy, still deploy Sieberg's (1932) Mercalli–Cancani–Sieberg scale (MCS). The basic intensity scale used in North America is the Modified Mercalli scale MMI. This scale has also been applied in countries elsewhere. In China, the China Seismic Intensity Scale (CSIS) has been used since 1999. This scale is strongly related to the test version of the EMS-98, that is, the EMS-92. While all these scales have 12 intensity degrees, the 1996 version of the JMA scale, introduced after the 1995 Great Hanshin earthquake, has been upgraded from originally 7° to 10°.

History of intensity scales

Historically, the use of intensities is important because it does not require instrumental measurements of earthquakes. In the last quarter of the nineteenth century, that is, before the establishment of instrumental seismology, the macroseismic intensity became a widespread parameter. The first internationally introduced scale was the 10° Rossi–Forel scale of 1883, which in some countries

Earthquakes, Intensity, Table 1 Short form of the EMS-98

EMS intensity	Definition	Description of typical observed effects (abstracted)
I	Not felt	Not felt.
II	Scarcely felt	Felt only by very few individuals at rest in houses.
III	Weak	Felt indoors by a few people. People at rest feel a swaying or light trembling.
IV	Largely observed	Felt indoors by many people, outdoors by very few. A few people are awakened. Windows, doors, and dishes rattle.
V	Strong	Felt indoors by most, outdoors by few. Many sleeping people awake. A few are frightened. Buildings tremble throughout. Hanging objects swing considerably. Small objects are shifted. Doors and windows swing open or shut.
VI	Slightly damaging	Many people are frightened and run outdoors. Some objects fall. Many houses suffer slight nonstructural damage like hairline cracks and fall of small pieces of plaster.
VII	Damaging	Most people are frightened and run outdoors. Furniture is shifted and objects fall from shelves in large numbers. Many well-built ordinary buildings suffer moderate damage: small cracks in walls, fall of plaster, parts of chimneys fall down; older buildings may show large cracks in walls and failure of fill-in walls.
VIII	Heavily damaging	Many people find it difficult to stand. Many houses have large cracks in walls. A few well-built ordinary buildings show serious failure of walls, while weak older structures may collapse.
IX	Destructive	General panic. Many weak constructions collapse. Even well-built ordinary buildings show very heavy damage: serious failure of walls and partial structural failure.
X	Very destructive	Many ordinary well-built buildings collapse.
XI	Devastating	Most ordinary well-built buildings collapse, even some with good earthquake-resistant design are destroyed.
XII	Completely devastating	Almost all buildings are destroyed.

persisted up to the 1960s. This points to the relevance of knowing the different scales that have been used in different countries during different time periods. Here, only a short summary can be given. For a comprehensive compilation, see Musson et al. (2010) and Musson and Cecić (2002).

A later popular scale is Mercalli's version from 1902. The first 12° scale is the one by Cancani (1904) with the focus on associating peak ground acceleration with single intensities. Based on the macroseismic observations from the 1911 $M_L = 6.1$ Central European earthquake, Sieberg (1912) provided a completely revised 12° version with descriptions of the intensity degrees given in great detail. This scale, which overwhelmingly influenced all modern intensity scales, later became known as the Mercalli–Cancani–Sieberg scale. Sieberg's (1923) minor revision of this scale was translated into English and published by Wood and Neumann (1931) in a largely condensed form, at the cost of clarity. The title is also inappropriate: the Modified Mercalli scale, MM-31. Sieberg's (1932) further revised version, with some major innovations, the MCS scale, is still in use in southern Europe. The MM-31 scale was completely overhauled in 1956 by Richter (1958). He refrained from associating his name with the new version to avoid confusion with the "Richter" magnitude scale M_L, and it was to be known as the Modified Mercalli scale of 1956 (MM-56).

Today's macroseismic evaluation procedures at the USGS rely basically on the MM-31 scale, but in a newly edited and almost undocumented version, that is, lacking differentiations in the terminology used, and in which extra rules and conventions are being applied (see Musson et al., 2010; Musson and Cecić, 2002). Intensity data sets are often given only as MM or MMI.

In Europe, Medvedev, Sponheuer and Kárník developed the MSK scale (Medvedev et al., 1964), introduced as MSK-64, on the basis of the MCS, the Russian GEOFIAN scale, and the MM-56. It is more systematically arranged than the forerunners and was widely adopted in Europe.

In 1988, the European Seismological Commission (ESC) launched the updating of the MSK scale to bring it in line with modern earthquake-resistant building types and engineering requirements. After the new scale had been constructed it was obvious that it contained so many new elements with respect to the MSK scale that it should bear a new name, and so EMS-92 and EMS-98 were established.

An aspect connected with the history of intensity scales is the decline of interest in macroseismology in the middle of the twentieth century when instrumental monitoring was progressing. However, the 1970s saw a revival of the subject since intensity assignments are fundamental for parametrizing historical seismicity and developing intensity-based ground-motion models essential in seismic hazard (see *Seismic Hazard*) and risk assessments.

Conversion of intensity scales

The conversion between intensity scales is, due to their integer nature, not possible in the way scalar parameters like different magnitude types can be converted. Ideally, conversions between intensities according to different scales should be avoided and the original data should be

Earthquakes, Intensity, Table 2 Conversion from the JMA-96 scale to the EMS-98 (After Musson et al. [2010])

JMA-96	EMS-98
0	1
1	2 or 3
2	4
3	4 or 5
4	5
5 Lower	6
5 Upper	7
6 Lower	8
6 Upper	9 or 10
7	11

reassessed using one scale. This is frequently not possible. Therefore, conversion diagrams have been developed in which the intensities according to different scales are represented as a series of "intensity ranges." One scale, taken as a standard with equally large ranges, is, as it is perceived by different authors, compared with other scales with different ranges for the intensities. Musson et al. (2010) elaborate in more detail on this issue and present tables suggesting direct numerical correspondence. Among the 12° scales, Musson et al. (2010) conclude equality between EMS-98 and MCS, MMI, and MSK for the intensities 1–10. For intensities 11 and 12, the definitions in MCS, MSK, and MMI relate to phenomena, such as surface faulting, which do not represent strength of shaking, or to a saturation of classification criteria in the scale where, for example, total damage refers to total damage to buildings without antiseismic design. Therefore, the intensities 11 and 12 do not have an equivalent in the EMS-98. The relation of the JMA-96 (JMA, 1996) scale with the EMS-98 is given in Table 2.

Macroseismic surveys

After an earthquake is felt, there are several ways to gather macroseismic data. The main source of data collection is questionnaires designed according to the diagnostics of the intensity-degree definitions. Questionnaires are especially used to compile data for the vast majority of observations in the range of intensities ≤ 6, although many questionnaires are designed for diagnostics up to intensity 7 or 8. Information on earthquake damages to buildings and other constructions need to be collected in the field. Special forms are in use in several high seismicity countries to optimize this procedure with respect to certain intensity scales and the building stock in a region.

Printed questionnaires used in the past have today been almost entirely replaced by internet web-based versions. The "Did You Feel It" (DYFI) program by the USGS (Wald et al., 1999a) is the basis of similar current activities worldwide. Musson (2006) describes the DYFI procedure for EMS-98 intensities in detail. The idea of such DYFI programs is that persons who have observed an earthquake use an internet web-site to report the macroseismic effects through answering a multiple-choice questionnaire. Such DYFI programs are often not restricted to a certain country but can be applied internationally. They have been remarkably successful since their inception. In the U.S. alone, more than ¾ million domestic responses have been compiled in the first 7 years since the start of the program in 2000 (Atkinson and Wald, 2007). This vast amount of macroseismic information proves to be very useful as an essential new source of ground-motion data (see *Earthquakes, Strong-Ground Motion*).

Intensity attenuation and focal depth

The rate of decay of shaking strength with distance from the epicenter, that is the intensity attenuation, is strongly dependent upon the focal depth h and to a lesser extent on the regionally dependent absorption of energy. Assuming a logarithmic proportionality between intensity and ground acceleration after Cancani (1904), an isotropic intensity-attenuation model results (Kövesligethy, 1906), rewritten by Jánosi (1907) as

$$I_0 - I_i = 3\log\left(\frac{r}{h}\right) + 3\alpha \log e(r-h) \tag{1}$$

with $r = \sqrt{R^2 + h^2}$, where R is the radius of the isoseismal intensity, I_i, I_0 the epicentral intensity, and e is Euler's constant. The same structure as Equation 1. results for a logarithmic proportionality between intensity and seismic energy density (Howell and Schultz, 1975; *Earthquakes, Energy*).

The first and, for a long time, the most important use of intensity attenuation was focused on depth determinations. For this purpose, graphical schemes by Sponheuer (1960) were frequently used in the pre-computer era. In the case of good macroseismic data, the calculated depths coincide very well with instrumental depth determinations, at least for shallow (5–25 km) events with small source dimensions (Musson, 1993). Intensity-attenuation relations of the type of Equation 1. were determined from a few mean isoseismal radii. Only within the last few years have modern statistical procedures been applied, which rely directly on a huge amount of intensity data points. One of these relations based on Central European data after Stromeyer and Grünthal (2009) reads

$$I_* - I = 2.95\log\left(\frac{r}{h}\right) + 0.0025(r-h) \tag{2}$$

where I is the intensity at epicentral distance r and I_* is any estimate of the intensity at $r = 0$, but not necessarily epicentral intensity I_0. Gomez Capera (2006) compiled 15 intensity-attenuation relations of this type.

Conversion of intensities to magnitudes

Intensity data give surprisingly robust measures of earthquake magnitudes (see *Earthquake, Magnitude*). This important fact qualifies the macroseismic method to be used for magnitude determinations of historical earthquakes occurred in early or pre-instrumental times. Seismic hazard analysis greatly benefits from long-term

earthquake catalogues. There are three ways of macroseismic magnitude determinations based on the following: (1) the epicentral intensity, (2) the areas affected by certain intensity levels, and (3) the entire field of intensity data points. These ways are described below.

1. A few years after the introduction of the earthquake magnitude, the first empirical relation with the epicentral intensity I_0 was given (Gutenberg and Richter, 1942). The maximum or epicentral intensity is in most cases the only information at hand for historical earthquakes. I_0 is strongly affected by the focal depth h. Therefore, h should be taken into consideration or the seismogenic depth in an area should be narrowly constrained at the conversion to magnitude. A modern relation based on well constrained moment magnitude M_w, I_0, and h of European earthquakes given by Grünthal et al. (2009) reads

$$M_w = 0.667I_0 + 0.30\log(h) - 0.10 \quad (3)$$

or omitting h

$$M_w = 0.682I_0 + 0.16 \quad (4)$$

These relations are valid for the ranges $5 < I_0 < 9.5$, $3.0 < M_w < 6.4$, and $5 \leq h[\text{km}] < 22$.

2. For earthquakes with known total felt area or known area enclosed by one or more of the n preferably outer isoseismals (An), the magnitude can be determined as a function of $\log An$. Johnston (1996) derived regression relations for MMI isoseismal areas for stable continental regions to predict seismic moments M_0 according to a functional regression form proposed by Frankel (1994)

$$\log M_0 = a + b\log An + c\sqrt{An} \quad (5)$$

From the published relations for different isoseismal areas, the coefficients for the areas of intensity 3 and 4 read

	a	b	c
A3	17.59	1.020	0.00139
A4	18.10	0.971	0.00194

The uncertainties in the derived moment magnitudes can be halved by applying weighted averaging of a suite of isoseismal areas. The uncertainties are then in the range of $\pm 0.15 \ldots 0.23M_w$ units.

3. The most advanced methods to derive magnitude from macroseismic data use the intensity data points of an event to make a joint determination of epicenter location (see *Earthquake, Location Techniques*) and magnitude. These methods need a well-calibrated intensity-attenuation model valid for a certain region. The most important methods of this type are by Bakun and Wentworth (1997), originally developed for Californian earthquakes $(4.4 < M_w < 6.9)$, and by Gasperini et al. (1999) for Italian earthquakes with $M_s \geq 5.5$. The latter additionally provides the fault azimuth, essentially derived from the direction of maximum elongation of the highest intensity data points. An adaption of the Bakun and Wentworth (1997) method to the intensity attenuation in France was made by Bakun and Scotti (2006).

Magnitudes derived from macroseismic data should be indicated as such, for example, as $M_w(I_0)$ or $M_0(An)$.

Intensity-attenuation models

Rapid post-earthquake information in the form of shake maps and earthquake early warnings, and other quantitative scientific studies like intensity-based seismic hazard and risk assessments, require well-established models of intensity attenuation with distance. The macroseismic data accumulated by the "Did You Feel It" (DYFI) program (Wald et al., 1999a) represent a vast new source of engineering seismology data. The DYFI intensity data "make up in quantity what they may lack in quality... They are also robust and of surprisingly high utility" (Atkinson and Wald, 2007). Their regression equation reads as:

$$MMI = C_1 + C_2(M_w - 6) + C_2(M_w - 6)^2 + C_4\log R + C_5R + C_6B + C_7M_w\log R \quad (6)$$

where C_1 through C_7 are determined constants, $R = \sqrt{D^2 + h^2}$ where D is rupture distance and h effective depth. $B = log(R/R_t)$ where R_t is the transition distance in the attenuation shape. The values in Table 3 are yielded for California and Central and Eastern U.S. (Atkinson and Wald, 2007).

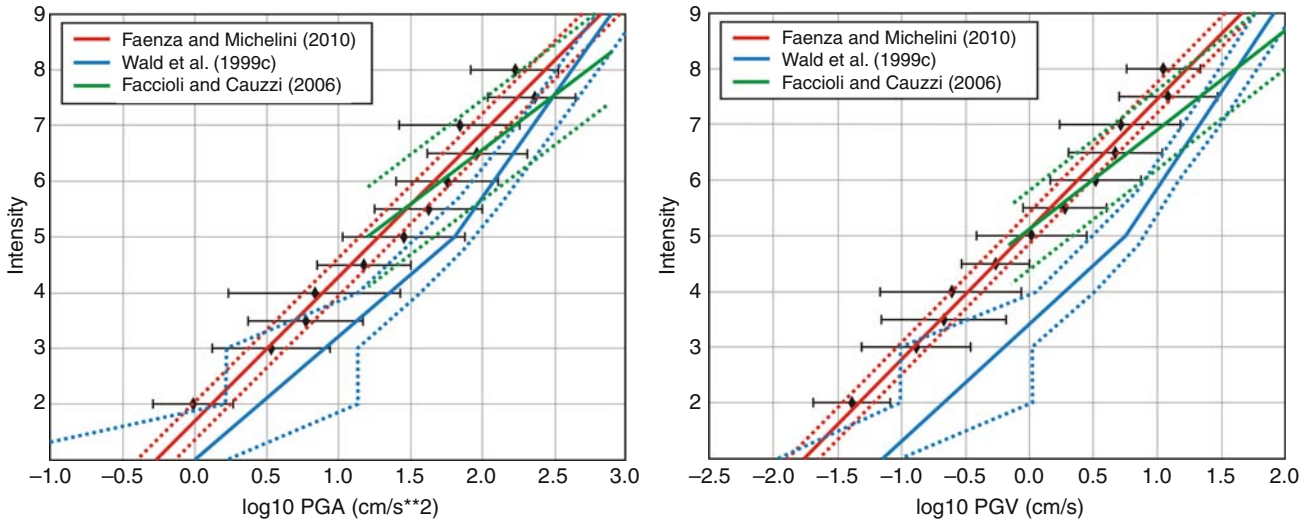
In recent years, intensity-attenuation models have especially been derived for regions where shake map programs (see *Earthquakes, Shake Map*) and early warning systems exist or are under development.

Relation of intensities to recorded strong ground motion

Attempts to express the strength of an earthquake and its observed effects in the form of physical parameters date

Earthquakes, Intensity, Table 3 Coefficients of equation 6

Coefficient	California	Central and Eastern U.S.
C_1	12.27(±0.24)	11.72(±0.36)
C_2	2.270	2.36
C_3	0.1304	0.1155
C_4	−1.30	−0.44
C_5	−0.0007070	−0.002044
C_6	1.95	2.31
C_7	−0.577	−0.479
h	14.0	17.0
R_t	30.0	80.0



Earthquakes, Intensity, Figure 1 Relations between intensity versus PGA (*left*) and PGV (*right*) after Wald et al. (1999c), Faccioli and Cauzzi (2006), and Faenza and Michelini (2010). Although MMI values and MCS values are used in these studies, the relations are given in a common scheme according to their de facto equivalence (Musson et al., 2010). (Redrawn after Faenza and Michelini [2010].)

back to the last quarter of the nineteenth century (cf. summary in Grünthal, 1984). Omori (1900) and Cancani (1904) were the first who related intensity degrees to peak ground acceleration (PGA) (see *Earthquakes, Strong-Ground Motion*). Kövesligethy (1906) and Jánosi (1907) invented a functional form that is still used today – a linear regression between intensity I and the logarithm of Cancani's PGA $(I = a + b\log(PGA))$. Grünthal (1984) compiled 30 published relations of intensity to peak ground motion, which showed a larger scatter in PGA compared to peak ground velocity (PGV).

For the implementation of the USGS Shake Map (see *Earthquakes, Shake Map*) software package (Wald et al., 1999b), representing shaking through maps of PGA, PGV, and ground-motion shaking intensity, it was necessary to derive "instrumental intensity" by converting peak ground motion to MMI intensity values (Wald et al., 1999c; Figure 1). In Europe, notably Italy, well-established relations for MCS intensities have recently been developed by Faenza and Michelini (2010) and references therein. Their orthogonal relations (Figure 1) read:

$$I = 1.68 + 2.58\log PGA\,[\mathrm{cm/s^2}] \tag{7}$$

$$I = 5.11 + 2.35\log PGV\,[\mathrm{cm/s}] \tag{8}$$

The densest network to generate shake maps in terms of instrumental intensity exists in Japan. Here, the instrumental intensity according to the JMA scale (I_{JMA}) is computed from Fourier acceleration spectra multiplied with different filter functions (Kuwata and Takada, 2002; Karim and Yamazaki, 2002). Empirical relationships between I_{JMA} and different ground-motion parameters are given by Karim and Yamazuki (2002). Their relation between I_{JMA} and PGA was derived as

$$I_{JMA} = 1.58 + 1.38\log SI + 0.59\log \mathrm{PGA} + 0.02M \tag{9}$$

where SI is the "spectrum intensity" ($\mathrm{cm/s^2}$) introduced by Housner (1961) as the integral over the velocity response spectrum in the period range from 0.1 to 2.5 s, and PGA is in $\mathrm{cm/s^2}$. However, correspondence between "instrumental intensity" and the wording of the macroseismic scale cannot be taken for granted.

Conclusions

Macroseismic intensity is a robust measure of perceptible and potentially damaging ground motion and is essential in engineering seismology. Modern developments of macroseismic scales, defining intensity degrees with persons and buildings as sensors, fully consider the range of vulnerability of buildings and differentiate structural and nonstructural damage patterns. Intensity provides the principal means by which one can parameterize historical or pre-instrumental earthquakes, that is, to determine their location, magnitude, and focal depth. Good quality macroseismic data enables a similar range of uncertainty in the derived focal parameters as standard instrumental seismology. The increased relevance of the parameter macroseismic intensity is strongly related to the enhanced needs in reliable seismic hazard and risk assessments. With the internet-based platforms Shake Map and "Did You Feel It?" the macroseismic intensity has become omnipresent.

Bibliography

Atkinson, G. M., and Wald, D. J., 2007. "Did You Feel It?" intensity data: a surprisingly good measure of earthquake ground motion. *Seismological Research Letters*, **78**(3), 362–368.

Bakun, W. H., and Scotti, O., 2006. Regional intensity attenuation models for France and the estimation of magnitude and location of historical earthquakes. *Geophysical Journal International*, **164**(3), 596–610.
Bakun, W. H., and Wentworth, C. M., 1997. Estimating earthquake location and magnitude from seismic intensity data. *Bulletin. Seismological Society of America*, **87**(6), 1502–1521.
Cancani, A., 1904. Sur l'emploi d'une double echelle sismique des intensites, empirique et absolute. *Beiträge zur Geophysik, Ergänzungsband (supplementary volume) 2, Verhandlungen der zweiten internationalen seismologischen Konferenz*, Annex A10, 2, 281–283.
Faccioli, E., and Cauzzi, C., 2006. Macroseismic intensities for seismic scenarios, estimated from instrumentally based correlations. In Abstract Book 1st ECEES, http://www.ecees.org/abstracts_book.pdf, p. 125.
Faenza, L., and Michelini, A., 2010. Regression analysis of MCS intensity and ground motion parameters in Italy and its application in ShakeMap. *Geophysical Journal International*, **180**(3), 1138–1152.
Frankel, A., 1994. Implications of felt area-magnitude relations for earthquake scaling and the average frequency of perceptible ground motion. *Bulletin. Seismological Society of America*, **84**(2), 462–465.
Gasperini, P., Bernardini, F., Valensise, G., and Boschi, E., 1999. Defining seismogenic sources from historical earthquake felt reports. *Bulletin. Seismological Society of America*, **89**(1), 94–110.
Gomez Copera, A. A., 2006. Seismic hazard map for the Italian territory using macroseismic data. *Earth Science Research Journal*, **10**(2), 67–90.
Grünthal, G., 1984. Seismische Gefährdung. In Hurtig, E., and Stiller, H. (eds.), *Erdbeben und Erdbebengefährdung*. Berlin: Akademie Verlag, pp. 169–238.
Grünthal, G., (ed.) 1998. European Macroseismic Scale 1998 (EMS-98). Cahiers du Centre Européen de Géodynamique et de Séismologie 15, Centre Européen de Géodynamique et de Séismologie, Luxembourg, 99 pp. (http://www.gfz-potsdam.de/EMS98).
Grünthal, G., Wahlström, R., and Stromeyer, D., 2009. Harmonization check of M_w within the central, northern, and northwestern European earthquake catalogue (CENEC). *Journal of Seismology*, **13**(4), 613–632.
Gutenberg, B., and Richter, C. F., 1954. *Seismicity of the Earth*. Princeton: Princeton University Press. 310 pp.
Housner, G. W., 1961. *Vibration of Structures Induced by Seismic Waves, Shock and Vibration Handbook*. New York: McGraw Hill, pp. 1–32.
Howell, B. F., Jr., and Schultz, T. R., 1975. Attenuation of modified Mercalli intensity with distance from the epicenter. *Bulletin. Seismological Society of America*, **65**(3), 651–655.
Jánosi, v. I., 1907. Bearbeitung der makroseismischen Erdbeben auf Grund der "Cancanischen Gleichung," Technical report, K. u. K. Reichsanstalt für Meteorologie und Erdmagnetismus, Budapest.
Japanese Meteorological Agency, 1996. Explanation table of JMA seismic intensity scale. http://www.jma.go.jp/jma/kishou/know/shindo/explane.html
Johnston, A. C., 1996. Seismic moment assessment of earthquakes in stable continental regions – II. Historical seismicity. *Geophysical Journal International*, **125**(3), 639–678.
Karim, K. R., and Yamazaki, F., 2002. Correlation of JMA instrumental seismic intensity with strong motion parameters. *Earthquake Engineering and Structural Dynamics*, **31**(5), 1191–1212.
Kövesligethy de, R., 1906. A makroszeizmikus rengések feldolgozása. *Math. és Természettudományi Értesítõ*, **24**, 349–368.
Kuwata, Y., and Takada, S., 2002. Instantaneous instrumental seismic intensity and evacuation. *Journal of Natural Disaster Science*, **24**(1), 35–42.
Medvedev, S., Sponheuer, W., and Kárník, V., 1964. Neue seismische Skala. Intensity scale of earthquakes, 7. Tagung der Europäischen Seismologischen Kommission vom 24.9. bis 30.9.1962. In: Jena, Veröff. Institut für Bodendynamik und Erdbebenforschung in Jena, vol 77. Deutsche Akademie der Wissenschaften zu Berlin, pp. 69–76.
Musson, R. W. M., 1993. Macroseismic magnitude and depth for British earthquakes. BGS Global Seismology Report no. WL/93/13.
Musson, R. M. W., 2006. Automatic assessment of EMS-98 intensities. British Geological Survey, Seismology and Geomagnetic Programme, Internal Report IR/06/048, 16 pp.
Musson, R. M. W., and Cecić, I., 2002. Macroseismology. In Lee, W. H. K., Kanamori, H., Jennings, P. C., and Kisslinger, C. (eds.), *International Handbook of Earthquake and Engineering Seismology*. San Diego: Academic, pp. 807–822.
Musson, R. W. M., Grünthal, G., and Stucchi, M., 2010. The comparison of macroseismic intensity scales. *Journal of Seismology*, **14**(2), 413–428.
Omori, F., 1900. Seismic experiments on the fracturing and overturning of columns. *Publication of the Earthquake Investigation Committee in Foreign Languages*, **4**, 69–141.
Richter, C. F., 1958. *Elementary Seismology. A Series of Books in Geology*. San Francisco: Freeman. 768 pp.
Sieberg, A., 1912. Über die makroseismische Bestimmung der Erdbebenstärke. *Gerlands Beitrage zur Geophysik*, **11**, 227–239.
Sieberg, A., 1923. *Geologische, Physikalische und Angewandte Erdbebenkunde*. Jena: G. Fischer.
Sieberg, A., 1932. Geologie der Erdbeben. *Handbuch der Geophysik*, **2**(4), 550–555.
Sponheuer, W., 1960. Methoden zur Herdtiefenbestimmung in der Makroseismik. Freiberger Forschungshefte C 88, Geophysik. Akademieverlag Berlin, 117 pp.
Stromeyer, D., and Grünthal, G., 2009. Attenuation relationship of macroseismic intensities in Central Europe. *Bulletin. Seismological Society of America*, **99**(2A), 554–565.
Wald, D. F., Quiroriano, V., Heaton, T. H., and Kanamori, H., 1999a. Relationships between Peak Ground Acceleration, Peak Ground Velocity, and Modified Mercalli Intensity in California. *Earthquake Spectra*, **15**(3), 557–564.
Wald, D. F., Quitoriano, V., Dengler, L. A., and Dewey, J. W., 1999b. Utilization of the internet for rapid community intensity maps. *Seismological Research Letters*, **70**(1), 680–697.
Wald, D. F., Quitoriano, V., Heaton, T. H., Kanamori, H., Scrivner, C. W., and Worden, C. B., 1999c. TriNet "shake maps": rapid generation of peak ground motion and intensity maps for earthquake in southern California. *Earthquake Spectra*, **15**(3), 537–556.
Wood, H. O., and Neumann, F., 1931. Modified Mercalli intensity scale of 1931. *Bulletin. Seismological Society of America*, **21**, 277–283.

Cross-references

Earthquake, Location Techniques
Earthquake, Magnitude
Earthquakes, Energy
Earthquakes, Shake Map
Earthquakes, Strong-Ground Motion
Seismic Hazard

EARTHQUAKES, PAGER

David J. Wald
U.S. Geological Survey, Denver Federal Center, Lakewood, CO, USA

Definition and introduction

Prompt Assessment of Global Earthquakes for Response (PAGER) is an automated system that produces content concerning the impact of significant earthquakes around the world within approximately 30 min of any magnitude 5.5 or larger event. PAGER rapidly assesses earthquake impacts by comparing populations exposed to estimates of shaking intensity, and models of economic and fatality losses based on past earthquakes in each country or region of the world. The main purpose of the PAGER system is to inform emergency responders, government and aid agencies, and the media to the scope of the potential disaster. Earthquake alerts – which were formerly sent based on event magnitude and location, or population exposure to shaking – are generated based on the estimated range of fatalities and/or economic losses.

The U.S. Geological Survey's National Earthquake Information Center (NEIC), located in Golden, Colorado, locates over 30,000 earthquakes a year. Tragically, about 25 of these cause significant damage, injuries, or fatalities. The U.S. Geological Survey (USGS) often detects earthquakes well before eyewitness reports are available. The USGS must then decide rapidly whether Federal and international agencies should be alerted to a potentially damaging event. In the past, the USGS primarily relied on the experience and intuition of its on-duty seismologists to estimate the impact of an event. To quantify and improve the accuracy of the assessment, the USGS has developed PAGER, an automated system to rapidly estimate the shaking distribution, the number of people and settlements exposed to severe shaking, and an estimate of the range of possible fatalities and economic losses. The estimated losses, in turn trigger the appropriate color-coded alert (green, yellow, orange, red), which determines suggested levels of response, whether no-response-needed, or local/regional, national, or international response, respectively.

In addition to direct alert notifications, PAGER provides important supplementary information, including comments describing the dominant types of vulnerable buildings in the region, exposure, and any fatality reports from a previous nearby earthquake, and a summary of regionally specific information concerning the potential for secondary hazards, such as earthquake-induced landslides, tsunami, and liquefaction. A map provides quick visual overviews of shaking levels and population densities, and a table summarizes the predicted shaking intensity at nearby cities and towns. This impact summary information is available on the USGS earthquake web site [http://earthquake.usgs.gov/pager/] and as a printable, one-page report with an accompanying description such as that shown on the following pages.

Fundamental to such a system, the USGS operates the robust computational and communication infrastructure necessary for earthquake response. PAGER results are generally available within 30 min of a significant earthquake, shortly after the determination of its location and magnitude. However, information on the extent of shaking will be uncertain in the minutes and hours following an earthquake and typically improves as additional sensor data and reported intensities are acquired and incorporated into models of the earthquake's source. Users of PAGER should account for the inherent uncertainty in shaking and loss estimations, and always seek the most current PAGER release on the USGS web site for any earthquake.

The PAGER process

In general, the shaking-related impact of an earthquake is controlled by the distribution and severity of shaking, the population exposed at each shaking intensity level, and how vulnerable that population is to building damage at each intensity level. Population vulnerability is dominated by the degree of seismic resistance of the local building stock. The PAGER system takes all these factors into account. At the heart of PAGER are the timely and accurate earthquake location and magnitude determinations that the USGS has been producing for decades. PAGER uses these earthquake parameters to calculate estimates of ground shaking by using the methodology and software developed for ShakeMap (http://earthquake.usgs.gov/shakemap/). The ShakeMap system produces maps of regional ground shaking using a site-specific ground-motion amplification map, seismic wave attenuation equations, and reported or recorded intensities and ground motions (Wald et al., 2005). The number of people exposed to each shaking intensity levels is then calculated by combining the maps of estimated ground shaking with a comprehensive worldwide population database (Landscan, from Oak Ridge National Lab).

Next, based on the population exposed to each intensity level of shaking, the PAGER system estimates total losses based on country-specific models developed from economic and casualty data collected from past earthquakes (Jaiswal et al., 2009). To calibrate the loss models, the USGS has generated an atlas of approximately 6,000 ShakeMaps for significant global earthquakes that have occurred during the last 40 years (Allen et al., 2009). The calibration of loss methodologies relies on this atlas and on fatality and damage data collected by the NEIC. Finally, the alert level, determined by estimated ranges of fatalities and economic loss are produced, with the higher of the two levels setting the overall alert level. The alert level determines which users are actively notified, and, at the same time, all PAGER information and content is automatically distributed to the Web on the USGS Earthquake web pages, as part of the earthquake summary information, for immediate consumption.

USGS science for a changing world

Summary Alert Red

GSN USAID FROM THE AMERICAN PEOPLE

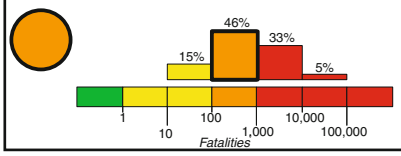
M 8.8, OFFSHORE MAULE, CHILE

Origin Time: Sat 2010-02-27 06:34:14 UTC (03:34:14 local)
Location: 35.91°S 72.73°W Depth: 35 km

PAGER Version 5

Created: 6 days, 10 hours after earthquake

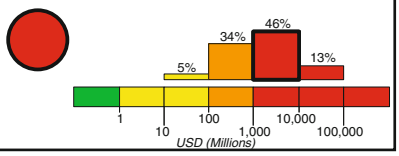
Estimated Fatalities



Red alert level for economic losses. Widespread damage is likely and the disaster is potentially widespread. Past events with this alert level have required a national or international level response.

Orange alert level for fatalities. Significant casualties are likely.

Estimated Economic Losses



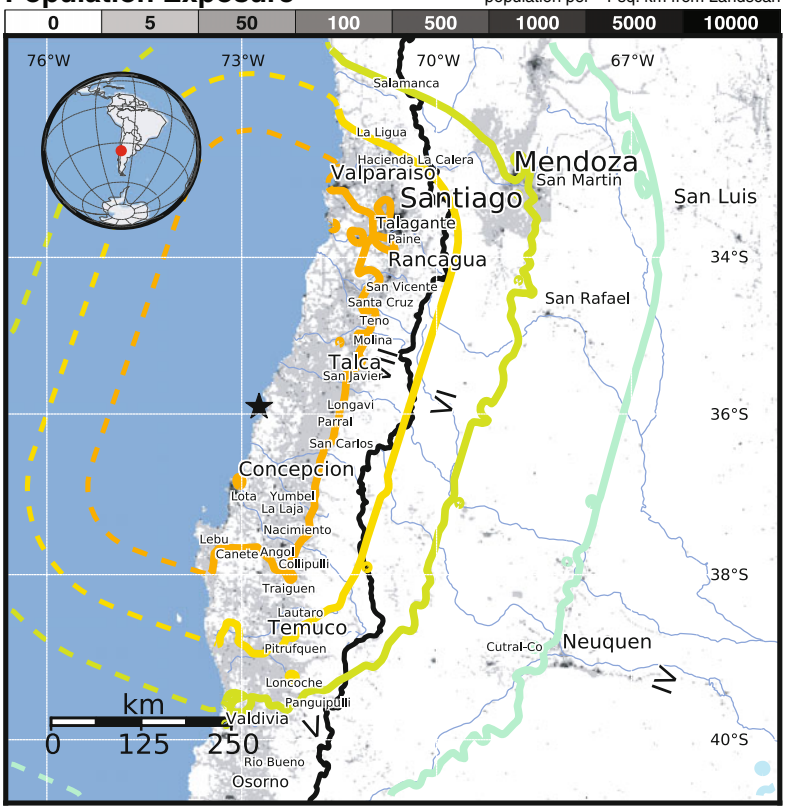
Estimated Population Exposed to Earthquake Shaking

ESTIMATED POPULATION EXPOSURE (k = x1000)		- -*	0*	1,101k*	2,795k*	967k*	9,169k	3,585k	0	0
ESTIMATED MODIFIED MERCALLI INTENSITY		I	II-III	IV	V	VI	VII	VIII	IX	X+
PERCEIVED SHAKING		Not felt	Weak	Light	Moderate	Strong	Very Strong	Severe	Violent	Extreme
POTENTIAL DAMAGE	Resistant Structures	none	none	none	V. Light	Light	Moderate	Moderate/Heavy	Heavy	V. Heavy
	Vulnerable Structures	none	none	none	Light	Moderate	Moderate/Heavy	Heavy	V. Heavy	V. Heavy

*Estimated exposure only includes population within the map area.

Population Exposure

population per ~1 sq. km from Landscan



Structures:
Overall, the population in this region resides in structures that are vulnerable to earthquake shaking, though some resistant structures exist. The predominant vulnerable building types are low-rise reinforced/confined masonry and adobe block construction.

Historical Earthquakes:
On May 22, 1960 (UTC), a magnitude 9.5 earthquake 273 km South of this one struck Valdivia, Chile, with estimated population exposures of 216,000 at intensity IX and 230,000 at intensity VIII, resulting in a reported 3,263 deaths from the earthquake and tsunami. Recent earthquakes in this area have caused tsunamis, landslides, and liquefaction that may have contributed to losses.

Selected City Exposure

from GeoNames.org

MMI	City	Population
VIII	Arauco	25k
VIII	Coronel	93k
VII	**Yumbel**	**11k**
VII	Curanilahue	31k
VII	Cauquenes	31k
VII	**Nacimiento**	**21k**
VII	**Talca**	**197k**
VII	**Valparaiso**	**282k**
VII	**Santiago**	**4,837k**
VII	**Rancagua**	**213k**
VII	**Concepcion**	**215k**

bold cities appear on map (k = x1000)

This information was automatically generated.

http://earthquake.usgs.gov/pager

Event ID: us2010tfan

Earthquakes, PAGER, Figure 1 Example of the PAGER summary product showing the color-coded impact scale indicated the alert levels (*top*), the total population exposed per color-coded intensity level (*center, near top*), a map of the population density and contoured intensity level (*lower left*), summary statements on country-specific structures and historical earthquakes, and selected cities with population and intensity level (*lower right*).

The PAGER earthquake impact scale

PAGER employs a new earthquake impact scale (EIS, Wald et al., 2010) that is based on two complementary criteria. One, based on the estimated cost of damage, is most suitable for domestic events; the other, based on estimated ranges of fatalities, is generally more appropriate for global events, particularly in developing countries. Simple thresholds, derived from the systematic analysis of past earthquake impact and associated response levels, turn out to be quite effective in communicating predicted impact and response needed after an event, characterized by alerts of green (little or no impact), yellow (regional impact and response), orange (national-scale impact and response), and red (international response). Corresponding fatality thresholds for yellow, orange, and red alert levels are 1, 100, and 1,000, respectively. For damage impact, yellow, orange, and red thresholds are triggered by estimated losses reaching $1 million, $100 million, and $1 billion, respectively. The rationale for a dual approach to earthquake alerting stems from the recognition that relatively high fatalities, injuries, and homelessness dominate in countries where local building practices typically lend themselves to high collapse and casualty rates, and it is these impacts that lend to prioritization for international response. In contrast, it is often financial and overall societal impacts that trigger the level of response in regions or countries where prevalent earthquake resistant construction practices greatly reduce building collapse and resulting fatalities.

Since PAGER calculations are available well in advance of ground-truth observations or news accounts, PAGER information can play a primary alerting role for domestic as well as international earthquake disasters. An example of the PAGER summary product, or *onePAGER*, is shown on the following pages for a recent destructive earthquake in early 2010 near central Chile that killed approximately 300 people. This earthquake reached an orange-alert level based on both projected fatalities and economic losses. The alert level is based on the median loss estimate; the uncertainty in the alert level can be gauged by the histogram showing the percent likelihood that adjacent alert levels (or loss/fatality ranges) occur. Accompanying text clarifies the nature of the alert based on experience from past earthquakes and provides context on the total economic losses in terms of the percent of the gross domestic product (GDP) of the country affected.

Though PAGER uses simple and intuitive color-coded alerting criteria, it preserves the necessary uncertainty measures by which one can gauge the likelihood for the alert to be over- or underestimated. Utilizing the earthquake impact scale, PAGER's rapid loss estimates can adequately recommend alert levels and suggest appropriate response protocols, despite their uncertainties; demanding or awaiting observations or loss estimates with a high level of accuracy may increase the losses. Over the past 38 years (1973–2010), there would have been approx. 17,792 green (not shown in Figure 1), 568 yellow, 52 orange, and 49 red alerts. This frequency equates to approximately 15 yellow, 1–2 orange, and 1–2 red alerts per year.

Ongoing PAGER developments

USGS is improving the PAGER system to include more comprehensive loss-estimate methodologies that take into account detailed building inventories that account for subcountry level regional variations, more complete population demographics (including time of day population shifts), and better tools to compute building damage (e.g., Jaiswal et al., 2010). Such data sets are very difficult and time consuming to acquire, and are effectively unavailable in many areas of the globe. In particular, more detailed within-country building inventories are key for describing the dominant vulnerable structures that contribute to casualties in the region. Knowledge of the main collapse "culprits" is vital for response, public safety, recovery, and long-term mitigation. Related USGS developments under the auspices of the PAGER Project include rapid determination of fault geometry, size, and rupture characteristics; refined global estimates of seismic soil-amplification patterns; ShakeMap enhancements; ground-motion and loss-uncertainty analyses; and earthquake-induced landslide and liquefaction probability mapping.

Bibliography

Allen, T. I., Wald, D. J., Earle, P. S., Marano, K. D., Hotovec, A. J., Lin, K., and Hearne, M. G., 2009. An Atlas of ShakeMaps and population exposure catalog for earthquake loss modeling. *Bulletin of Earthquake Engineering*, **7**, 701–718.

Jaiswal, K. S., Wald, D. J., and Hearne, M., 2009. Estimating casualties for large earthquakes worldwide using an empirical approach: U.*S Geological Survey Open-File Report*, of 2009–1136, 78pp.

Jaiswal, K. S., Wald, D. J., and D'Ayala, D., 2010. Developing Empirical Collapse Fragility Functions for Global Building Types, *Earthquake Spectra,* in press.

Wald, D. J., Worden, B. C., Lin, K., and Pankow, K., 2005. "ShakeMap manual: technical manual, user's guide, and software guide." *U.S. Geological Survey, Techniques and Methods 12-A1*, 132pp.

Wald, D. J., Jaiswal, K., Marano, K. D., and Bausch, D., 2010. An earthquake impact scale. *Natural Hazards Review*, in press.

Cross-references

Earthquakes, Intensity
Earthquakes, Shake Map

EARTHQUAKES, SHAKE MAP

David J. Wald
U.S. Geological Survey, Denver Federal Center, Lakewood, CO, USA

Definition and introduction

ShakeMap® is an open-source software program employed to automatically produce a suite of maps and

products that portray the geographical extent and severity of potentially damaging shaking following an earthquake. ShakeMap's primary purpose is to provide post-earthquake situational awareness for emergency management and response as well as damage loss estimation. The availability of ShakeMaps immediately after a significant earthquake is critical for the identification of areas likely to be most damaged. Principal users include first responders, utility companies, response and aid agencies, scientists and engineers, and the media. Maps are made publicly available via the Internet within several minutes of an earthquake's occurrence. ShakeMap is widely deployed in seismically active, well-instrumented portions of the USA, internationally in Italy, Iceland, Greece, Romania, Switzerland, among other countries, and the U.S. Geological Survey (USGS) global system provides approximate maps for all events of magnitude 5.5 and larger worldwide.

Making a ShakeMap entails deriving a detailed description of the shaking over a large region with a variable distribution of seismic stations; this requires intelligently guided interpolation of measured ground motions unless the recordings are extremely abundant (as in Japan, for instance). In the ShakeMap algorithms, empirically based ground-motion estimates, modified by simple geologically based, frequency- and amplitude-dependent site-correction factors provide useful first-order shaking constraints to guide interpolation in areas that are not well instrumented. Thus, essential elements of a ShakeMap system (Wald et al., 2005) include: (1) near-real time peak ground motion and/or intensity observations; (2) a ground-motion, site-amplification map for the region; (3) the ShakeMap algorithms, which combine and guide interpolation of ground-motion observations with built-in shaking estimates to produce the suite of shaking-hazard maps; and (4) mechanisms for automatically transferring ShakeMap products to users, including via web pages and through internet-based product feeds.

Historical context

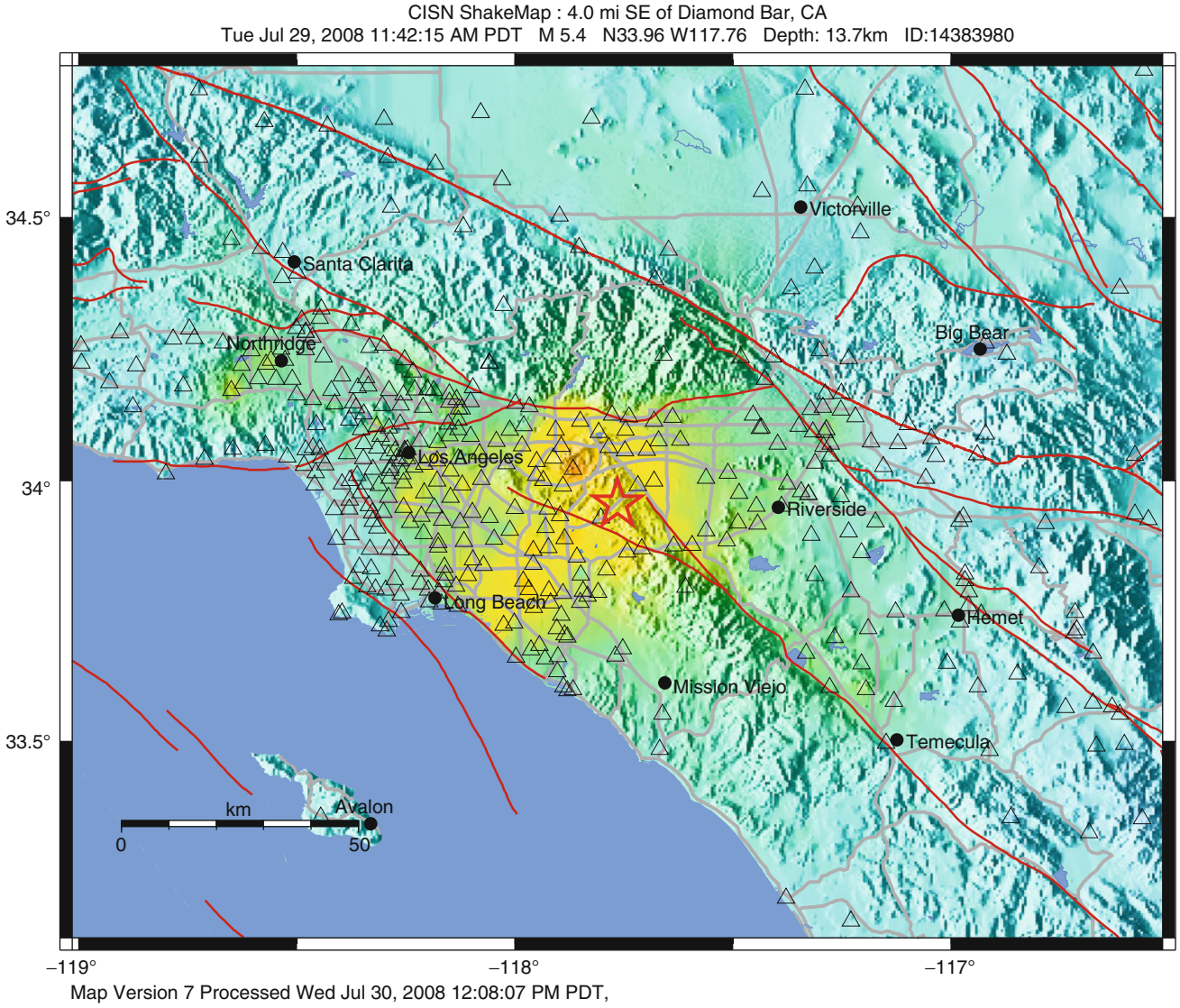
The most common information available immediately following a damaging earthquake has traditionally been its magnitude and the location of its epicenter. However, the damage pattern is not a simple function of these two parameters alone, and more detailed information is necessary to properly evaluate the situation. ShakeMap fills this role, and it has proven to be a useful, descriptive display for rapidly assessing the scope and extent of shaking and potential damage potential damage, following an earthquake. Because large earthquakes happen over a fault surface, not at a single point, the location of the earthquake (the epicenter) tells us only where the earthquake started, not necessarily where the shaking was the greatest. For a large earthquake, damage can occur hundreds of miles from the epicenter. Other factors, such as rupture direction, underground earth structure, and shallow geology, influence the amount of shaking in a particular area. For example, although emergency responders identified many areas of heavy damage soon after both the 1994 Northridge and the 1989 Loma Prieta, California, earthquakes, additional regions of severe damage were only belatedly discovered. A ShakeMap displays the distribution of ground shaking within minutes after an earthquake so that emergency managers can properly deploy resources to those locations with the greatest damage.

Part of the strategy for generating rapid-response ground-motion maps was to determine the best format for reliable presentation of the maps, given the diverse audiences. In an effort to simplify and maximize the flow of information to the public, ShakeMap was designed to produce not only peak-ground acceleration and velocity maps, but also an instrumentally derived, estimated Modified Mercalli Intensity map. The color-coded instrumental intensity map makes it easier to relate the recorded ground motions to the felt experience and damage distribution. ShakeMap was first developed for earthquakes in Southern California as part of the TriNet Project, a joint effort by the USGS, the California Institute of Technology, and the California Geological Survey (Wald et al., 1999b). Following the Northridge earthquake, older analog instruments were replaced with a modern seismic network employing near-real time digital communications. A by-product of that new network, ShakeMap was made possible by advances in telecommunications and computer-processing speed, as well as research that improved empirical ground-motion predictions and the relations among recorded ground motions and shaking intensities. A prototype ShakeMap system was first deployed in Southern California in 1997; refinements to the ShakeMap system continue today.

Related systems and uses

ShakeMap has facilitated the development and implementation of additional post-earthquake analyses and tools. Separate map layers are generated, each comprised of the spatial distribution of a specific peak ground-motion parameter (specifically acceleration, velocity, or response-spectral acceleration) as well as a layer for instrumentally derived seismic intensities. In addition to maps, ShakeMap produces an equally spaced grid containing the value of each parameter, as well as geographic information system (GIS) files and a host of other products for intended uses and users. Peak ground and spectral acceleration are fundamental parameters for earthquake-engineering and loss-estimation analyses; grids of shaking values can be combined with building inventory or lifeline databases to rapidly produce maps of estimated damage.

One such commonly used loss-modeling approach, both for post-earthquake response and pre-earthquake mitigation, is the Federal Emergency Management Agency's (FEMA) Hazards U.S. (HAZUS) software. FEMA routinely uses HAZUS to compute earthquake



Perceived shaking	Not felt	Weak	Light	Moderate	Strong	Very strong	Severe	Violent	Extreme
Potential damage	none	none	none	Very light	Light	Moderate	Moderate/Heavy	Heavy	Very Heavy
Peak acc.(%g)	*<.17*	*.17–1.4*	*1.4–3.9*	*3.9–9.2*	*9.2–18*	18–34	34–65	65–124	>124
Peak vel.(cm/s)	<0.1	0.1–1.1	1.1–3.4	3.4–8.1	*8.1–16*	*16–31*	*31–60*	*60–116*	*>116*
Instrumental intensity	I	II-III	IV	V	VI	VII	VIII	IX	X+

Earthquakes, Shake Map, Figure 1 ShakeMap for the magnitude 5.4 Diamond Bar, Los Angeles, California earthquake in July 2008 made by the California Integrated Seismic Network (CISN). Triangles indicate locations of seismic stations contributing to the map. Due to radiation pattern and site amplification, the strongest shaking is not at the epicenter; amplification due to the Los Angeles basin is particularly apparent at large distances southwest of the epicenter. Shaking variations in areas without stations are calculated from mapped soil variations and their amplification properties.

losses based on ShakeMap input within the USA. A secondary system for ShakeMap, ShakeCast (short for ShakeMap Broadcast), is a fully automated system for delivering specific ShakeMap products to critical users and for triggering established post-earthquake response protocols (Lin and Wald, 2008). ShakeCast allows utility and transportation agencies, businesses, and other large organizations to control and optimize the earthquake

information they receive. With ShakeCast, users can automatically determine the shaking value at their facilities, set thresholds for notification of different degrees of damage for each facility, and then automatically notify (via pager, cell phone, or email) specified operators, inspectors, or responders within their organizations who are responsible for those particular facilities so they can prioritize response. Having deployed ShakeCast, the California Department of Transportation (Caltrans) can prioritize inspection of thousands of bridges and overpasses statewide, an otherwise potentially overwhelming task.

Another secondary system, also developed primarily as a by-product of the USGS global implementation of ShakeMap, is called PAGER (Prompt Assessment of Global Earthquakes for Response). PAGER combines shaking intensity with exposed populations to rapidly and automatically estimate economic and human impacts worldwide. Unlike regional ShakeMap systems that rely on dense instrumentation, the global ShakeMap system relies primarily on the predictive aspects of the ShakeMap system; it incorporates site amplifications based on approximate soil conditions derived from topographic slope and utilizes macroseismic intensity data, if available, from near-real time, Internet-based data collection systems such as the USGS "Did You Feel It?" system (Wald et al., 1999a).

Outlook

Though developed primarily for post-earthquake, near-real time situational awareness, ShakeMap algorithms, which incorporate empirical ground-motion estimation capabilities, are widely used for other purposes. One side-benefit of the built-in ability to estimate ground motions for a specified earthquake is for generating hypothetical earthquake ground-motion maps. Such "Scenario ShakeMaps" are highly useful for response planning and long-term earthquake mitigation since portraying the overall pattern of shaking for potential future events is highly useful for earthquake drills and loss assessments. Calibration of earthquake loss models is facilitated by running ShakeMaps for historical earthquakes, which has been done for significant earthquakes around the globe from 1970 to 2008 (the ShakeMap Atlas; Allen et al., 2009). Some of these historic ShakeMap events have constraints from ground motion, intensity, and fault extent; others are purely predictive. Nonetheless, such historic ShakeMaps can be used for loss calibration or for better scientific understanding of the nature of past earthquakes. Ongoing refinements and the addition of earlier past or more recent events will continue. In addition, refinements to ShakeMap algorithms, products, formats, and associated tools are ongoing, and ShakeMap systems are being deployed and operated in numerous countries. At the time of this article, systems operate or are in development in countries including the USA, Italy, Iceland, Switzerland, Greece, Romania, China, and Indonesia.

Bibliography

Allen, T. I., Wald, D. J., Earle, P. S., Marano, K. D., Hotovec, A. J., Lin, K., and Hearne, M. G., 2009. An Atlas of ShakeMaps and population exposure catalog for earthquake loss modeling. *Bulletin of Earthquake Engineering*, **7**, 701–718.

Lin, K., and Wald, D. J., 2008. *ShakeCast Manual*. U.S. Geological Survey Open File Report 2008–1158, 90 pp.

Wald, D. J., Quitoriano, V., Dengler, L., and Dewey, J. W., 1999a. Utilization of the internet for rapid community intensity maps. *Seismological Research Letters*, **70**(6), 680–697.

Wald, D. J., Quitoriano, V., Heaton, T. H., Kanamori, H., Scrivner, C. W., and Worden, C. B., 1999b. TriNet "ShakeMaps": rapid generation of peak ground motion and intensity maps for earthquakes in southern California. *Earthquake Spectra*, **15**(3), 537–556.

Wald, D. J., Worden, B. C., Lin, K., and Pankow, K., 2005. "ShakeMap manual: technical manual, user's guide, and software guide." *U. S. Geological Survey, Techniques and Methods 12-A1*, 132 pp.

Cross-references

Earthquakes, Intensity
Earthquakes, PAGER
Earthquakes, Strong-Ground Motion

EARTHQUAKES, SOURCE THEORY

Raul Madariaga
Laboratoire de Géologie, Ecole Normale Supérieure, Paris, France

Definition

Seismic moment. A measure of the size of an earthquake. In the simplest situation it represents the moment of the couples of forces that make up a dipolar source. In more general cases it is a 3 by 3 symmetric tensor of elementary force couples. For planar faults it is the product of rigidity by slip times the source area.

Seismic radiation. The seismic waves emitted by an earthquake. For point sources these are spherical P and S waves emitted by the point source.

Seismic spectrum. Fourier transform of the seismic waves in the far field from the source. In general, when it is not noted otherwise, it refers to the amplitude of the Fourier transform of the far-field displacement.

Kinematic earthquake model. A model of slip history of rupture on a fault determined from near and far-field observations. In kinematic models, the propagation of rupture is arbitrarily specified.

Dynamic earthquake model. A model of rupture propagation on an earthquake fault based on fracture mechanics.

Introduction

Earthquake source theory studies the dynamics of seismic sources in the Earth. The simplest possible model of a seismic source is that of a point source buried in an elastic medium. A proper understanding of the exact nature of seismic source took more than 50 years since the first

efforts by Nakano and colleagues in the 1920s in Japan. Earthquakes were initially modeled as simple explosions, then as point forces and, finally, as the result of fast transformational strains inside a sphere. In the early 1950s, it was recognized that P waves radiated by earthquakes presented a spatial distribution similar to that produced by single couples of forces, but it was very soon recognized that this type of source could not explain S wave radiation. The next level of complexity was to introduce a double couple source, a source without resultant force or moment. The physical origin of the double couple model was established in the early 1960s. Soon after that, finite source models derived from dislocation theory were introduced as a simple way to model observations of seismic radiation. Dislocation models present some severe mechanical problems that we will discuss later. They were soon replaced by more mechanically sound models based on dynamic shear cracks. These are the two kinds of models that are currently being used to model earthquakes over a broad range of frequencies and length scales.

Point source models

The radiation from a point force or Green function is the fundamental solution of the equation of elastodynamics. With few rare exceptions, seismic sources are due to fast internal deformation in the earth, for instance, faulting or fast phase changes on localized volumes. For a seismic source to be of internal origin, it has to have zero net force and zero net moment. It is not difficult to imagine seismic sources that satisfy these two conditions. The simplest such source is an explosion or implosion, but they do not produce shear waves, at least in uniform media. The next group of possible sources is dipoles, quadrupoles, or combinations thereof. Experimental observation has shown that simple dipoles are not good models of seismic sources. It is possible, however, to combine three orthogonal linear dipoles in order to form a general seismic source. These three dipoles represent the principal directions of a symmetric tensor of rank 2 that we call the seismic moment tensor:

$$\mathbf{M} = \begin{bmatrix} \mathrm{Mxx} & \mathrm{Mxy} & \mathrm{Mxz} \\ \mathrm{Mxy} & \mathrm{Myy} & \mathrm{Myz} \\ \mathrm{Mxz} & \mathrm{Myz} & \mathrm{Mzz} \end{bmatrix} \tag{1}$$

This moment tensor has a structure that is identical to that of a stress tensor, but it is not of elastic origin. The off-diagonal elements of the moment represent the torque produced by two point forces of direction i separated by an infinitesimal distance in direction j. The diagonal terms represent linear dipoles, two opposite forces separated by a very small distance.

Radiation produced by a point moment tensor source located at point $\mathbf{r}_0$:

$$\mathbf{M}(\mathbf{r}, t) = \mathbf{M}_0(t)\delta(\mathbf{r} - \mathbf{r}_0) \tag{2}$$

form a set of Green functions. In a homogeneous, linear elastic medium, these Green functions are well known and contain both near-field and far-field terms. The most important are the far-field terms, the radiation observed at large distance from the source. These are:

$$u_i^{\mathrm{P}}(\mathbf{r}, t) = \frac{1}{4\pi\rho\alpha^3}\frac{1}{R}\sum_{jk} R_{ijk}^{\mathrm{P}}\dot{M}_{jk}(t - R/\alpha) \tag{3a}$$

for P waves, and

$$u_i^{\mathrm{S}}(\mathbf{r}, t) = \frac{1}{4\pi\rho\beta^3}\frac{1}{R}\sum_{jk} R_{ijk}^{\mathrm{S}}\dot{M}_{jk}(t - R/\beta) \tag{3b}$$

for S waves. Here R is the distance from the source to the observer. R^{P} and R^{S} are the radiation patterns of P and S waves, respectively; $\rho, \alpha,$ and β denote the density and the P and S wave speeds of the elastic medium. The radiation patterns are different for every element of the moment tensor with its own set of symmetries and nodal planes. We observe that the far-field signal carried by both P and S waves is the time derivative of the seismic moment components, so that far-field seismic waves are proportional to the moment rate of the source.

Very often in seismology, it is assumed that the geometry of the source can be separated from its time variation, so that the moment tensor can be written in the simpler form:

$$\mathbf{M}_0(t) = \mathbf{M}_0\Omega(t) \tag{4}$$

where, $\mathbf{M}_0$ is a time-invariant tensor that describes the geometry of the source and $\Omega(t)$is the time variation of the moment, the source time function determined by seismologists, often called also the scalar seismic moment rate. In this case, we can now write a simpler form of Equation 3:

$$u_i^c(\mathbf{r}, t) = \frac{1}{4\pi\rho c^3}\frac{1}{R}R^{\mathrm{c}}(\theta, \phi)\dot{\Omega}(t - R/c) \tag{5}$$

where, c stands for either the P or the S wave velocity. For P waves, u^c is the radial component; for S waves, it is the appropriate transverse component for SH or SV waves. We observe that the signal emitted by the source in the far field is the derivative of the scalar moment time function $\Omega(t)$. The term $R^{\mathrm{c}}(\theta, \phi)$is the radiation pattern, a function of the takeoff angle of the ray at the source.

With small modifications designed to take into account smooth variations of elastic wave speeds in the earth, these expressions are widely used to generate synthetic seismograms in the far field. The main changes that are needed are the use of travel time instead of R/c in the waveform (Equation 5), and a more accurate geometrical spreading to replace the $1/R$ term. The complete expressions are given by Aki and Richards (2002). In most work with local earthquakes, the approximation (Equation 5) is frequently used with a simple correction for free surface response.

What does a seismic moment represent? A number of mechanical interpretations are possible. In the previous sections, we introduced it as a simple mechanical model of double couples and linear dipoles. Other authors (Backus and Mulcahy, 1976) have explained them in terms of the

distribution of inelastic stresses (sometimes called stress "glut"). Let us first notice that a very general distribution of force of internal origin must necessarily derive from a symmetrical seismic moment density of the form

$$\mathbf{f}(r,t) = \nabla \bullet \mathbf{M}_0(\mathbf{r},t) \tag{6}$$

where, $\mathbf{M}_0$ is the moment tensor density per unit volume. It is not difficult to prove that such a force distribution has neither net force nor torque. The moment tensor represents a stress tensor due to internal stresses as proposed by Eshelby (1957). From the moment tensor one can define the associated inelastic strain by

$$\mathbf{M}_0 = \lambda\theta_I\mathbf{I} + 2\mu\boldsymbol{\varepsilon}_I \tag{7}$$

where, $\mathbf{I}$ is the identity matrix in 3D, λ and μ are the elastic constants. $\boldsymbol{\varepsilon}_I$ represents the inelastic strain at the source. This is the strain that would occur in the body in the absence of elastic forces. We can now define ε_I for a fault segment. Let the source be reduced to a 2D fault element of area dS and thickness h. Then the inelastic strain that occurs across the fault element dS is approximately ½ D/h, the ratio of the displacement discontinuity or slip, D across the fault by the thickness of the fault. Thus $\varepsilon_I h \to D/2$ as h decreases to zero fault width. For a planar moment distribution we get a 2D seismic moment distribution

$$dm_0 = \mu\, D\, dS \tag{8}$$

That is, the moment density of a shear fault is just the product of the elastic constant by slip, a classical result obtained by Burridge and Knopoff (1964) using a different argument based on the representation theorem.

Seismic radiation in the spectral domain. In seismically active areas the waves radiated by earthquakes become quite complex because of multipathing, scattering, attenuation, site response, etc., so that the actually observed seismogram resembles the source time function $\Omega(t)$only at long periods. Complexities in wave propagation affect much less the spectral amplitudes in the Fourier transformed domain. Radiation from a simple point moment tensor source can be obtained from Equation 5 by straightforward Fourier transformation:

$$u_i^c(\mathbf{r},\omega) = \frac{1}{4\pi\rho c^3}\frac{1}{R}R^{\mathrm{c}}(\theta,\phi)\,\tilde{\Omega}(\omega)\exp(-i\omega R/c) \tag{9}$$

where, $\tilde{\Omega}(\omega)$ is the Fourier transform of the source time function, $\Omega(t)$.

From the scaling of moment with earthquake size, Aki (1967), and from spectral observations, Brune (1970) concluded that the seismic spectra decayed as ω^{-2} at high frequencies. A simple model of the spectral amplitude is:

$$\Omega(\omega) = \frac{M_0}{1+\omega^2/\omega_0^2} \tag{10}$$

where, ω_0 is the corner frequency of the earthquake, the intersection between the low and high frequency asymptotes of expression (11). This is the so-called omega-squared model of seismic radiation. In this model, seismic sources are characterized by only two independent scalar parameters: the seismic moment and the corner frequency. As mentioned earlier, not all earthquakes have displacement spectra as simple as this, but the omega-squared model is a simple starting point for understanding seismic radiation.

Finite source models

Earthquakes are due to the fast propagation of slip on a fault, or fault system, at speeds of the order of that of shear waves. There is evidence that a few events have propagated at speeds faster than that of shear waves, but these are exceptional events. For most others, the speed of rupture propagation is close to that of Rayleigh waves on the fault. Traditionally, finite source earthquake models are either kinematic or dynamic. In the first class the distribution of seismic moment and rupture time are specified in order to satisfy certain observational constraints on the source. The most common way is to specify the rupture process based on simple models of rupture propagation, for instance, at constant speed and then determining the distribution of moment (slip) on the fault from the waveforms observed in different seismic stations. As the data has improved kinematic inversion methods have become increasingly sophisticated, but the basic problem of how does a rupture grow cannot be determined from basic kinematic principles. The alternative is to use dynamic models where rupture is computed numerically from the specification of the stresses that prevail on the fault and the friction law that acts on the fault surface.

Kinematic models

One of the most widely used kinematic models was introduced by Haskell (1964). In this model a uniform displacement discontinuity spreads at constant rupture velocity inside a rectangular fault. At low frequencies, for wavelengths longer than the size of the fault, this model is a reasonable approximation to the rupture of a strike slip fault. In Haskell's model at time $t = 0$, a line dislocation of width W appears suddenly and propagates along the fault length at a constant rupture velocity v until a region of length L of the fault has been broken. As the dislocation moves it leaves behind a zone of constant slip D. Assuming that the fault lies on a plane of coordinates (x, y), the slip function can be written as

$$D(x,y,t) = D\,s(t-x/v) \quad \text{for} \quad -W < y < W \text{ and } 0 < x < L \tag{11}$$

where, $s(t)$ is the source time function that, in the simplest version of Haskell's model, is invariant with position on the fault. The most important feature of this model is the

propagation of rupture implicit in the time delay of rupture x/v, where v is the rupture velocity. An obvious unphysical feature of this model is that rupture appears instantaneously in the y-direction; the other is that near the borders of the fault slip suddenly jumps from D to zero. This violates material continuity and produces very strong unphysical stress concentrations. In spite of these two shortcomings, Haskell's model gives a simple, first-order approximation to seismic slip, fault finiteness, and finite rupture speed. The seismic moment of Haskell's model is just $M_0 = \mu DLW$. This model has been extensively used to invert for seismic source parameters both in the near and far field from seismic and geodetic data. Some of the limitations of Haskell's model can be improved by admitting that slip is variable on the fault, so that D in Equation 11 is no longer assumed to be constant but a function of position on the fault. Slip can be tapered as suggested by dynamic models of rupture. Rupture speed can also be made variable.

Dynamical fault models

The most frequently used source models are dynamic models, based on some basic concepts of fracture mechanics introduced by seismologists in the middle 1960s. The simplest of these dynamic models is a circular shear crack. This model was introduced by several authors in order to relate fault slip to stress drop inside a fault. A natural approach to model earthquakes was to use well-known results from fracture mechanics. Among dynamic models we consider here a circular shear crack, assuming that rupture starts from a point and then propagates self-similarly until it finally stops at a certain source radius. This model was carefully studied in the 1970s and a complete understanding of it is available without the need of getting into the details of the numerical modeling of earthquake sources.

Kostrov's self-similar circular crack. The simplest model of how an earthquake starts is a circular-shaped rupture front that starts from a point and spreads at constant rupture speed v without ever stopping. Slip on this fault is driven by stress drop inside the fault. The solution for slip across the fault, found by Kostrov (1964), is surprisingly simple:

$$D(r,t) = C(v)\frac{\Delta\sigma}{\mu}\sqrt{v^2t^2 - r^2} \tag{12}$$

where, r is the radius of the fault measured from the point of rupture initiation, vt is the instantaneous radius of the rupture at time t, $\Delta\sigma$ is the constant stress drop inside the rupture zone, μ is the elastic shear modulus, and $C(v)$ is a very slowly varying function of the rupture velocity. For most practical purposes $C \sim 1$. Inside the fault, for $r < vt$, the stress change produced by the slip function is constant and equal to $\Delta\sigma$. Thus for constant stress drop, slip inside the fault scales with the "ratio of stress drop over rigidity 'times' the instantaneous radius of the fault." As rupture propagates, displacement around the fault scales with the instantaneous radius of the rupture zone. The dynamic model (Equation 12) describes rupture propagation, but it has no mechanism to stop the rupture.

A finite fault model can be developed from Equation 12 by assuming that rupture suddenly stops after the fault has grown to a certain fixed radius. This simple dynamic model was proposed by Madariaga (1976). The circular crack problem is posed in terms of stresses not of slip, but the rupture process is fixed in advance so that rupture does not develop spontaneously. This is the only unrealistic feature of this model and it is the reason it was qualified as quasidynamic, that is, rupture is kinematically defined, but slip is computed solving the elastodynamic equations.

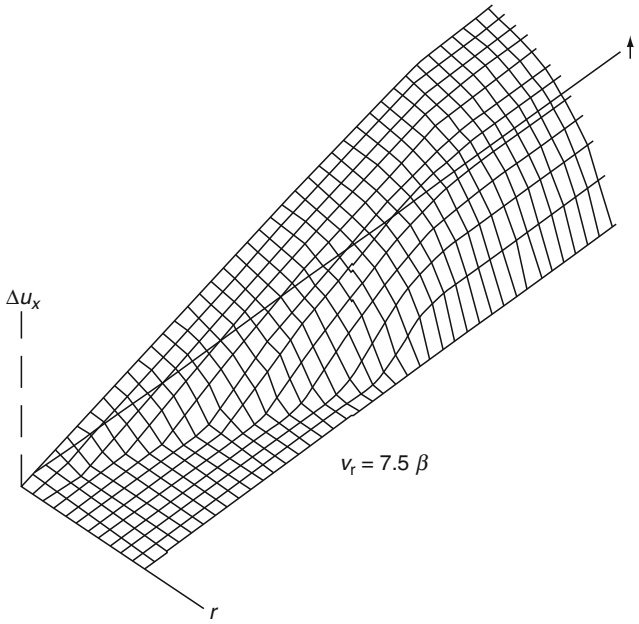
The static circular shear crack. We start by a quick study of a simple circular crack from which we derive some of the most fundamental properties of dynamic source models. Let us consider a static circular ("penny shaped") crack of radius a lying on the x, y plane. Assuming that the fault is loaded by an initial shear stress σ_0 and that the stress drop $\Delta\sigma$ is uniform inside the fault, slip on the fault is given by

$$D(r) = \frac{24}{7\pi}\frac{\Delta\sigma}{\mu}\sqrt{a^2 - r^2} \tag{13}$$

where, r is the radial distance from the center of the fault, a is the radius of the shear crack, and μ is the elastic rigidity of the medium surrounding the crack. Slip in this model has the typical elliptical shape that we associate with cracks. The taper of slip near the edges of the crack is the hallmark of dynamic source models. The slip distribution near the edges has a parabolic shape that is closely related to stress concentrations outside the fault zone (see Madariaga and Olsen, 2002).

From Equation 13 we can determine the scalar seismic moment for a circular fault, $M_0 = 16/7\ \Delta\sigma\ a^3$, so that the moment is the product of the stress drop times the cube of the fault size. This simple relation is the basis of the seismic scaling law proposed by Aki (1967). The circular crack model has been used to quantify numerous small earthquakes for which the moment was estimated from the amplitude of seismic waves, and the source radius was estimated from corner frequencies, aftershock distribution, etc.; the result is that for shallow earthquakes in crustal seismogenic zones like the San Andreas fault, or the North Anatolian fault in Turkey, stress drops are of the order of 1–10 MPa. For deeper events in subduction zones, stress drops can reach several tens of MPa. Thus, in earthquakes, stresses do not change much, at most a couple of orders of magnitude, while source radius varies over several orders of magnitude from meters to 100 km or more. It is only in this sense that the usual assertion "stress drop in earthquakes is constant" should be taken; it actually changes but much less than the other parameters in the scaling law.

The quasi-dynamic circular crack. There are no simple analytical solutions like Equation 13 for finite dynamic cracks. We are forced to use numerical solutions that are actually very simple to obtain using either finite difference or boundary integral equation techniques. The full solution to the circular crack problem is shown in Figure 1. Initially, until the sudden arrest of rupture at the final radius a, the slip distribution can be accurately computed using



Earthquakes, Source Theory, Figure 1 Slip distribution as a function of time and position for Madariaga's (1976) quasi-dynamic circular crack model.

Kostrov's self-similar solution (Equation 12). The stopping of rupture generates strong healing waves that propagate inward from the rim of the fault. These waves are of three types: P, S, and Rayleigh waves. Soon after the passage of the Rayleigh waves, slip rate inside the fault decreases to zero and the fault heals. After healing, we assume that frictional forces are sufficiently strong that no slip will occur until the fault is reloaded. As observed in Figure 1, it is clear that slip and rise time are functions of position on the fault, the rise time being much longer near the center where slip is also larger than near the edges of the fault where slip is clamped. Finally, let us note that the slip after healing is very similar to that of a static circular crack, except that there is a slight overshoot of slip with respect to the static solution (Equation 13). The overshoot is of course a function of the rupture speed, but its maximum value is of the order of 15% for a rupture speed of 0.75.

Future work

Seismic observations have steadily improved in the last 20 years due to the wide availability of digital, continuously recording accelerograms that can be integrated to obtain velocity and band limited displacement fields near active faults. Additional information about rupture is provided by field studies of faulting associated with earthquakes and from geodetic observations. Modeling and inversion of such data has been usually done using kinematic models derived from the original Haskell model. This approach gives solutions that are very nonunique because the rupture process (the rupture velocity) is difficult to estimate. Rupture velocity actually controls seismic radiation, so that poor velocity control leads to poor seismic inversions. There is hope that dynamic models, where rupture is inverted as part of the inversion process will provide better estimates of the stress field around major fault zones along with some essential parameters like the friction law that controls rupture propagation.

Bibliography

Aki, K., 1967. Scaling law of seismic spectrum. *Journal of Geophysical Research*, **73**, 5359–5376.

Aki, K., and Richards, P. G., 2002. *Quantitative Seismology*, 2nd edn. Sausalito: University Science Books.

Backus, G., and Mulcahy, M., 1976. Moment tensors and other phenomenological descriptions of seismic sources. I: continuous displacements. *Geophysical Journal of the Royal Astronomical Society*, **46**, 321–361.

Brune, J., 1970. Tectonic stress and the spectra of seismic shear waves from earthquakes. *Journal of Geophysical Research*, **75**, 4997–5009.

Burridge, R., and Knopoff, L., 1964. Body force equivalents for seismic dislocations. *Bulletin of the Seismological Society of America*, **54**, 1875–1878.

Eshelby, J. D., 1957. The elastic field of an ellipsoid inclusion and related problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, **241**, 376–396.

Haskell, N. A., 1964. Total energy and energy spectral density of elastic wave radiation from propagating faults. *Bulletin of the Seismological Society of America*, **54**, 1811–1841.

Kostrov, B. V., 1964. Self-similar problems of propagation of 1379 shear cracks. *Journal of Applied Mathematics and Mechanics*, **28**, 1077–1087.

Kostrov, B., and Das, S., 1988. *Principles of Earthquake Source Mechanics*. Cambridge: Cambridge University Press.

Madariaga, R., 1976. Dynamics of an expanding circular fault. *Bulletin of the Seismological Society of America*, **66**, 639–667.

Madariaga, R., and Olsen, K. B., 2002. Earthquake dynamics. In Lee, W. H. K., Kanamori, H., and Jennings, P. C. (eds.), International Handbook of Earthquake and Engineering Seismology (Chapter 7), Amsterdam: Academic Press.

Cross-references

Earthquake Rupture: Inverse Problem
Earthquake, Focal Mechanism

EARTHQUAKES, STRONG-GROUND MOTION

Giuliano F. Panza[1,2], Cristina La Mura[1], Fabio Romanelli[1], Franco Vaccari[1]
[1]Department of Geosciences, University of Trieste, Trieste, Italy
[2]Earth System Physics Section/Sand Group, The Abdus Salam International Centre for Theoretical Physics, Trieste, Italy

Definition

Strong-ground motion. An earthquake-induced ground motion capable of damaging man-made environment.

Introduction

Earthquake ground motion is a natural phenomenon associated with a sudden release of energy due to a fault rupture. Strong-motion seismology deals with seismic waves radiated by large earthquakes. More precisely, it is concerned with measurement, interpretation and prediction of strong shaking generated by damaging earthquakes.

Because of the engineering need for estimates of future strong-ground motion, the scope of strong-motion seismology encompasses seismic source theory and all other aspects of seismology that can provide insights helpful in making the best possible earthquake ground-shaking scenarios given the limited observational data.

Accumulated data have been providing very important knowledge about rupture processes of earthquakes, propagation paths and site effects on ground motion, and the relation between ground motion, damage, and other earthquake related phenomena (e.g., landslides earthquake induced).

The most important goals of strong-motion seismology are the improvement of the physical understanding of the processes controlling strong-ground motion and the reliable assessment of seismic hazard.

To assess the hazard for engineering design applications and due to the limited availability of strong-ground motion records, it has become increasingly common to compute broadband synthetic seismograms that allow us to perform realistic waveform modeling for different seismotectonic environments. The modeling has to take simultaneously into account the properties (e.g., dimensions, directivity, and near-field effects) of the radiating source, lateral heterogeneities along the path, and local site effects.

The joint use of reliable synthetic signals and observations can be fruitfully used for design purposes. In fact, even if recently strong-motion records in near-fault, soft soils, or basin conditions have been obtained, their number is still very limited to be statistically significant for seismic engineering applications.

The lack of a representative set of observations is due to the low frequency of large earthquakes and to the difficulty of providing a proper instrumental coverage of all the areas prone to strong shaking. Despite of the growing strong-motion registration networks, the existing database is still not complete with respect to the possible scenario earthquakes, and the lack of completeness will last for many years if not centuries. The variability of ground motion, due to different causes, e.g., spatial variability, source parameter variability, azimuthal variability (Strasser and Bommer, 2009), show how the current available strong-motion data still represent only a small sample of all physically possible ground motion. The mathematical modeling, with different degrees of complexity, based on probabilistic concepts cannot fill in the gap due to the lack of knowledge about the physical process behind an earthquake (Panza et al., 2004). Therefore, to resort to broadband synthetic seismogram is a fundamental step: where no records are available, synthetic signals can be used to estimate the ground motion without having to wait for a strong earthquake to occur.

Description of strong-ground motion

A seismic record of ground acceleration, due to a damaging shaking, is usually defined a strong-motion datum. There is not a quantitative rule, but it is generally assumed that damaging earthquakes are those of magnitude 5 and above. The threshold to be exceeded in terms of peaks acceleration for ground motion to be considered strong motion is quite blurred. The strongest recorded earthquake motions have peaks between 1 and 3 g, but these are very rare events. Usually, damages are caused by ground motion, which exceeds the threshold of 0.1 g, but Bolt and Abrahamson (2003) lower this limit to 0.05 g.

Due to the importance of strong motion for earthquake engineering, a number of different parameters have come into use to represent various characteristics of strong motion. Actually, the most appropriate way to describe or specify the characteristics of ground motion for the purpose of seismic design is still an open problem. A large number of parameters have been proposed for measuring the capacity of earthquakes to damage structures. However, to account for the complex characteristics of earthquake-induced strong-ground motion in the engineering analysis and design, the need arose for a more inclusive definition of the existing parameters and for the introduction of new ones. In fact, the adoption of inadequate parameters can lead to the definition of unrealistic design earthquakes and consequently to the unreliable evaluation of the seismic risk.

The parameters fundamentally involved in the evaluation of the level of severity associated with strong motion are, for engineering purposes, the frequency content, the amplitude, and the effective duration.

Beginning about the 1960s, the most used ground-motion parameter for quantification of ground motion was the peak ground acceleration (PGA), as the inertia forces depend directly on acceleration. PGA is simply obtained from unprocessed accelerograms. However, PGA is a poor indicator of damage, since, as observed, time histories with the same PGA could be very different in their frequency content, strong-motion duration, and energy level. In fact, PGA may be associated with high frequency pulses, which do not produce significant damage to the buildings as most of the impulse is absorbed by the inertia of the structure with little deformation. On the other hand, a more moderate acceleration may be associated with a long-duration pulse of low frequency (acceleration pulse), which gives rise to a significant deformation of the structure. The peak ground velocity (PGV) is another useful parameter for the characterization of ground-motion amplitude. The velocity is less sensitive to the higher-frequency components of the ground motion, so the PGV should be more useful than the PGA in characterizing the damaging potential of ground motion. For design purposes, the peak ground displacement (PGD),

generally associated with the lower-frequency components of an earthquake ground motion, has been recognized as useful parameter, mostly for the design of seismically isolated structures.

The response spectrum, firstly proposed by Benioff (1934), defined as the maximum response of a damped harmonic system to input motion (Anderson, 2003), is a parameter widely used for engineering purposes.

Another ground-motion parameter, very important for engineers, is the duration of shaking (Villaverde, 2009). Duration can impact structures and response of soils: structures that can stand one or two large oscillations may collapse under prolonged motions, and while a second of shaking of sandy soil can be tolerated, repeated oscillations can induce liquefaction (Jennings, 2003). But there is not a unique definition of duration, as clearly evidenced in Bommer and Martinez-Pereira (1998), where 30 definitions of strong-motion duration are reported. The available definitions can be classified in three different groups, or as a combination of them (Montejo and Kowalsky, 2008): (1) "bracketed" durations (Bolt, 1969), in which the duration is defined to be the time interval between the first and the last exceedances of a threshold of acceleration; (2) "significant" durations that define the duration of strong motion as the time interval during which a given percentage of the total energy of the record is released; and (3) frequency-dependent durations, in which the duration of strong-ground motion is analyzed separately in several narrow-frequency bands (Bolt, 1969; Trifunac and Westermo, 1977). The following two examples of duration measurement show the limits of this parameter. A way to estimate duration is to measure the interval of time within which acceleration exceeds some threshold, usually 0.05 g (bracketed duration, Bolt, 1969); another measure is the interval of time in which 90% of the integral of the acceleration-squared is recorded. These two estimations yield opposite results as the distance from the source increases. Peak ground motions decrease, as a rule, with increasing distance from the source (for exceptions, see e.g., Panza and Suhadolc, 1989; Fah and Panza, 1994), but energy becomes dispersed, so the time that is necessary to wait for the arrival of the 90% of the total energy in the seismogram increases (Anderson, 2003).

Several equations, commonly referred to as attenuation relations or ground-motion prediction equations (GMPE), have been proposed over the years from strong-ground motion observations and theoretical studies. These mathematical expressions give ground-motion intensity measures as a function of earthquake magnitude and distance. They may be determined empirically, by a regression of recorded ground motion, or theoretically, using seismological models to generate synthetic ground motions that account for the source, site, and path effects. Theoretical attenuation relations between various source parameters are used to estimate source parameters for earthquakes that have not yet occurred, or for which parameters of interest are unknown (e.g., Geller, 1976; Oth et al., 2007; Somerville et al., 1997). The other approach uses empirical relations between source parameters compiled for a huge number of earthquakes (e.g., Wells and Coppersmith, 1994). In the past few years, the availability of additional strong-motion data has allowed the incorporation of additional parameters, such as type of fault, fault orientation; further recent relationships have increasingly become more refined, although, at the same time, more complicated. Attenuation relations are usually grouped into three sets (Panza et al., 2004): shallow crustal earthquakes in active tectonic regions, shallow crustal earthquakes in stable tectonic regions, and subduction zone earthquakes. These relations can differ in the assumed functional form, the data selection criteria, the number and definition of independent variables, and the statistical treatment of the data. A review can be found in Douglas (2003), where the complete procedure that needs to be followed to derive equations for the estimation of ground motion from recorded strong-motion data is outlined as well. As the coefficients of the attenuation relations are determined empirically by means of regression analysis, they are quite sensitive to the used data set, an example of the dependence of attenuation relations on the procedure followed in the data processing is given by Parvez et al. (2001) for the Himalayas. The Izmit, Turkey, earthquake of August 17, 1999 and the Chi Chi, Taiwan, September 20, 2000, did show the inadequacy of the existing database: as explained in Anderson (2007), the GMPEs available at that moment underestimated the peak accelerations for those events. The discrepancies could be due to a wide variety of factors, so these examples suggest that a large amount of data and a deep physical understanding of the processes involved in seismic events are needed to improve the quality of GMPEs (Anderson, 2007). A common main source of error can be mathematical shortcomings that arise when attenuation relations are used in geologically complex areas (e.g., Peresan et al., 2010).

As stressed earlier by Aki: "a major goal of strong-motion seismology is to be able to synthesize strong-motion seismograms suitable for use in engineering analysis" (Anderson, 2003, p. 942). The use of computational modeling has the aim to simulate the wave generation and propagation process in complex media. The formulation of physical-mathematical models aimed to represent the effective complexity of the phenomenon and their resolution is a complicated affair. With the exception of a very limited number of simple situations, the physical-mathematical approach does not lead to closed form solutions, i.e., fully analytical, so it is necessary to employ numerical procedures, more or less complex, or approximated theories (Bolt, 1999).

Earthquake source effects on strong-ground motion: near-field effects

Source effects on far-field seismograms are discussed by Aki and Richards (1980, Chap. 14). Seismic data used in earthquake engineering are occasionally collected in the

near-field, but the question where the near-field ends and the far-field begins is gradational and will not be addressed here.

Near-fault ground motions often contain large long-period pulses (Somerville, 2002). There are two causes for these pulses: one is constructive interference of the dynamic shaking due to rupture directivity effects, the other is due to the movement of the ground associated with the permanent offset of the ground. For keeping these two causes separated, the terms "directivity pulse" and "fling step" have been introduced (Bolt and Abrahamson, 2003).

Large earthquakes are produced by rupture that starts at a point and travels with a velocity that is generally only slightly smaller than the shear-wave velocity in the medium. Therefore, when rupture is towards the station, the propagation causes most of the seismic energy to arrive in a single large pulse occurring at the beginning of the record. Forward-directivity effects occur when the rupture front propagates towards the recording site and the slip direction on the fault is aligned with the site. Backward-directivity effects, occurring when the rupture front propagates away from the recording site, generates opposite effects, i.e., longer-duration motions, lower amplitudes, and longer periods. This phenomenon has been first explained by H. Benioff in 1955, when analyzing the records of the 1952 Kern County, California, earthquake. The conditions for generating forward-rupture directivity effects are readily met in strike-slip faulting, where the rupture propagates horizontally along the strike either unilaterally or bilaterally, and the fault slip direction is oriented horizontally in the direction along the strike of the fault. As described by Bolt and Abrahamson (2003), for strike-slip earthquakes, the rupture directivity is observed on the fault normal component, and the static displacement fling step is observed on the fault parallel component. The conditions required for forward directivity are also met in dip-slip faulting, including both reverse and normal faults. The alignment of both the rupture direction and the slip direction updip on the fault plane produces rupture directivity effects at sites located around the surface exposure of the fault (or its updip projection if it does not break the surface). Consequently, it is generally the case that all sites located near the surface exposure of a dip-slip-fault experience forward-rupture directivity when an earthquake occurs on that fault.

For dip-slip earthquakes, the separation of the effects (rupture directivity and fling step) on the fault components is more complicated: the rupture-directivity effect will be strongest on the fault normal component at a location direct updip from the hypocenter, and the fling-step will also be observed on the horizontal component perpendicular to the strike of the fault, i.e., directivity-pulse effects and fling-step effects occur on the same component for dip-slip faults. The horizontal records of the 1966 Parkfield and 1971 San Fernando, California, earthquakes were the first to be discussed (Bolt, 1975) as examples of near-fault velocity pulses. The pulses from both directivity and fling-step may attenuate differently. Hisada and Bielak (2003) investigated the combined effects of fling-step and rupture directivity by paying special attention to the contribution of static and dynamic Green's functions. Fling effects are found to be stemmed mainly from the static Green's function. They are dominant in the vicinity of the surface fault and negligible for buried faults since the static Green's function decreases rapidly, as $(1/r^2)$, with the distance from the fault. When the observation point is above the buried fault, the fling step disappears because the slip dislocation of the fault cannot fling the ground, due to the presence of a continuous medium above the fault itself. Directivity effects derive mainly from the dynamic Green's function and attenuate as $(1/r)$ to $(1/\sqrt{r})$, more slowly than the fling. The combination of both the effects of rupture directivity and fling-step results in inclined directions, with respect to the fault plane, of maximum velocities, and displacements.

The destructive potential of near-fault ground motions was manifested in the 1994 Northridge and 1995 Kobe earthquakes. In each of these earthquakes, peak ground velocities as high as 175 cm/s were recorded. The period of the near-fault pulses recorded in both of these earthquakes lies in the range from 1 to 2 s, comparable with the natural periods of structures, such as bridges and buildings, many of which were indeed severely damaged.

Near-fault recordings from recent earthquakes indicate that this pulse is a narrow band pulse whose period increases with magnitude. This magnitude dependence of the pulse period causes the response spectrum to have a peak whose period increases with magnitude, such that the near-fault ground motions from moderate-magnitude earthquakes may exceed those of larger earthquakes at intermediate periods (around 1 s).

Wave-propagation effects

Wave propagation in heterogeneous media is addressed in another chapter herein (see *Propagation of Elastic Waves: Fundamentals*), so, here we limit the analysis to site effects due to their impact on strong-ground motion.

Ground motion can be highly amplified, or deamplified, by local effects. The essential aspects of site effects on strong-ground motion are widely addressed in literature (see e.g., Kawase, 2003). The pioneering works where site effects are well-recognized date back to 1930s and are due to Sezawa and Ishimoto (Kawase, 2003), but a first comment on the effect of near-surface conditions on shaking intensity during the 1906 San Francisco earthquake dates back to 1910: "Experience shows that the damage done by destructive earthquakes is much greater on alluvial soil than on solid rock" (Reid, 1910). Almost all recent destructive earthquakes have shown the relevance of site effects in the amplification of the seismic ground motion. The classical example of such a situation is the September 19, 1985 earthquake that, in Mexico City, about 400 km far away from the epicenter, caused a large victims toll (about 8,000), destroyed or heavily damaged hundreds of buildings, and caused few billion dollars damage (Fah et al., 1994; Anderson et al., 1986).

The presence of lacustrine sediments is the cause of the large ground motion, up to 40 times larger than that of firm ground (Chàvez-Garcìa and Salazar, 2002). The modeling of wave propagation in basins, and more generally in complex 2D and 3D regions, using numerical methods, such as finite difference (FDM), finite element (FEM), spectral element method (SEM), boundary integral equation, and boundary element method (BIEM and BEM) is an active field of research (see e.g., Aagaard et al., 2001; Bielak et al., 1991; Bouchon and Sánchez-Sesma, 2007; Graves, 1996; Komatitsch and Tromp, 1999, 2002a, 2002b; Komatitsch and vilotte, 1998; Komatitsch et al., 2004; Oprsal and Zahradnik, 1999, 2002; Pitarka, 1999; Yoshimura et al., 2003). These methods differ in accuracy with respect to different structural features of the complex heterogeneous models and considerably in the computational efficiency. For comparison of several modeling methods, we refer to the review paper by Takenaka et al. (1998), but the most recent review of the computational- and numerical-modeling methods can be found in the book edited by Wu and Maupin (2007). Very recently, a new analytical approach has been developed and validated performing simulations in the area of the Kanto basin, Japan (La Mura et al., 2010). It is generally observed that at stations throughout the basin larger amplitudes and longer durations of ground shaking are observed, but the pattern can be very complicated. The amplification patterns depend not only on the site conditions, but also on the characteristics of the incoming wavefield and, therefore, on the source and the wave propagation path. Clear examples are given in Field (2000).

Last but not least, we mention the liquefaction phenomenon, largely observed, for example, during the Niigata earthquake, Japan, on June the 16th, 1964 (Youd, 2003; Bardet, 2003). Liquefaction of water-saturated sands is believed to occur when the pore pressure, increased by the strong-ground motion, reaches the effective confining pressure. The likelihood that an earthquake will liquefy a site depends on many site characteristics (e.g., mean grain size, percentage of fine-grained sediment, porosity and level of saturation) and on the regional geology, which influences the amplification and attenuation of strong-motion amplitude. Ground-motion modeling may help in defining a priori the liquefaction potential (e.g., Nunziata et al., 2008).

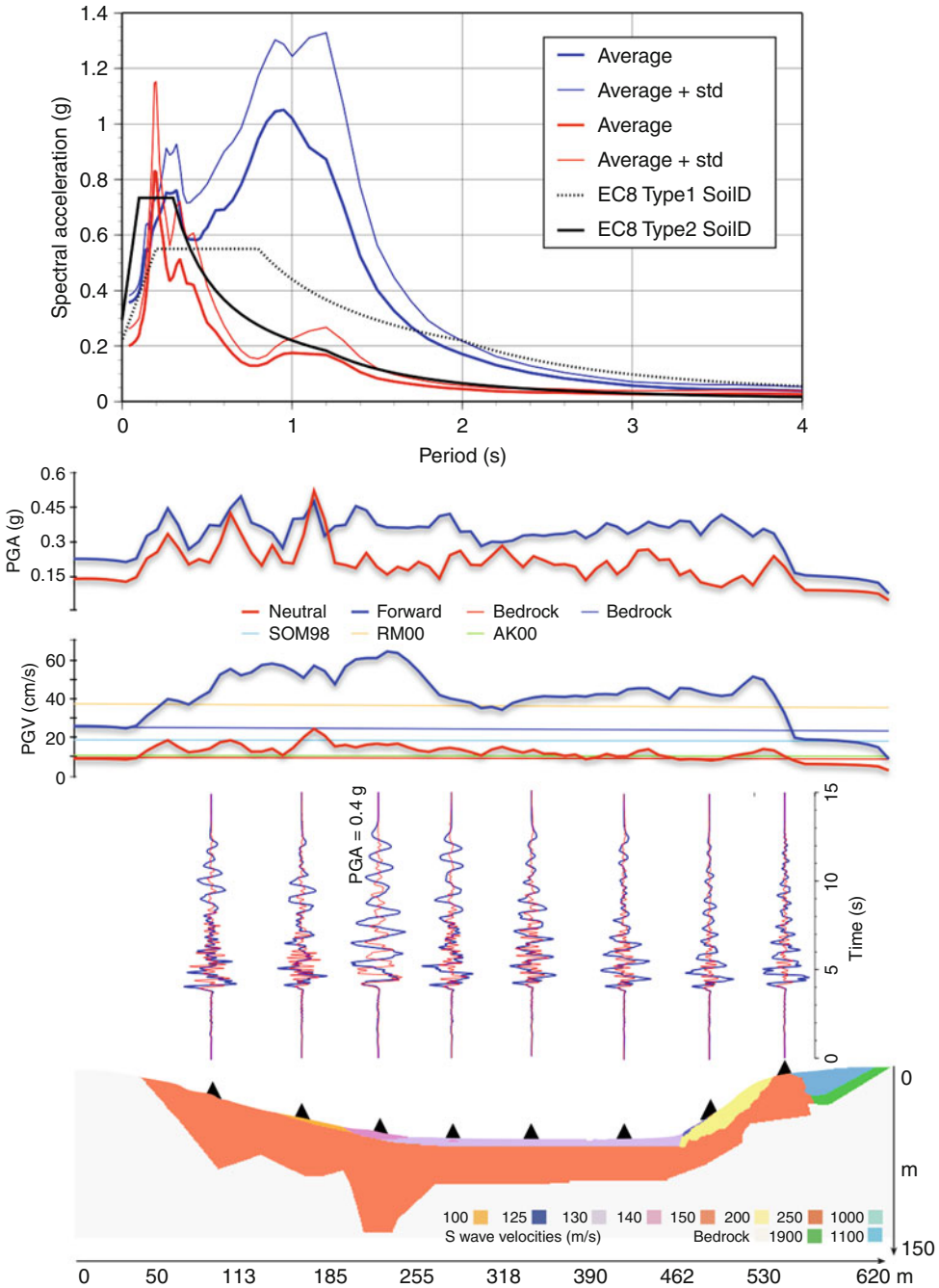
Strong-ground motion modeling

Due to the destructive potential of earthquakes, a goal is to develop methodologies for modeling strong-ground motions from seismic sources related to active faults and subduction zones with high probability of earthquake occurrence. Confidence in such modeling relies on the theoretical part of strong-motion seismology and on the calibration of the techniques against the empirical data.

One of the most common methods to simulate broadband strong-ground motion is the use of records from small earthquakes occurred in the vicinity of the source area of a large earthquake. This kind of modeling of the strong motion is known as empirical Green's function method, and it has been proposed by Hartzell in 1978 and largely reviewed and applied through the years (see e.g., Irikura, 1986; Dan et al., 1990). The empirical Green's function method has been applied in simulating strong-ground motions from a large event, even using observed records of just one or two small events. The records are used as the empirical Green's functions over the fault plane of the large event by correcting only attenuation effects due to difference in distance from each subfault to the site. The advantage of this approach relies in the fact that the seismograms contain all the complexities of wave propagation. On the other hand, the serious disadvantage is that the empirical Green's functions may not be available for the desired source-station path and may originate by an earthquake with a focal mechanism and/or depth different from the desired one.

Very recently, a procedure, called "recipe of strong-motion prediction," has been developed to the aim of obtaining ground-motion time histories from specific earthquakes (Irikura et al., 2004). Such a recipe is based on source characteristics obtained from the waveform inversion using strong-motion data. The idea is to follow a procedure, outlined as a recipe, for characterizing the source model estimating three kinds of source parameters: outer-, inner-, and extra-fault parameters. The outer are to outline the overall properties of the target earthquake, such as entire source area and seismic moment; the inner are parameters characterizing stress heterogeneity inside the fault area; the extra parameters are considered to complete the source model, such as the starting point and propagation pattern of the rupture.

The theoretical approach to the simulation of strong motion, based on the computation of synthetic seismograms, has experienced an impressive growth in the last decades. A description of both analytical and numerical methods can be found, e.g., in Anderson (2003), Bolt and Abrahamson (2003), Bielak et al. (2003). Hybrid techniques, obtained combining analytical and numerical methods, e.g., modal summation method and finite difference (Panza et al., 2001), wave number method and finite difference (Zahradnik and Moczo, 1996), finite element and boundary integral method (Bielak et al., 1991), have been developed and a review of these methods dating back to 1980 can be found in Moczo et al. (1997). Numerical-modeling methods for anelastic wave propagation, that take into consideration the earthquake source, propagation path, and local site effects, have become available for the estimation of the site responses and, more generally, for addressing the problem of solving the wave equation in 3D models. Then simulations based on finite difference (see e.g., Sato et al., 1998) or, more recently, finite element (see e.g., Bielak et al., 2003) and Spectral Element Method (SEM) (see e.g., Komatitsch and Tromp, 2002a, b; Komatitsch et al., 2004) have been largely used in order to simulate strong-ground motion in 3D structures. A very recent development (La Mura et al., 2010) extended the very efficient analytical modal summation (Panza et al., 2001) to 3D anelastic structures.



Earthquakes, Strong-Ground Motion, Figure 1 From *bottom* to *top*: local geotechnical model schematizing the surface geology along the Warth bridge. *Black triangles* show the sites of the abutments and of the piers along the section. Transverse acceleration time series computed at the eight pier sites. The amplitude of the signals is normalized with respect to the maximum one (0.4 g). *Blue* and *red* curves correspond to the case of forward and neutral directivity, respectively. Peak Ground Velocities (*PGV*) calculated along the profile, for forward (*blue curves*) and neutral (*red curves*) directivity, for a unilateral rupture and for the bedrock case (*BED*), that can be compared with the regression analysis for PGV made using a database of near-fault, forward-directivity motions, proposed by: Somerville (1998) -SOM98- *azure* curve, Alavi and Krawinkler (2000) -AK00- *orange* curve and Rodriguez-Marek (2000) -RM00- *green* curve, for a magnitude equal to 5.5. Peak Ground Accelerations (*PGA*) calculated along the profile, for forward (*blue curves*) and neutral (*red curves*) directivity, for a unilateral rupture. Average Response Spectra Accelerations (*ARSA*) and ARSA + one standard deviation computed for the eight sites; *blue* and *red curves* correspond to the case of forward and neutral directivity, respectively. The *black lines* correspond to the two (Type 1 and 2) Design Spectra suggested by EC-8 (EN, 1998) for a 5.5 magnitude earthquake and for a soil of class D.

Strong-motion simulation: an example

In this section, an example of simulation of strong ground motion is discussed. The determination of the seismic ground motion due to an earthquake with a given magnitude and epicentral distance from the site has been done following a theoretical approach. In order to perform an accurate and realistic estimate of ground shaking scenario, it is necessary to make a parametric study that takes into account the complex combination of the source and propagation parameters, in realistic geological structures.

The realistic modeling of ground motion requires the simultaneous knowledge of the geotechnical, lithological, geophysical parameters, and topography of the medium on one side, and tectonic, historical, paleoseismological, seismotectonic models on the other, for the best possible definition of the probable seismic source.

To deal both with realistic source and structural models, including topographical features, a hybrid method, that combines modal summation and the finite difference technique (e.g., Panza et al., 2001) and optimizes the use of the advantages of both methods, has been used. Wave propagation is treated by means of the modal-summation technique from the source to the vicinity of the local, heterogeneous structure that has to be modeled in detail. A laterally homogeneous, anelastic structural model is adopted that represents the average crustal properties of the region. The generated wavefield is then introduced in the grid that defines the heterogeneous area and it is propagated according to the finite differences scheme. With this approach, source, path, and site effects are all taken into account simultaneously.

An example of seismic-input computation is illustrated. The computations are performed for the Warth bridge, freeway A2, 63 km south of Vienna (Austria), a location where no seismic records are available (see Romanelli et al., 2003, 2004). The information about the possible seismic input is limited to an estimate of the macroseismic intensity, in the range from VI to VIII (MSK), the value of the magnitude, M, of the nearest largest recorded event, M = 5.5 and the most probable focal depth of strong earthquakes, in the range from 6 to 11 km.

The synthetic time signals are calculated for the three components of motion and the working magnitude is 5.5 (seismic moment equal to 1.8×10^{17} Nm). The study of possible directivity effects in the direction of the Warth bridge (see cross section in Figure 1) has been performed with the method developed by Gusev (2010), based on the modeling of a Haskell-type seismic source (Haskell, 1964; Haskell, 1966), where a stochastic component allows to build a spectrum (amplitude and phase) of the source function that takes into account both the rupture process and directivity effects. The near-source effect is shown in Figure 1, where the acceleration time series at the eight sites are plotted in the case of forward (blue curves) and neutral (red curves) directivity for a unilateral rupture. In the figure are also reported the Peak Ground Velocities (PGV) and Peak Ground Accelerations (PGA) along the profile, for neutral (N) and forward (F) directivity, calculated for a unilateral rupture, both for the bedrock case and taking into account the local site conditions. The curves can be compared with the regression analysis for PGV made using a database of near-fault, forward-directivity motions proposed by Somerville (1998), Alavi and Krawinkler (2000) and Rodriguez-Marek (2000). From the results it is evident the strong influence of the rupture and site effects on the ground-motion characteristics, in terms of amplitude and duration. The accelerograms exhibit the greatest peaks in the frequency range, from 1 to 6 Hz, and reach considerable peak values, around 0.4 g.

The analysis of the computed strong-ground motion can be carried out in the time domain (broadband ground-motion time series) and other domains (e.g., Fourier and response spectra). The results show that lateral heterogeneities and source effects can produce strong spatial variations (e.g., more than doubling the amplitudes) in the ground motion (also in terms of differential motion) even over distances of a few tens of meters. In other words, considering that an increment of one intensity degree in the MCS scale roughly corresponds to a doubling of the PGA (Cancani, 1904), a general result of the modeling is that a difference greater than one unit in the seismic intensity can be experienced at sites as close as a few tens of meters.

Conclusions

Solution to earthquake engineering problems, either design or estimation of seismic hazard, are today very different than they were before a strong interaction between seismologists and engineers. Today the differences between strong motion and weak motion have been softened, due to progress in instrumentation and the awareness that both kinds of motions are useful to the understanding of the nature of earthquakes. Today it is unthinkable to study seismic-rupture processes and seismic-wave generation of large earthquakes without near-source strong-motion data; however, while waiting for the enlargement of the strong-motion data set, a very useful approach is the development and use of modeling tools based, on one hand, on the theoretical knowledge of the physics of the seismic source and of wave propagation and, on the other hand, exploiting the rich database about the geotechnical, geological, tectonic, seismotectonic, and historical information already available. With these efforts, the prediction of strong motions through physical modeling is adequate and desirable.

Bibliography

Aagaard, B. T., Hall, J. F., and Heaton, T., 2001. Characterization of near source ground motions with earthquake simulations. *Earthquake Spectra*, **17**, 177–207.

Aki, K., and Richards, P. G., 1980. *Quantitative Seismology, Theory and Method*. San Francisco: W.H. Freeman and Co.

Alavi, B., and Krawinkler, H., 2000. Consideration of near-fault ground motion effects in seismic design. In *Proceedings of the*

12th World Conference on Earthquake Engineering. The New Zealand Society for Earthquake, Auckland, p. 8.

Anderson, J. G., 2003. Strong-motion seismology. In Lee, W. H. K., Kanamori, H., Jennings, P. C., and Kisslinger, C. (eds.), *International Handbook of Earthquake and Engineering Seismology.* San Diego: Academic. IASPEI, Vol. 81B, pp. 983–965.

Anderson, J. G., 2007. Physical processes that control strong ground motion. In Kanamori, H. (ed.), *Treatise on Geophysics, Volume 4, Earthquake Seismology.* Amsterdam: Elsevier. Vol. 4, pp. 513–565.

Anderson, J. G., Bodin, P., Brune, J. N., Prince, J., Singh, S. K., Quaas, R., and Onate, M., 1986. Strong Ground Motion from the Michoacan, Mexico, earthquake. *Science*, **233**, 1043–1049.

Bardet, J., 2003. Advances in analysis of soil liquefaction during earthquakes. In Lee, W. H. K., Kanamori, H., Jennings, P. C., and Kisslinger, C. (eds.), *International Handbook of Earthquake and Engineering Seismology.* San Diego: Academic Press. IASPEI, Vol. 81B, pp. 1175–1201.

Benioff, H., 1934. The physical evaluation of seismic destructiveness. *Bulletin of the Seismological Society of America*, **24**, 9–48.

Bielak, J., MacCamy, R. C., McGhee, D. S., and Barry, A., 1991. Unified symmetric BEM-FEM for site-effects on ground motion-SH waves. *Journal of Engineering Mechanics*, **117**, 2265–2285.

Bielak, J., Loukakis, K., Hisada, Y., and Yoshimura, C., 2003. Domain reduction method for three-dimensional earthquake modeling in localized regions, part I: theory. *Bulletin of the Seismological Society of America*, **93**, 817–824.

Bolt, B. A., 1969. Duration of strong motion. In *Proceedings 4th WCEE*, Rome, pp. 1304–1315.

Bolt, B. A., 1975. The San Fernando earthquake, 1971. Magnitudes, aftershocks, and fault dynamics. In *Bulletin 196, California Division of Mines and Geology*, Sacramento, Chap. 21.

Bolt, B. A., 1999. Estimating seismic ground motion. *Earthquake Spectra*, **15**, 187–197.

Bolt, B. A., and Abrahamson, N. A., 2003. Estimation of strong seismic ground motions. In Lee, W. H. K., Kanamori, H., Jennings, P. C., and Kisslinger, C. (eds.), *International Handbook of Earthquake and Engineering Seismology.* San Diego: Academic. IASPEI, Vol. 81B, pp. 983–1001.

Bommer, J. J., and Martinez-Pereira, A., 1998. The effective duration of earthquake strong motion. *Journal of Earthquake Engineering*, **3**(2), 127–172.

Bouchon, M., and Sánchez-Sesma, F. J., 2007. Boundary integral equations and boundary elements methods in elastodynamics. In Wu, R. S., Maupin, V., and Dmowska, R. (eds.), *Advances in Wave Propagation in Heterogeneous Earth, Advances in Geophysics*. New York: Elsevier–Academic, Vol. 48, pp. 157–166.

Cancani, A., 1904. Sur l'emploi d'une double echelle seismique des intesites, empirique et absolue. *G Beitr*, **2**, 281–283.

Chàvez-Garcìa, F. J., and Salazar, L., 2002. Strong Motion in Central Mexico: a Model Based on Data Analysis and Simple Modeling. *Bulletin of the Seismological Society of America*, **92**, 3087–3101.

Dan, K., Watanabe, T., Tanaka, T., and Sato, R., 1990. Stability of earthquake ground motion synthesized using different small-event records as empirical Green's functions. *Bulletin of the Seismological Society of America*, **80**, 1433–1455.

Douglas, J., 2003. Earthquake ground motion estimation using strong-motion. A review of equations for the estimation of peak ground acceleration response spectral ordinates. *Earth Science Review*, **61**, 43–104.

EN, 1998. *General Rules, Seismic Actions And Rules For Buildings*. Bruxelles, Belgium: CEN European Committee for Standardization.

Fah, D., and Panza, G. F., 1994. Realistic modeling of observed seismic motion in complex sedimentary basins. *Annali di Geofisica*, **37**, 1771–1797.

Fah, D., Suhadolc, P., Mueller, S., and Panza, G. F., 1994. A hybrid method for the estimation for ground motion in sedimentary basins: quantitative modeling for Mexico City. *Bulletin of the Seismological Society of America*, **84**, 383–399.

Field, E. H., and SCEC Phase III Working Group, 2000. Accounting for site effects in probabilistic seismic hazard analyses of Southern California: overview of the SCEC Phase III report. *Bulletin of the Seismological Society of America*, **90**(6B), S1–S31.

Geller, R. J., 1976. Scaling relations for earthquake source parameters and magnitudes. *Bulletin of the Seismological Society of America*, **66**, 1501–1523.

Graves, R. W., 1996. Simulating seismic wave propagation in 3D elastic media using staggered-grid finite differences. *Bulletin of the Seismological Society of America*, **86**, 1091–1106.

Gusev, A. A., 2010. Broadband kinematic stochastic simulation of an earthquake source: a refined procedure for application in seismic hazard studies. Submitted to *Pure and Applied Geophysics*.

Hartzell, S. H., 1978. Earthquake aftershocks as Green's functions. *Geophysical Research Letters*, **5**, 1–4.

Haskell, N. A., 1964. Total energy and energy spectra density of elastic wave radiation from propagating faults. *Bulletin of the Seismological Society of America*, **54**, 1811–1841.

Haskell, N. A., 1966. Total energy and energy spectral density of elastic wave radiation from propagating faults. 2. A statistical source model. *Bulletin of the Seismological Society of America*, **56**, 125–140.

Hisada, Y., and Bielak, J., 2003. A theoretical method for computing near-fault ground motions in layered half-spaces considering static offset due to surface faulting, with a physical interpretation of fling step and rupture directivity. *Bulletin of the Seismological Society of America*, **93**, 1154–1168.

Irikura, K., 1986. Prediction of strong acceleration motion's using empirical Green's function. *Proceedings of 7thJapan Earth. Eng. Symp.*, Tokyo, pp. 151–156.

Irikura, K., Miyake, H., Iwata, T., Kamae, K., Kawabe, H., and Dalguer, L. A., 2004. Recipe for predicting strong ground motions from future large earthquakes. *Proceedings 13th world conference of Earthquake Engineering*, Paper No. 1371.

Jennings, P. C., 2003. An introduction to the earthquake response of structures. In Lee, W. H. K., Kanamori, H., Jennings, P. C., and Kisslinger, C. (eds.), *International Handbook of Earthquake and Engineering Seismology.* San Diego: Academic. IASPEI, Vol. 81B, pp. 1097–1125.

Kawase, H., 2003. Site effects on strong ground motion. In Lee, W. H. K., Kanamori, H., Jennings, P. C., and Kisslinger, C. (eds.), *International Handbook of Earthquake and Engineering Seismology.* San Diego: Academic. IASPEI, Vol. 81B, pp. 1013–1030.

Komatitsch, D., and Tromp, J., 1999. Introduction to the spectral-element method for 3-D seismic wave propagation. *Geophysical Journal International*, **139**, 806–822.

Komatitsch, D., and Tromp, J., 2002a. Spectral-element simulations of global seismic wave propagation, I. Validation. *Geophysical Journal International*, **149**, 390–412.

Komatitsch, D., and Tromp, J., 2002b. Spectral-element simulations of global seismic wave propagation, II. 3-D models, oceans, rotation, and self-gravitation. *Geophysical Journal International*, **150**, 303–318.

Komatitsch, D., and Vilotte, J. P., 1998. The spectral-element method: an efficient tool to simulate the seismic response of 2D and 3D geological structures. *Bulletin of the Seismological Society of America*, **88**, 368–392.

Komatitsch, D., Liu, Q., Tromp, J., Peter Suss, P., Stidham, C., John, H., and Shaw, J. H., 2004. Simulations of Ground Motion in the Los Angeles Basin Based upon the Spectral-Element

Method. *Bulletin of the Seismological Society of America*, **94**, 187–206.
La Mura, C., Yanovskya, T. B., Romanelli, F., and Panza, G. F., 2010. Three-dimensional seismic wave propagation by modal summation: method and validation. *Pure and Applied Geophysics*, in press, doi:10.1007/s00024-010-0165-2.
Moczo, P., Bystricky, J., Kristek, J., Carcione, J. M., and Bouchon, M., 1997. Hybrid modeling of P-SV seismic motion at inhomogeneous visco-elastic topographic structures. *Bulletin of the Seismological Society of America*, **87**, 1305–1323.
Montejo, L. A., and Kowalsky, M. J., 2008. Estimation of frequency-dependent strong motion duration via wavelets and its influence on nonlinear seismic response. *Computer-Aided Civil and Infrastructure Engineering*, **23**, 253–264.
Nunziata, C., De Nisco, G., and Panza, G. F., 2008. Evaluation of liquefaction potential for building code. In Santini, A., and Moraci, N. (eds.), *Seismic Engineering Conference Commemorating the 1908 Messina and Reggio Calabria Earthquake*. New York: Melville, pp. 370–377.
Oprsal, I., and Zahradnik, J., 1999. Eòastic finite-difference method for irregular grids. *Geophysics*, **64**, 240–250.
Oprsal, I., and Zahradnik, J., 2002. Three-dimensional finite-difference method and hybrid modeling of earthquake ground motion. *Journal of Geophysical Research*, **107**(B8), 2161 doi: 10.1029/2000JB000082.
Oth, A., Wenzel, F., Wust-Bloch, H., Gottschammer, E., and Ben-Avraham, Z., 2007. Parametrization of a composite attenuation relation for the Dead Sea area based on 3D modeling of elastic wave propagation. *Pure and Applied Geophysics*, **164**, 23–37.
Panza, G. F., and Suhadolc, P., 1989. Realistic simulation and prediction of strong ground motion. In Carlomagno, G. M., and Brebbia, C. A. (eds.), *Computers and experiments in Stress Analysis*. New York: Springer-Verlag, pp. 77–98.
Panza, G. F., Romanelli, F., and Vaccari, F., 2001. Seismic wave propagation in laterally heterogeneous anelastic media: theory and applications to seismic zonation. In Dmowska, R., and Saltzman, B. (eds.), *Advances in Geophysics*. San Diego: Academic, Vol. 43, pp. 1–95.
Panza, G. F., Romanelli, F., Vaccari, F., Decanini, L., and Mollaioli, F., 2004. Seismic ground motion modeling and damage earthquake scenarios: a possible bridge between seismologists and seismic engineers. In Chen, Y. T., Panza, G. F., and Wu, Z. L. (eds.), *Earthquake Hazard, Risk, and Strong Ground Motion*. Beijing: Seismological, pp. 323–349.
Parvez, A. I., Gusev, A. A., Panza, G. F., and Petukhin, A. G., 2001. Preliminary determination interdependence among strong motion amplitude, earthquake magnitude and hypocentral distance for the Himalayan region. *Geophysical Journal International*, **144**, 577–596.
Peresan, A., Zuccolo, E., Vaccari, F., Gorshkov, A., and Panza, G. F., 2010. Neo-deterministic seismic hazard and pattern recognition techniques: time-dependent scenarios for north-eastern Italy. *Pure and Applied Geophysics*, in press, DOI 10.1007/s00024-010-0166-1.
Pitarka, A., 1999. 3D elastic finite-difference modeling of seismic motion using staggered grids with non-uniform spacing. *Bulletin of the Seismological Society of America*, **89**, 54–68.
Reid, H. F., 1910. *The California earthquake of April 18, 1906.* Report of the state earthquake investigation commission, Vol. II, The Mechanics of the earthquake, Carnegie Institution of Washington, Publication no. 87.
Rodriguez-Marek, A., 2000. *Near-Fault Seismic Site Response*. PhD dissertation, Berkeley, Department of Civil Engineering, University of California, p. 451.
Romanelli, F., Vaccari, F., and Panza, G. F., 2003. Realistic modelling of the seismic input: site effects and parametric studies. *Journal of Seismology and Earthquake Engineering*, **5**, 27–39.
Romanelli, F., Vaccari, F., and Panza, G. F., 2004. Realistic modelling of the effects of asynchronous motion at the base of bridge piers. *Journal of Seismology and Earthquake Engineering*, **6**, 17–26.
Sato, T., Helmberger, D. V., Somerville, P. G., Graves, R. W., and Saikia, C. K., 1998. Estimates of regional and local strong motions during the Great 1923 Kanto, Japan, earthquake (*Ms* 8.2). Part 1: source estimation of calibration event and modeling of wave propagation path. *Bulletin of the Seismological Society of America*, **88**, 183–205.
Somerville, P. G., 1998. Development of an improved representation of near fault ground motions. In *Proceedings SMIP98 Seminar on Utilization of Strong Motion Data*, California Strong Motion Instrumentation Program, Oakland, pp. 1–20.
Somerville, P. G., 2002. Characterizing near fault ground motion for the design and evaluation of bridges. In Nimis, R., and Bruneau, M. (eds.), *Proceedings of the Third National Seismic Conference and Workshop on Bridges and Highways*. MCEER, Buttalo, pp. 137–148.
Somerville, P. G., Smith, N. F., Graves, R. W., and Abrahamson, N. A., 1997. Modification of Modification of Empirical Strong Ground Motion Attenuation Relations to Include the Amplitude and Duration effects of rupture Directivity. *Seismological Research Letters*, **68**, 199–222.
Strasser, F. O., and Bommer, J. J., 2009. Review: Strong Ground Motions – Have we seen the worst? *Bulletin of the Seismological Society of America*, **99**, 2613–2637.
Takenaka, H., Furumura, T., and Fujiwara, H., 1998. Recent developments in numerical methods for ground motion simulation. In Irikura, K., Kudo, K., Okada, H., and Sasatani, T. (eds.), *The Effects of Surface Geology on Seismic Motion*. Rotterdam: Balkema, Vol. 2, pp. 91–101.
Trifunac, M. D., and Westermo, B. D., 1977. A note on the correlation of frequency dependent duration of strong earthquake ground motion with the modified Mercalli intensity and the geologic conditions at the recording site. *Bulletin of the Seismological Society of America*, **67**, 917–27.
Villaverde, R., 2009. *Fundamental Concepts of Earthquake Engineering*. Boca Raton: CRC.
Wells, D. L., and Coppersmith, K. J., 1994. New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement. *Bulletin of the Seismological Society of America*, **84**, 974–1002.
Wu, R. S., and Maupin, V., 2007. Advances in wave propagation in heterogeneous earth. In Dmowska, R. (ed.), *Advances in Geophysics.* San Diego: Elsevier–Academic, Vol. 48.
Yoshimura, C., Bielak, J., Hisada, Y., and Fernandez, A., 2003. Domain reduction method for three-dimensional earthquake modeling in localized regions part II: verification and applications. *Bulletin of the Seismological Society of America*, **93**, 825–840.
Youd, T. L., 2003. Liquefaction mechanisms and induced ground failure. In Lee, W. H. K., Kanamori, H., Jennings, P. C., and Kisslinger, C. (eds.), *International Handbook of Earthquake and Engineering Seismology.* San Diego: Academic. IASPEI, Vol. 81B, pp. 1159–1173.
Zahradnik, J., and Moczo, P., 1996. Hybrid seismic modeling based on discrete-wave number and finite-difference methods. *Pure and Applied Geophysics*, **148**, 21–38.

Cross-references

Earthquakes, Intensity
Earthquakes, PAGER
Earthquakes, Shake Map
Propagation of Elastic Waves: Fundamentals

EARTHQUAKES, VOLCANOGENIC

J. W. Neuberg
Institute of Geophysics and Tectonics, School of Earth and Environment, University of Leeds, Leeds, UK

Definition

Volcano-tectonic earthquakes. "Conventional" double-couple earthquakes of shallow tectonic origin, occurring in a volcanic setting, and comprising primary (*P*) and secondary (*S)* waves as well as surface waves.
Rock-fall events. High-frequency (above 2 Hz), "cigar-shaped" signals caused by the surface impact of solid, crystallized lava dome material, consisting mainly of surface waves.
Low-frequency earthquakes. Earthquakes with low-frequency content (0.5–2 Hz), generated at the interface between a magmatic fluid and solid rock, causing parts of the volcanic plumbing system such as conduits, dykes, and cracks to resonate.
Volcanic tremor. Either unstructured, random background noise on a volcano caused by magma movement, or merging swarms of low-frequency earthquakes that retain all characteristics of the single events.
Very-long-period earthquakes. Seismic transients with periods between tens of seconds and several minutes, caused by magma/gas movement in the shallow plumbing system.

Volcano seismology

Introduction

Volcano seismicity covers a large variety of seismic signals caused by the complex interaction of magmatic and hydrothermal fluids with solid rock within the volcanic edifice leading to stress changes, brittle failure of rock, pressure variations, magma fracturing, and fragmentation. Volcano seismology has the aim to identify and separate different seismic signals and attribute them to physical processes, a task that helps volcanologists to understand the dynamics of the magmatic systems.

The interpretation of different volcanogenic seismic signals and the development of conceptual and numerical models are at the core of any attempt to forecast volcanic eruptions. While volcano seismology emerged from earthquake seismology, it has evolved into a separate discipline in its own right, dealing with a large complexity of seismic source processes in a volcanic environment.

Crucial for the interpretation of seismic signals is the examination of their frequency content by Fourier or frequency analysis by which the signal is decomposed into harmonic waves and contributions of single frequencies are compared with each other. This decomposition is called a frequency spectrum, and its repeated application in consecutive time windows is referred to as a spectrogram. It turns out that different volcanogenic signals can be easily identified and separated by their frequency content; hence, we will make frequent use of this methodology in the following sections where the characteristics and the possible origin and source mechanism of different volcanogenic signals are described. Single event types are classified according to their source mechanism rather than using a phenomenological classification according to their waveform. A summary of the three main types of volcanogenic signals and their respective frequency spectra are shown in Figure 1, while volcanic tremor and very-long-period transients are depicted in Figures 3–5, respectively.

Volcano-tectonic earthquakes

Volcanoes are often located in tectonically active areas such as subduction zones, where earthquakes are generated merely by changes in the regional stress field due to the continuous movement of the tectonic plates. When magmatic and hydrothermal processes occur at the same location and time, the local stress field is further modulated and impacted upon by these processes. It is this complex interaction between the regional stress field and local stress modulation by volcanic processes such as dyke intrusion and (de)pressurization of the volcanic plumbing system that leads to the generation of volcano-tectonic earthquakes, referred to in the following as VTs for brevity. These stress changes can be created directly by magma emplacement, or indirectly by temperature changes or pressure changes as a result of magma–water interaction in a hydrothermal system.

Assuming that the pressure modulation through a magmatic body produces only a localized effect on the overall stress field impacting only on close distance of a few hundreds meters from the magmatic source, the earthquake location in itself indicates the emplacement of fresh magma (or its withdrawal). Hence, the locations of VT earthquakes over a certain time period are an indicator of changes in the system rather than providing a picture of a so-called magma chamber. VTs are generated by *stress changes*, as a consequence of *magma movement*. In this way, Battaglia et al. (2005) traced the rising magma dyke through time at Piton de la Fournaise, La Reunion.

Being generated by brittle fracture of rocks along a fault plane, VTs have all the characteristics of local earthquakes: containing a clear primary (P) wave, followed by the secondary (S) wave, both in the frequency range of 1–20 Hz (see Figure 1a). If the recordings of several seismic stations are available, the hypocenter of the event can be located using standard seismological tools, as well as other seismic source parameters which can be obtained by the fault plane solution. Typical VTs range in magnitude Mb between 1 and 3.

Rock-fall events

This event type is produced by the impact of falling rock fragments on the surface of the volcanic edifice, and therefore, is typical for dome-building eruptions involving highly viscous magma. There is a continuum between small rock avalanches to highly energetic, so-called pyroclastic

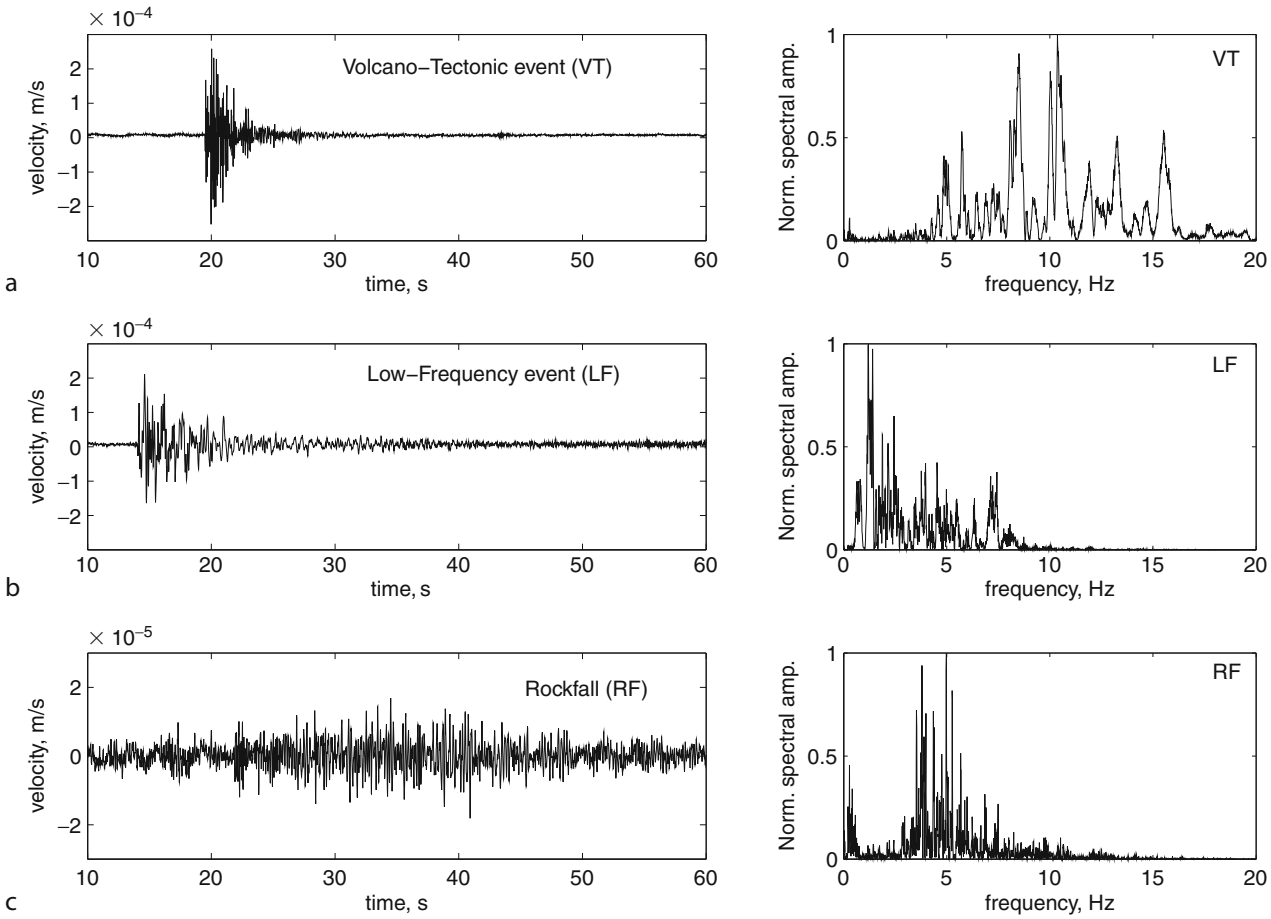
flows, which are a mixture of hot gas and rock fragments, ranging from huge boulders to fine volcanic ash particles. The impact that generates the seismic signal depends on rock size, impact angle, and the property of the surface.

The signal, a typical example of which is depicted in Figure 1c, has a "cigar-shaped" waveform, increasing in amplitude as more and more rocks form a rock avalanche, gaining speed and momentum. Secondary phases can be identified in the waveform as the flow reaches a change in topography slope converting its kinetic energy into a seismic signal. The frequency content is very high, containing mostly surface waves. The end of the "cigar" is reached once the material involved in the flow has settled. Rock-fall signals occur sometimes as secondary events in connection with low-frequency events that are able to cause significant seismic shaking of a dome structure (see below) or volcanic explosions resulting in a shower of volcanic ejecta and collapsing eruption columns. Another signal with a similar waveform is caused by so-called *lahars*, high-density mudflows that can keep big boulders in suspension. These mudflows occur often after heavy rainfall or through melting icecaps that cover parts of the volcano. While rock-fall events are seismologically the least interesting signals on a volcano, they can often be used as a proxy for magmatic extrusion rate, assuming that the lava disintegrates and falls off the dome at the same rate as it is extruded.

Low-frequency earthquakes

Low-frequency earthquakes, depicted in Figure 1b and in the following referred to as LF, proved to be pivotal to modern volcano seismology. Several examples have been quoted where these events have been successfully used to forecast volcanic eruptions (e.g., Chouet, 1996). Their occurrence alone has been used as an indicative tool even though a full understanding of their source mechanism had not been achieved. On Galeras volcano, Colombia, the occurrence of a few low-frequency earthquakes played a major role in the heated scientific and public debate regarding whether lives could have been saved had their occurrence been recognized as a warning sign for an imminent volcanic eruption (Gil Cruz and Chouet, 1997).

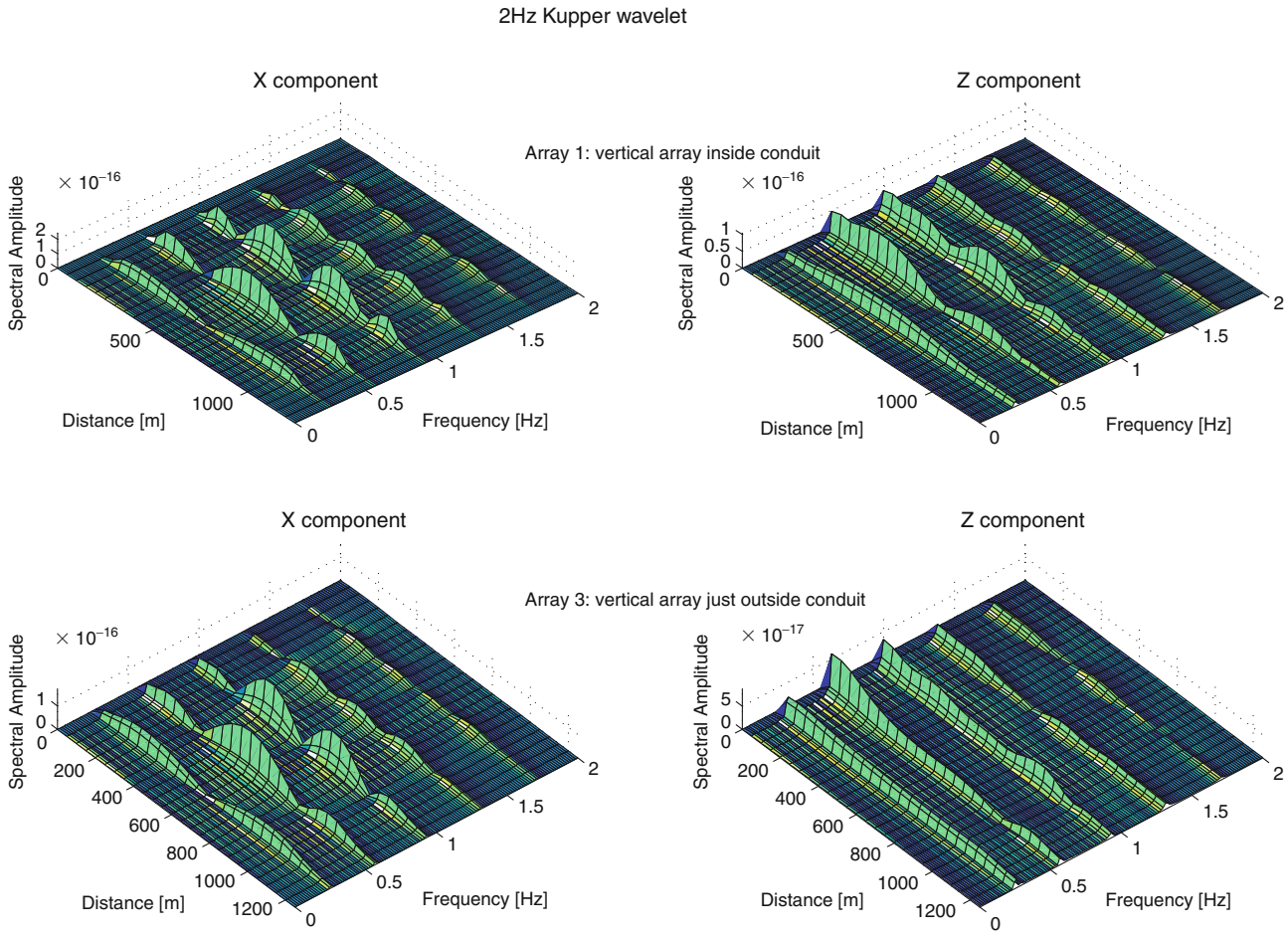
Low-frequency earthquakes occupy a spectral range of approximately 0.2–10 Hz and form a spectral



Earthquakes, Volcanogenic, Figure 1 Different event types in time and frequency domain, recorded on Soufriere Hills volcano on Montserrat; (**a**) volcano-tectonic earthquake (VT), (**b**) low-frequency earthquake, and (**c**) rock-fall event.

continuum between two end-members of seismic categories: *long-period events* and so-called *hybrid events*, which are similar to long-period events but carry in addition a high-frequency onset. On Soufriere Hills volcano in Montserrat, it has been demonstrated that the amount of high-frequency energy observed in the seismic signal can vary from station to station, and also at one station from event to event. Therefore, long-period and hybrid earthquakes have been grouped together under a more generic term as *low-frequency earthquakes* (Neuberg et al., 2000). These events have been observed on many, mostly dome-building volcanoes, and have been empirically associated with the pressurization of the volcanic system (Lahr et al., 1994; Neuberg et al., 1998).

Very different from tectonic events, low-frequency earthquakes originate from a boundary between a fluid, such as gas or magma contained in conduits, dykes, or cracks, and the solid surrounding rock (e.g., Ferrazzini and Aki, 1987; Chouet, 1988; Neuberg et al., 2000). Most of the seismic energy is trapped in the fluid-filled plumbing system – conduit, dyke or crack – and resonates within this system while only a small part can escape and propagate through the volcanic edifice and is recorded at the surface. Hence, low-frequency earthquakes are generated by resonating interface waves – hence the term *conduit resonance* – however, these interface waves must not be confused with acoustic waves in the fluid, often referred to as organ pipe modes. Acoustic and interface waves differ in one important aspect: Interface waves are dispersive; hence, the velocity varies with wavelength, and it is the slow moving part of the wave field that explains the low-frequency character (Ferrazzini and Aki, 1987). Figure 2 shows the horizontal and vertical seismic component of a synthetic, resonating interface wave for a two-dimensional dyke model. The single overtones, or modes, are separated by frequency in the depth–frequency plot. The low-frequency seismic wave can be considered as the superposition of these modes. The frequency content



Earthquakes, Volcanogenic, Figure 2 Eigen modes of a low-frequency earthquake source: spectral amplitude distribution for horizontal (X) and vertical (Z) component inside (*above*) and outside (*below*) the conduit for single frequencies. The superposition of the frequency contributions represents the source signal inside the conduit and outside. Note that the vertical component is affected by the interaction with the free surface, while the horizontal component is not.

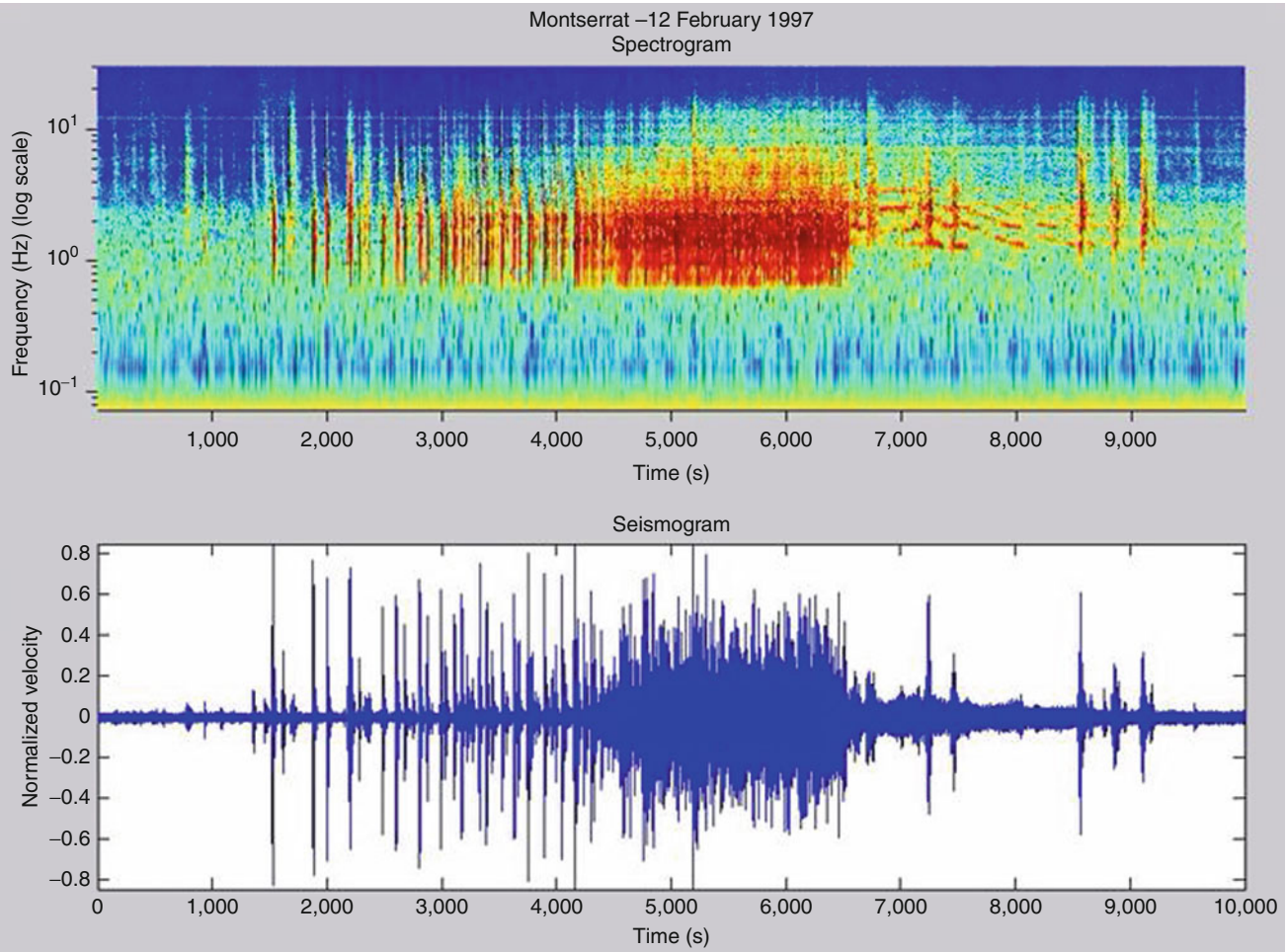
is controlled by the interface wave velocity which, in turn, is determined by the rigidity μ of the surrounding rock, the bulk modulus of the fluid κ, the wavelength of the interface wave, and the width of the fluid-filled container.

Low-frequency earthquakes are observed on many volcanoes, and most observations include one or several of the following characteristics: (1) LFs occur in swarms, (2) the waveforms are very similar, (3) their occurrence is highly regular, (4) with increasing occurrence, the LF swarms merge into continuous tremor, and most importantly, (5) major dome collapses or other eruptive phases are preceded by LF swarms (Miller et al., 1998; Aspinall et al., 1998; Hellweg, 2003). The highly periodic occurrence of LFs in swarms of similar waveform can be seismologically interpreted as a stationary seismic source with a repetitive non-destructible source mechanism (Green and Neuberg, 2006). The other exciting aspect is their potential as a forecasting tool for volcanic eruptions or dome collapses. Both Galeras volcano in Colombia and Soufriere Hills volcano in Montserrat are examples where LF swarms occurred before such events; however, the occurrence of the earthquakes does not guarantee a volcanic event. This demonstrates that the mere occurrence of LFs is not yet a reliable forecasting tool unless the actual trigger mechanism is understood and a direct link between the seismogram and source mechanism can be made.

Several trigger mechanisms for LFs have been proposed, including magma–water interaction (Zimanowski, 1998), stick-slip motion of magma plugs (Goto, 1999; Iverson et al., 2006), magma flow instabilities (Julian, 1994), periodic release of gas–ash mixtures into open cracks (Molina et al., 2004), and repeated, brittle fracturing of magma in glass-transition (Neuberg et al., 2006).

Volcanic tremor

As indicated in the previous section, one definition of volcanic tremor is the superposition of low-frequency earthquakes as they merge in time. Figure 3 depicts an



Earthquakes, Volcanogenic, Figure 3 Seismic low-frequency earthquake swarms merging into tremor: spectrogram (*upper part*) and seismogram (*lower part*). Note the *lines* in the spectrogram toward the end of the signal, representing a highly periodic repeated excitation of identical events. The spectral distance between the lines corresponds to the excitation frequency.

example in time and frequency domain from Soufriere Hills volcano in Montserrat where single LF earthquakes merge into tremor. If the highly periodic occurrence, or trigger mechanism, is maintained, the tremor signal exhibits a spectral signature of integer harmonic spectral lines (see Figure 3) if the time window from which the spectral slice is computed is long enough to cover several events. The distance between the spectral lines is inversely related to the time separation between LFs of the same waveform, origin, and time history. As the frequency of occurrence increases, the distance between spectral lines will increase as well. In this way, one can measure the inter-event time, or event rate, by using the separation of spectral lines, even though the single events have merged in time into a continuous tremor signal. For this type of volcanic tremor, it has been shown that the spectral signature of the tremor signal is identical to the low-frequency earthquakes of which it is composed (Latter, 1981).

Another definition of volcanic tremor is more general and refers to the background noise caused by magmatic processes. Figure 4 shows an example from Piton de la Fournaise, La Reunion, where volcanic tremor is produced by lava extruding from fissures. The onset and end of the tremor signal are remarkably sharp and coincide with the start and finish of the eruption. The spectrum of the signal is "white"; hence, it is composed of a large number of frequencies and does not have any spectral structure. The origin of the tremor is due to pressure fluctuations during turbulent magma flow (Julian, 1994). Another tremor source of this kind can be attributed to hydrothermal activity or degassing. While such a tremor signal does not have any distinct seismic phases (like *P* and *S* wave arrival) that could be used to determine the source location on the base of relative travel times, the relative amplitudes of the signal can be utilized to obtain an approximate location of the tremor source.

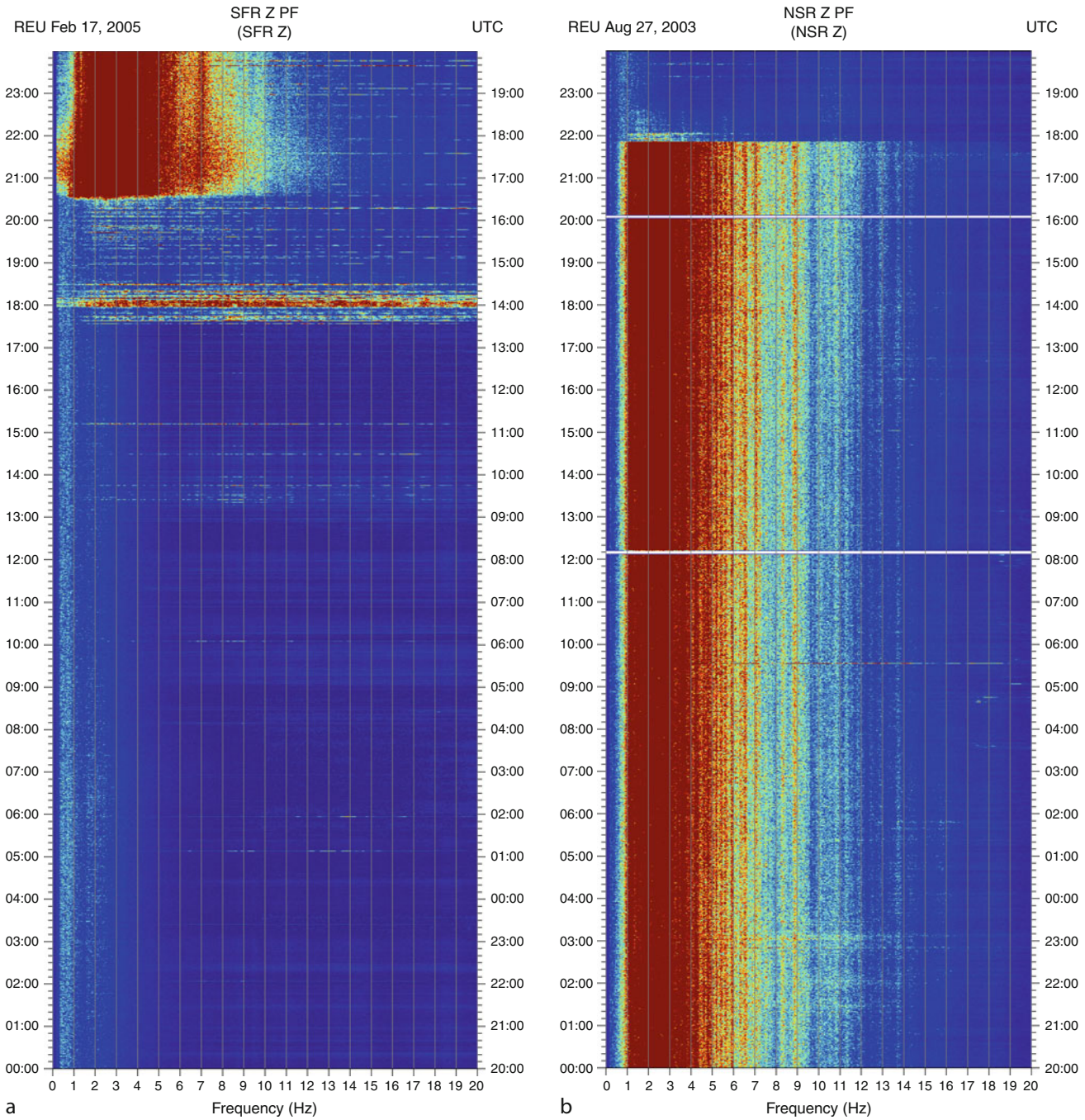
Very-long-period earthquakes

With the availability of broadband seismic sensors, seismology opened a new window to very-long-period signals caused by transients of pressure variations in the volcanic edifice with periods ranging from tens of seconds to several minutes (Neuberg et al., 1994; Wielandt and Forbriger, 1999). Measurements in this frequency range close the gap between classic seismology and ground deformation as determined by strain-, tilt-meters, and GPS. However, while ground deformation can be measured directly as ground *displacement*, a seismometer detects the ground *velocity*, or the temporal change of the displacement, due to its electromagnetic measurement principle, reacting to the *change* of the magnetic flux by inducing an electric current. The advantage of such a system lies in the high sensitivity of the detection system; the disadvantage is that several data processing steps are necessary to retrieve the ground displacement, if this is the quantity required for modeling or further interpretation. For short-period signals, this does not pose any problem as the instrument response is constant in this narrow (high-) frequency band; hence, the seismogram has only to be integrated to convert ground velocity into displacement. However, for a broadband seismometer, the instrument response is highly frequency dependent and lower frequencies are treated by the seismometer in a different way than high frequencies. In order to retrieve the ground displacement from the measured ground velocity, the effect of the instrument response has to be taken into account and, if possible, to be removed. This is in addition to the integration step in the data processing sequence.

Modern broadband seismometers have a "flat" frequency response typically in the range between 50 and 0.008 Hz, expressed in periods, between 0.02 and 120 s. Figure 5 shows the effect of the removal of the instrument response in an example from Stromboli volcano. The seismic record of a Strombolian eruption shows an ultra-long-period transient in the vertical direction, caused by an ascending magma/gas plug with an eruptive, high-frequency seismic signature superimposed. In order to represent the long-period part of the signal properly, the original velocity seismogram has to be corrected for the instrument response as well as integrated. This processing sequence is shown in Figure 5, where the velocity seismogram is converted to displacement by integration (Figure 5a), and where finally, after the instrument response has been successfully removed, the seismogram (Figure 5d) represents the true ground displacement at the location of the seismic sensor. The comparison between corrected and uncorrected displacement demonstrates the importance of this processing steps. An interpretation or model of the uncorrected trace (Figure 5a) would comprise an upward movement of the volcanic edifice prior to the Strombolian eruption followed by a deflation and slow recovery of the edifice. The corrected trace offers a much simpler interpretation (Neuberg and Luckett, 1996): The edifice inflates while the magma plug is ascending, and once the gas is released through, the eruption deflates to its original level. This example demonstrates that each seismic sensor, and indeed each recording instrument, has to be considered a *filter* that affects and modifies the original physical signal in a certain way, often introducing a frequency-dependent change in amplitude and phase. To correct for these effects is absolutely essential when dealing with ultra-long-period seismic signals.

Volcano seismology in a wider volcanological context

Volcanogenic earthquakes have been the first signals to be used to monitor and assess volcanic activity, and are today still at the core of any volcano monitoring program. However, if combined with other observations and integrated into wider conceptual and numerical volcanological models, their potential as a forecasting tool increases even further. In the following section, we give a few examples

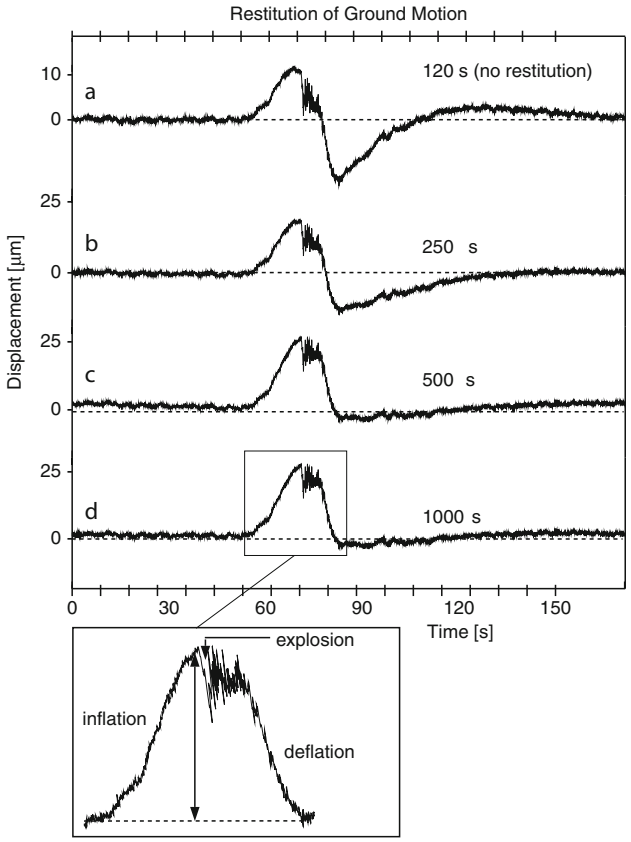


Earthquakes, Volcanogenic, Figure 4 Spectrograms of continuous tremor at Piton de la Fournaise, La Reunion. (**a**) Onset of tremor is preceded by a short swarm of VT events, (**b**) continuous tremor during lava extrusion, note the extremely sharp end of the tremor signal when magma extrusion stops.

how volcanogenic earthquakes have been used to assess several aspects of volcanic activity.

Estimation of magma ascent rate

A very exciting example from Montserrat has been reported by Voight et al. (1998), where a radial tilt signal recorded on Chances Peak, an old dome structure adjacent to the active lava dome of Soufriere Hills volcano, correlated perfectly with the cyclic occurrence of seismic swarms. Neuberg et al. (2006) reinvestigated this correlation and found that seismic activity, in this case, low-frequency earthquake swarms, commences when the tilt



Earthquakes, Volcanogenic, Figure 5 Very-long-period signal from Stromboli volcano after removal of instrument response. (**a**) through (**d**) indicate the period limit to which the spectral energy has been included in the restitution. While the trace in (**a**) contains only spectral components up to 120 s period, the true ground displacement is reconstructed by including all spectral components to a period of 500 s (**c**) or even 1,000 s (**d**). The final result in the lowermost trace shows the ground displacement associated with the rising gas slug, the Strombolian explosion as well as the deflation after the explosion.

signal goes through an inflection point (Figure 6). Furthermore, low-frequency seismicity ceases when the tilt signal goes through a second turning point. The tilt, which is the rotation of the volcano's flank, is caused by a combination of a pressure buildup in a shallow magma reservoir and the shear stress acting across the conduit wall by the motion of viscous magma at shallow depth (Green et al., 2006). Hence, the cyclic occurrence of LF swarms can be directly related to the magma ascent rate: As the pressure in the magma has reached a certain threshold, magma starts to move causing low-frequency seismicity by either a stick-slip motion (Iverson et al., 2006) or by magma rupture (Neuberg et al., 2006). This causes the turning point in the tilt behavior; however, the ground deformation (tilt) still increases to its maximum where the magma ascent rate peaks. At this point, the magma ascent slows down as gas bubble diffusion cannot keep up with the pressure drop, caused by the magma movement. This can be clearly seen in both the tilt signal and the seismicity rate depicted in the inset of Figure 6. At the second turning point, magma ascent rate has become so low or even stopped completely that no further LF seismicity is produced. This gives gas diffusion a chance to catch up, leading to a renewed pressurization of the system again, and a new cycle has started.

This example demonstrates how LF seismicity can be interpreted following a conceptual model that allows to estimate the changes in magma ascent rates on the base of seismicity, even if no tilt signal is available. Using such a model, Hammer and Neuberg (2009) found an accelerating seismicity rate in consecutive LF swarms on Montserrat – as a proxy for an accelerated magma ascent – which led to a dome collapse in 1997.

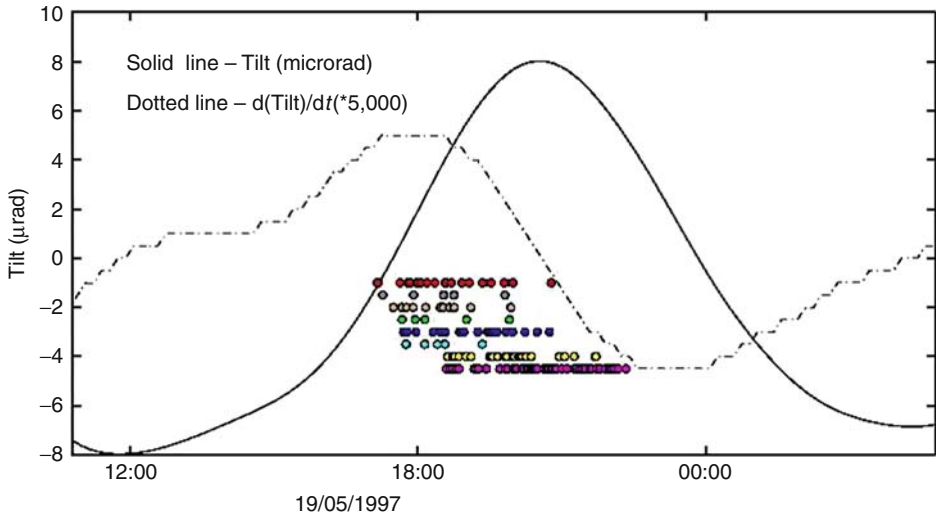
Combining seismicity with infrasonic signals

A second example of a combined interpretation uses infrasonic acoustic signals with seismicity. The monitoring of infrasonic signals on active volcanoes has become routine in the last 10 years and turned out extremely helpful for open volcanic systems.

See Johnson (2003) for an overview on infrasonic monitoring techniques. Compared to seismic waves, infrasonic signals are less affected by the medium (steam and air) through which they propagate, and the relative timing between the two types of signals can be used to dissect a complex volcanic event such as a dome collapse, that can last for several hours and exhibit distinct phases. Green and Neuberg (2005) modeled a partial dome collapse at Montserrat in 2004 and used the relative timing between seismic and infrasonic pulses to separate several distinct explosions during the collapse. Furthermore, they identified a plume forming degassing event as the underlying cause of the collapse. A further application of a combined interpretation can be found in Ripepe et al. (2002) where relative timing between infrasound and seismic signal revealed the relative height of a magma column on Stromboli volcano from where an explosion occurred.

Detection of lahars and pyroclastic flows

Pyroclastic flows, lava flows, and lahars often escape direct visual observation, but the fast determination of their location can be crucial in terms of risk mitigation. As pointed out above, the seismic signals produced by such events lack any clear seismic phase that could be used for a classic travel-time-based location procedure. Jolly et al. (2002) devised a method based on a relative seismic amplitude estimation across a seismic monitoring network, that allowed to follow the path of a flow in near real-time. The basic idea is to determine the location of the seismic surface source by assuming that the amplitudes of the surface waves caused by the flow decay by the inverse of the distance to a seismic station and by inelastic attenuation. They performed a so-called grid search inversion



Earthquakes, Volcanogenic, Figure 6 Correlation between tilt and seismicity on Soufriere Hills volcano, Montserrat. The low-frequency seismicity is sorted in different event "families." The different colors indicate different families of similar seismic waveform which implies that each family has a common source mechanism and origin. It commences and ceases at the inflection points of the tilt signal (*solid line*), equivalent to the maximum and minimum of its time derivative (*dotted line*).

across the volcanic edifice to locate the flow. Through this technique, they were able to resolve fine details of the flow propagation such as a "piling-effect" that occurred when the front of the pyroclastic flow came to a halt while the rest of the flow piled up behind it, causing the seismic source travelling backup-hill!

On Piton de la Fournaise, a similar method is used to detect the source of basaltic lava flows, which are responsible for the generation of a tremor signal. In this way, the observatory staff is able to inform the authorities in near real-time about the source location of the lava flow.

Detecting stress changes in volcanic settings

Recent advances have made it possible to detect temporal variations in properties of seismic signals that are caused by stress changes in the medium that they penetrate. These changes can be seen in directional variations of the wave speed (also referred to as anisotropy), energy loss, and seismic source parameters (fault plane solution) of volcano-tectonic events.

Gerst and Savage (2004) detected a change in anisotropy at Mt Ruapehu, New Zealand, which coincided with the eruptive activity in 1995 and 1996. They interpreted these changes to be caused by changes in the stress due to magmatic activity that altered the orientation and volume of cracks in the volcanic edifice. Using newly developed noise correlation techniques, Brenguier et al. (2008) observed at Piton de la Fournaise, La Reunion, similar changes in seismic velocity caused by the pressurization of the volcanic edifice due to ascending magma. Roman et al. (2006) found systematic variations in the orientation of the main pressure axis, which was derived from fault plane solutions of volcano-tectonic events on Soufriere Hills volcano on Montserrat. They identified distinct episodes in the eruptive history of Soufriere Hills volcano, where the regional stress field was superimposed by local stress perturbations that caused the pressure axis determined from VT fault plane solutions to swing around by 90°. Again, these stress changes were interpreted as caused by ascending magma, and indeed, these episodes preceded in some cases the restart of dome growth and magma extrusion. While the methods mentioned in this section do not necessarily utilize only volcanogenic earthquakes, they employ seismological tools to monitor the stress perturbations caused by volcanic activity, and therefore, could lead to new methods of volcanic forecasting.

Summary and conclusions

Monitoring and interpretation of volcanogenic earthquakes are core tasks of any volcano observatory. The separation into different event types allows us to attribute seismic activity to certain volcanic processes. While some event types are directly caused by brittle failure during magma ascent, or by extrusion of magma at the surface, other seismic signals provide us with indirect evidence of changes caused by the volcanic-magmatic system. Volcano seismology emerged from classic earthquake seismology as a separate discipline focusing on the special character of seismo-volcanic source processes. This development is ongoing, and classic seismic techniques such as the determination of seismic moment tensors will be applied in the future to the wider variety of seismo-volcanic sources which can be significantly different from usual earthquake sources acting on a single fault plane. With the availability of seismic broadband sensors, volcano seismology overlaps with high-frequency acoustic

emissions on one side of the seismic spectrum and with deformation signals on the low-frequency end. Multidisciplinary approaches combining seismology with other signals recorded on volcanoes will lead to more detailed insights into volcanic processes.

Bibliography

Aspinall, W., Miller, A., Lynch, L., Latchman, J., Stewart, R., White, R., and Power, J. A., 1998. Soufriere Hills eruption, Montserrat, 1995–1997: volcanic earthquake locations and fault plane solutions. *Geophysical Research Letters*, **25**, 3397–3400.

Battaglia, J., Ferrazzini, V., Staudacher, T., Aki, K., and Cheminee, J.-L., 2005. Pre-eruptive migration of earthquakes at the Piton de la Fournaise volcano (Reunion island). *Geophysical Journal International*, **161**, 549–558.

Brenguier, F., Shapiro, N. M., Campillo, M., Ferrazzini, V., Duputel, Z., Coutant, O., and Nercessian, A., 2008. Towards forecasting volcanic eruptions using seismic noise. *Nature Geoscience*, **1**, 126–130.

Chouet, B., 1988. Resonance of a fluid-filled crack: radiation properties and implications for the source of long-period events and harmonic tremor. *Journal of Geophysical Research*, **101**, 4375–4400.

Chouet, B., 1996. Long-period volcano seismicity: its source and use in eruption forecasting. *Nature*, **380**, 309–316.

Ferrazzini, V., and Aki, K., 1987. Slow waves trapped in a fluid-filled crack: implications for volcanic tremor. *Journal of Geophysical Research*, **92**, 9215–9223.

Gerst, A., and Savage, M. K., 2004. Seismic anisotropy beneath RuapehuVolcano: a possible eruption forecasting tool. *Science*, **306**, 1543–1547.

Gil Cruz, F., and Chouet, B., 1997. Long-period events, the most characteristic seismicity accompanying the emplacement and extrusion of a lava dome in Galeras Volcano, Columbia, in 1991. *Journal of Volcanology and Geothermal Research*, **77**, 121–158.

Goto, A., 1999. A new model for volcanic earthquakes at Unzen Volcano: melt rupture model. *Geophysical Research Letters*, **26**, 2541–2544.

Green, D. N., and Neuberg, J., 2005. Seismic and infrasonic signals associated with an unusual collapse event at the Soufriere Hills volcano Montserrat. *Geophysical Research Letters*, **32**, L07308.

Green, D., and Neuberg, J., 2006. Waveform classification of volcanic low-frequency swarms and its implication. *Journal of Volcanology and Geothermal Research*, **153**, 51–63.

Green, D. N., Neuberg, J., and Cayol, V., 2006. Shear stress along the conduit wall as a plausible source of tilt at Soufriere Hills volcano Montserrat. *Geophysical Research Letters*, **33**, L10306.

Hammer, C., and Neuberg, J., 2009. On the dynamical behaviour of low-frequency earthquake swarms prior to a dome collapse of Soufriere Hill volcano Montserrat. *Geophysical Research Letters*, **36**, L06305.

Hellweg, M., 2003. The polarization of volcanic seismic signals: medium or source? *Journal of Volcanology and Geothermal Research*, **128**, 159–176.

Iverson, R. M., Dzurisin, D., Gardner, C. A., Gerlach, T. M., LaHusen, R. G., Lisowski, M., Major, J. J., Malone, S. D., Messerich, J. A., Moran, S. C., Pallister, J. S., Qamar, A. I., Schilling, S. P., and Vallance, J. W., 2006. Dynamics of seismogenic volcanic extrusion at Mount St Helens in 2004–05. *Nature*, **444**, 439–443.

Johnson, J. B., 2003. Generation and propagation of infrasonic air waves cc from volcanic explosions. *Journal of Volcanology and Geothermal Research*, **121**, 1–14.

Jolly, A. D., Thompson, G., and Norton, G. E., 2002. Locating pyroclastic flows on Soufriere Hills Volcano, Montserrat, West Indies, using amplitude signals from high dynamic range instruments. *Journal of Volcanology and Geothermal Research*, **118**, 299–317.

Julian, B., 1994. Volcanic tremor: nonlinear excitation by fluid flow. *Journal of Geophysical Research*, **99**, 11859–11877.

Lahr, J. C., Chouet, B. A., Stephens, C. D., Power, J. A., and Page, R. A., 1994. Earthquake classification, location and error analysis in a volcanic environment. *Journal of Volcanology and Geothermal Research*, **62**, 137–151.

Latter, J. H., 1981. Volcanic earthquakes and their relationship to eruptions at Ruapehuand Ngauruhoe volcanoes. *Journal of Volcanology and Geothermal Research*, **9**, 293–309.

Miller, A., Stewart, R., White, R., Luckett, R., Baptie, B., Aspinall, W., Latchman, J., Lynch, L., and Voight, B., 1998. Seismicity associated with dome growth and collapse, at the Soufriere Hills volcano Montserrat. *Geophysical Research Letters*, **25**, 3401–3404.

Molina, I., Kumagai, H., and Yepes, H., 2004. Resonances of a volcanic conduit triggered by repetitive injections of ash-laden gas. *Geophysical Research Letters*, **31**, L03603.

Neuberg, J., and Luckett, R., 1996. Seismo-volcanic sources on Stromboli volcano. *Annali Geofisika*, **39**, 377–391.

Neuberg, J., Luckett, R., Ripepe, M., and Braun, T., 1994. Highlights from a seismic broadband array on Stromboli volcano. *Geophysical Research Letters*, **21**, 749–752.

Neuberg, J., Baptie, B., Luckett, R., and Stewart, R., 1998. Results from the broad-band seismic network on Montserrat. *Geophysical Research Letters*, **25**, 3661–3664.

Neuberg, J., Luckett, R., Baptie, B., and Olson, K. B., 2000. Models for tremor and low-frequency events on Montserrat. *Journal of Volcanology and Geothermal Research*, **101**, 83–104.

Neuberg, J. W., Tuffen, H., Collier, L., Green, D., Powell, T., and Dingwell, D., 2006. The trigger mechanism of low-frequency earthquakes on Montserrat. *Journal of Volcanology and Geothermal Research*, **153**, 37–50.

Ripepe, M., Harris, A. J. L., and Carniel, R., 2002. Thermal, seismic and infrasonic evidences of variable degassing rates at Stromboli volcano. *Journal of Volcanology and Geothermal Research*, **118**, 285–297.

Roman, D. C., Neuberg, J., and Luckett, R., 2006. Assessing the likelihood of volcanic eruption through analysis of volcanotectonic earthquake fault-plane solutions. *Earth and Planetary Science Letters*, **248**, 229–237.

Voight, B., Hoblitt, R., Clarke, A., Lockhart, A., Miller, A., Lynch, L., and McMahon, J., 1998. Remarkable cyclic ground deformation monitored in real time on Montserrat, and its use in eruption forecasting. *Geophysical Research Letters*, **25**, 3405–3408.

Wielandt, E., and Forbriger, T., 1999. Near-field seismic displacement and tilt associated with the explosive activity of Stromboli. *Annali Geofisika*, **42**, 407–416.

Zimanowski, B., 1998. Phreato-magmatic explosions. In Freundt, A., and Rosi, M. (eds.), *From Magma to Tephra*. Amsterdam: Elsevier, pp. 25–53.

Cross-references

Earthquake, Focal Mechanism
Earthquake, Location Techniques
Earthquakes, Source Theory
GPS, Tectonic Geodesy
Propagation of Elastic Waves: Fundamentals
Seismic Anisotropy
Seismic Instrumentation
Seismological Networks

ELECTRICAL PROPERTIES OF ROCKS

Takashi Yoshino
Institute for Study of the Earth's Interior, Okayama University, Misasa, Tottori, Japan

Definition

Electrical conduction. The movement of electrons or electrically charged particles through a medium.
Electrolyte. Any substance containing free ions that make the substance electrically conductive. The most typical electrolyte is an ionic solution.
Fermi level. The highest energy of occupied states in a solid at 0 K. The Fermi level in conductors lies in the conduction band, in insulators it lies in the valence band, and in semiconductors it falls in the gap between the conduction band and the valence band.
Stern layer. One of two electrically charged layers of electrolyte ions; the layer of ions immediately adjacent to the surface, and the layer in the neighborhood of a negatively charged surface.

Introduction

The interpretation of electromagnetic data obtained from the electromagnetic (EM) induction method requires knowledge of electrical properties of rocks comprising the Earth's interior. The electrical properties of rocks are widely used in electrical conductivity (resistivity) and induced polarization methods to identify zones of ore deposits and to probe the thermal and compositional structure of the Earth's interior. For example, electrical resistivity measurements were first used in the 1960s to prospect for sulfide ore deposits. Furthermore, the use of EM induction techniques to determine the electrical conductivity of the crust and mantle can provide useful information regarding the processes in the Earth's interior independent from seismological studies. This results from the fact that electrical conductivity is sensitive to small amounts of accessory phases, such as fluid, melt and metal in rocks, and small amounts of hydrogen and ferric irons in rock-forming minerals.

The bulk electrical properties of rocks depend on complex combinations of various electrical components consisting of solids (silicates, oxides, and metal), pores filled with air or liquid, and interfaces between solids and liquids. Thus, rocks are composites of a wide variety of minerals occasionally containing fluids and molten material. Since the 1960s, many studies have investigated the electrical properties of rocks and rock-forming minerals as a function of frequency, temperature, pressure, oxygen fugacity, water content, applied electric field, and other variables. Recent developments of analytical techniques and improvements in the resolution of electromagnetic observations have led to an increase in interest with regard to the electrical properties of rocks at depth within the Earth. The high pressures and temperatures within the Earth's interior are difficult to replicate in the laboratory, but experimental results are required to predict the electrical properties in these environments. Within the last two decades, many efforts have been made to measure physical properties of high-pressure phases to obtain knowledge of temperature, composition, and water content in the Earth's mantle. As a result, the maximum range of pressure where the electric conductance can be measured currently covers the pressure conditions at the core–mantle boundary (2,900 km depth).

Definition of electrical conductivity (resistivity)

When an electric potential difference V is applied across a material, the electrical current I in the material is proportional to the voltage across it according to Ohm's law, $V = IR$, where R is the electrical resistance, which depends on the intrinsic resistivity ρ of the material and on the geometry (length l and area A through which the current passes):

$$R = \rho l / A \tag{1}$$

The electrical conductivity σ is the inverse of the resistivity. The current density J can be expressed as $J = E/\rho = E\sigma$. Conductivity has one of the widest ranges for material property, from 10^{7} S/m (metals) to 10^{-20} S/m (good electrical insulators). Semiconductors have conductivities in the range of 10^{-6} to 10^{4} S/m.

Electrical conduction

An electrolytic process can be considered as a dominant mechanism for electron charge transport in shallower crustal rocks, which contain pores and cracks filled with electrolyte fluids. At high saturation conditions, conduction currents in electrolytes primarily arise from the flow of ions in the electric field. The conductivity of an electrolyte is controlled by the solute concentration, ionic charge, and other physical parameters, such as temperature and pressure. Fluid saturated rocks contain interfaces between the minerals and the fluid leading to surface conductivity, which can be considered as an additional fluid electrical conductivity in the vicinity of the charged surface. The concentration of ions increases with increasing proximity to the charged surfaces as a result of attractive electrostatic forces, which are induced by the surface charge. This high concentration layer is called the electrochemical double layer (DL), and it is composed of the Stern layer in contact with solids, and a diffusive layer in contact with fluids. Higher concentrations of ions in the Stern layer lead to higher conductivity. The significant influence of surface conductivity on bulk electrical properties appears to be the same for porous samples saturated with dilute electrolytes and low porosity samples where the thickness of the DL is the same as the width of a crack. A mixed conduction process of electrolyte and surface conductivity is most likely to appear in low saturation conditions.

As pressure and temperature increase, the electrical conduction in rock-forming minerals is likely to control

the bulk electrical conductivity of rocks. This is achieved by a decrease in the effect of rock open space and interface effects. Solids can be classified in terms of their conductivity as conductors, semiconductors, or insulators. The classification for solid materials can be visualized in terms of band theory, and is strongly correlated with the interatomic spacing in the solid. The electron energy level in a solid is often expressed in relation to the Fermi energy. In metals, the electrons occupy states up to the Fermi level. Conduction occurs by promoting electrons into the conduction band that starts at the Fermi level. The promotion energy is negligibly small so that at any temperature, electrons can be found in the conduction band. The number of electrons participating in electrical conduction is extremely low. In semiconductors and insulators, the valence band is filled, and no more electrons can be added, following Pauli's principle. In insulators, there is a large energy gap between the valence and conduction bands, so that a large quantity of energy is needed to promote an electron to the conduction band. Semiconductors can be divided into intrinsic or extrinsic conductors. Intrinsic conduction means that electrical conductivity is pure, and does not depend on impurities. In extrinsic semiconductors, the conductivity depends on the concentration of impurities. Electronic conduction through extrinsic semiconductors takes place through electrons and electron holes. The electrical conductivity of semiconductors can be enhanced by doping to produce n (negative)-type or p (positive)-type semiconductors.

Silicates and oxides usually are insulators at ordinary temperatures. Since the large energy gap of an insulator could be overcome by heat, the silicate minerals can behave as a semiconductor at high temperatures corresponding to conditions of the Earth's interior. Mobility of lattice defects and a small amount of impurity, such as hydrogen and ferric iron, enhances electrical conduction of silicate minerals at high temperatures. The conductivity of ionic-bonded crystals is proportional to the number and mobility of charge carriers. It is known from the Nernst–Einstein equation that electrical conductivity σ depends on the number N of electric charge carriers per unit volume:

$$\sigma = Nze\mu \tag{2}$$

where z is the charge number, e is the charge of electron and μ is mobility.

The dependence of conductivity on temperature is similar to other thermally activated processes:

$$\sigma = A\exp(-\Delta H/2kT) \tag{3}$$

where A is a pre-exponential constant, ΔH is an activation enthalpy, k is the Boltzmann constant and T is the absolute temperature. Plotting log normal σ versus reciprocal temperature ($1/T$) produces a straight-line slope ($H/2\ k$) from which the band gap energy can be determined. In addition to this dependence, extrinsic semiconductors rely on the thermal promotion of electrons from donor levels or holes to acceptor levels. The dependence on temperature also is exponential, but it eventually saturates at high temperatures, where all the donors are emptied or all the acceptors are filled. This means that at low temperatures, extrinsic semiconductors have higher conductivities than intrinsic semiconductors. At high temperatures, both impurity levels and valence electrons are ionized, but since the impurities are very low in number and they are exhausted, eventually the behavior is dominated by the intrinsic type of conductivity. The activation enthalpy (H) also can be defined as a function of pressure (P) as follows:

$$\Delta H = \Delta E + P\Delta V \tag{4}$$

where ΔE is the activation energy and ΔV is the activation volume.

Ionic (intrinsic) conduction can occur at high temperatures through the creation of cation vacancies. In the ferromagnesian mantle minerals (silicates and oxides), the charge carriers are generally vacancies in magnesium sites. The activation enthalpy of ionic conduction usually is higher than 2 eV. For this reason, extremely high temperatures are needed to observe the ionic conduction. Proton conduction is a type of ionic conduction. Because the proton is an extremely small ion, but massive compared with an electron, the proton is the only ion which is expected to be highly mobile in solids. The energy barrier for the transportation of charge carriers for proton conduction is relatively low because of its small bonding energy and small ionic radius. The activation enthalpy of proton conduction usually is less than 1 eV, and tends to decrease with an increasing number of hydrogen atoms in minerals at very low concentrations. The proton conduction is a dominant conduction mechanism at low temperatures.

The most dominant conduction mechanism for the ferromagnesian minerals, such as olivine, is hopping (small polaron) conduction. Hopping conduction occurs through charge transfer between neighboring ions of different valences. A ferrous ion (Fe^{2+}) generally substitutes an Mg ion in an Mg site. However, a certain proportion of ferric ions, which have an electron hole, depend on oxygen fugacity, temperature, pressure, and crystal structure. Transfer of an electron hole between ferrous and ferric iron sites carries an electric charge, which is the hopping conduction. If only electron holes migrate in crystals, the energy barriers for the migration would be low. However, presence of an electric hole significantly affects the local structure of the ionic crystals. The presence of ferric iron generates an extra positive charge, which repulses cations and attracts anions. This complex of the local strains is called "small polaron." The migration of electron holes is associated with migration of small polarons. A relatively large activation enthalpy is needed to move a small polaron. The hopping conduction of common ferromagnesian silicates has relatively large activation enthalpy (>1 eV). The activation enthalpy for hopping conduction decreases with increasing iron concentration. The conductivity of ferromagnesian minerals also is

controlled by the oxidation state of iron. A variation of the oxygen fugacity from an oxidizing to a reducing environment decreases the conductivity, due to the reduction of Fe^{3+} to Fe^{2+}. Moreover, at the same oxygen buffer, the conductivity tends to increase with increasing iron content in minerals.

Dielectric behavior

A dielectric is an electrical insulator that can be made to exhibit an electric dipole structure. When a dielectric is placed in an electric field, electric charges do not flow through the material, but only slightly shift from their average equilibrium positions, causing dielectric polarization. In this instant, positive charges are displaced along the field and negative charges shift in the opposite direction. This creates an internal electric field, which partly compensates the external field inside the dielectric. The term dielectric is commonly used for the electrically insulating material between the metallic plates of a capacitor. When two parallel plates of area A, separated by a small distance l, are charged by $+Q$, $-Q$, an electric field develops between the plates:

$$E = Q/A\varepsilon\varepsilon_0 \tag{5}$$

where ε_0 is called the vacuum permittivity and ε the relative permittivity, or dielectric constant ($\varepsilon = 1$ for vacuum). In terms of the voltage between the plates, $V = El = Q/C$. The constant $C = A\varepsilon\varepsilon_0/l$ is called the capacitance of the plates.

Polarization of dielectric materials in an electric field can be categorized by three types: electronic, ionic, and orientation polarizations. Since electrons have much smaller mass than ions, they respond more rapidly to a changing electric field. For electric fields that oscillate at very high frequencies, electronic polarization can occur. At smaller frequencies, the relative displacement of positive and negative ions also can take place. Orientation of permanent dipoles, which require the rotation of a molecule, can occur only if the oscillation is relatively slow (MHz range or slower). The time needed by the specific polarization to take place is called the relaxation time.

Most rock-forming minerals have characteristics of capacitors requiring dielectrics of high ε that can function at high frequencies (small relaxation times). Some ore minerals are ferroelectric minerals exhibiting permanent polarization in the absence of an electric field. This is due to the asymmetric location of positive and negative charges within the unit cell. Two possible arrangements of this asymmetry result in two distinct polarizations. In a piezoelectric material, like quartz, an applied mechanical stress causes electric polarization by the relative displacement of anions and cations.

Impedance spectroscopy

Electrical properties of rocks have strong frequency dependence. Impedance spectroscopy (IS) is a very useful and versatile electrochemical tool to characterize electrical properties of any material and its interfaces. The basis of IS is the analysis of the impedance (resistance of alternating current) of the observed system with regard to the applied frequency and exciting signal. This analysis provides quantitative information about the conductance, the dielectric coefficient, the static properties of the interfaces of a system, and its dynamic change due to adsorption or charge-transfer phenomena. IS usually uses alternating current with low amplitude, and measurements can be made with cells with two electrodes on the sample faces.

A common way to describe the complex impedance is to use the R (resistor)–C (capacitor) circuit elements in series and parallel, indicating a relaxation process. When the system has a distribution of relaxation time differing in time constants, the system represents a combination of R–C circuits (Cole and Cole, 1941). The R–C parallel circuit generally has been used to determine the electrical conductivity of minerals, based on the assumption that the system has a single relaxation process. In the complex impedance plane, the magnitude and direction of a planar vector can be defined by an impedance:

$$Z(\omega) = Z' + jZ'' \tag{6}$$

where $Z(\omega)$ is impedance as a function of frequency for alternating current, and Z' and Z'' are real and imaginary parts of impedance, respectively. The phase angle (θ) in a complex impedance plane can be obtained from $\theta = \tan^{-1}(Z''/Z')$. When the phase angle is 0°, the sample resistance for the R–C parallel circuit equals the real part of impedance.

Various frequency dispersive polarization processes should be accounted for in mixed media, such as rocks. Roberts and Tyburczy (1991) observed three impedance arcs in the complex impedance plane from the frequency-dependent electrical measurements (10^{-4} to 10^4 Hz) of polycrystalline olivine compacts at atmospheric pressure. The authors interpreted these impedance arcs as grain interior (first arc), grain boundary (second arc) and electrode reaction mechanism (third arc) from high to low frequency, respectively. At high pressures, pores, cracks, or grain boundaries in a sample are tightly closed. Under these conditions, the conduction along grain boundaries disappears in the impedance spectra. If electrodes reacted with a sample, the electrode reaction part may appear on the lower frequency side. The next step is to consider the complex impedance spectra for the system with a fluid/melt phase. In the absence of an interconnected fluid/melt, the impedance spectra contain three components similar to the polycrystalline process. If an interconnected fluid/melt phase does exist, it forms an electrical pathway in parallel with the solid matrix. The resultant impedance spectra show only the single arc if the electrode reaction is negligible.

Electrical conductivity of liquid-bearing rocks

The effective conductivity of liquid-bearing rocks is very sensitive to the pore geometry in rocks. The enormous

contrast in conductivity between silicates and liquids implies that electrical currents flow almost totally through the pore liquid within rocks. Archie (1942) proposed an empirical equation that relates the bulk conductivity σ_r to the conductivity of a pore saturating fluid σ_f. The rocks were assumed to exhibit interconnected porosity, and no disturbance due to the presence of reactive minerals was considered. The ratio σ_f/σ_r is known as the formation factor F. Archie's law forms the basis for resistivity log interpretation:

$$F = \frac{\sigma_f}{\sigma_r} = C\phi^{-n}, \tag{7}$$

where C and n are constants. Archie's law is not a geometrical model. There are various theoretical models for estimating the bulk conductivity of liquid-bearing rocks from the conductivity of liquid and solid phases. A frequently used model to predict the maximum effective bulk conductivity of an interconnected two-phase mixture is the Hashin and Shtrikman (HS) upper bound (Hashin and Shtrikman, 1962). This upper bound is representative for liquid phases distributed along grain boundaries and filling triple junctions of spherical grains. The opposite case representing isolated spheres of liquids in solid grains is known as the HS lower bound.

Electrical conductivity of crustal rocks

The shallower zone of the upper crust consists of a sequence of sedimentary rocks along with fractured crystalline rocks. Because these rocks contain pores and cracks that are filled with fluid electrolytes, they are moderately good conductors of electricity. As well as electrolyte conductivity, surface conductivity caused by electrochemical reactions between the fluid electrolyte and the mineral surface contribute additionally to the bulk conductivity. This zone has conductivities varying from about $10^{-0.5}$ S/m in recent sediments to 10^{-3} S/m or more in weathered crystalline rocks. As pressure increases, the rock conductivity decreases due to the reduction of porosity. Remarkable changes in electrical properties often occur at mid-crustal depths, probably caused by the development of incipient fractures. If the fractures are filled with fluid electrolytes, electrical conductivity would largely be enhanced compared to the surrounding dry crystalline rocks. Secular changes in electrical conductivity might be used in predicting earthquakes (Varotsos and Alexopoulos, 1984).

Lower crustal rocks mainly consist of crystalline igneous or metamorphic rocks, which are thought to be much denser, and which have little pore space for water collection. Since the main rock-forming minerals are highly insulating at crustal temperatures, electrical conductivity in such rocks should be expected to have lower values than the upper crust. However, magnetotelluric (MT) and deep geomagnetic soundings (GDS) have revealed an unexpectedly high conductivity anomaly in the lower crust (Shankland and Ander, 1983). A very small portion of well-conducting material may cause enhanced conductivity of the whole rock mass, provided that the material fills interconnected pore space, or covers grain surfaces. To generate a conductive layer, an indispensable interconnectivity of the well-conducting phase must be established and maintained over large distances under deep crustal conditions.

There are two popular candidates (water and graphite) to explain the high conductivity anomaly. Geophysicists traditionally have considered aqueous fluids as the likely origin of this anomaly (Hyndman and Hyndman, 1968). This is because in active tectonic environments, aqueous fluid can be supplied *via* metamorphic dehydration, igneous activity, and introduction of fluids derived from dehydration of subducting plates. However, in stable craton, fluids would be gravitationally driven but also would be able to gain access only to the upper 10 km or so of the crust. Below this depth the crust would be "dry." The discovery of grain-boundary films on minerals in high-grade rocks (Frost et al., 1989) has led to a very reasonable alternative to aqueous fluids as a cause for conductivity anomalies. Duba and Shankland (1982) used the upper bound formula to calculate the absolute amount of well-ordered graphite required to enhance the conductivity of dry rocks. They found that a volume fraction of carbon of only 5×10^{-6}% could cause a conductivity of 0.1 S/m, if well interconnected. However, high wetting angles of CO_2-quartz systems cannot establish along grain boundaries to form a connecting graphite film upon reduction with CO_2. Although both water and graphite can produce high conducting layers at deeper levels of the crust, any a priori preference for only one of them appears inadequate in view of the great and obvious diversities of geological situations.

Electrical conductivity of mantle rocks

Upper mantle

Upper mantle rocks are composed of olivine, orthopyroxene, clinopyroxene, and aluminous phases, such as plagiocase, spinel, and garnet. These variations in mineral proprtion are dependent on pressure. Olivine is the most abundant mineral in the upper mantle, extending to 410 km depth. Elecrical conduction behavior of dry olivine shows weak anisotropy (a factor of 2–3), oxygen fugacity dependence, and a change of conduction mechanism from hopping (small polaron) to ionic conduction at high temperatures. The conductivity data of mantle peridotites are not significantly different from measurements made on olivine single crystals and polycrystals, even though the rocks contain large amounts of pyroxene (Duba and Constable, 1993). The conductivity values of olivine at high temperatures corresponding to the upper mantle condition usually are consistent with those obtained from electromagnetic observations.

The conductivity of the high-conductive layer at the top of the asthenosphere (60–100 km depth for oceanic plates) is too high to be explained by the conductivity of ordinary

anhydrous olivine. One possibility is that partial melting occurs in this region, and the melt forms an anisotropic, interconnected conductive path in the partially molten peridotites. Another possibility is that olivine in the region contains considerable amounts of hydrogen inside crystal structures, and thus electrical conductivity is substantially elevated by proton conduction (Karato, 1990). Wang et al. (2006) reported large effects of water on electrical conductivity of olivine, and concluded that 10^{-2} wt.% water is enough to explain the high conductivity anomaly of oceanic asthenosphere. On the other hand, Yoshino et al. (2006) reported that the electrical conductivity of single crystal hydrous olivine is nearly isotropic and distinctly lower than the observed conductivity. Thus, hydration of olivine cannot account for the geophysical observations, which instead may be explained by the presence of partial melt with basaltic composition elongated in the direction of plate motion. Because electrical conductivities of molten carbonates are three orders of magnitude higher than those of molten silicate and five orders of magnitude higher than those of hydrated olivine, high conductivities in the asthenosphere may be explained by the presence of very small amounts of carbonate melt in peridotite (Gaillard et al., 2008).

Mantle transition zone

Wadsleyite and ringwoodite, which are high-pressure polymorphs of olivine, are the main constituent minerals in the mantle transition zone (410–660 km depth). Xu et al. (1998a) reported that electrical conductivities of wadsleyite and ringwoodite are similar, and two orders of magnitude higher than that of olivine. The laboratory-based model obtained from previous studies showed a single large conductivity jump at the 410 km discontinuity. However, recent electromagnetic observations (Kelbert et al., 2009) did not show such a distinct conductivity jump at this depth. In addition, their conductivity values are too high to explain the conductivity–depth profiles in the transition zone obtained by the semi-global electromagnetic induction studies (Shimizu et al., 2010). This inconsistency would be caused by a significant amount of water in their samples.

The Earth's mantle transition zone also has been considered as a large water storage area because wadsleyite and ringwoodite can aquire significant amounts of water in their crystal structures. Therefore, the effect of water on electrical conductivity was investigated to determine the water content in this region. Huang et al. (2005) were the first to report water content dependence on electrical conductivity of their minerals, but without separation of the effect of hopping conduction, and estimated 0.1–0.2 wt.% water in the mantle transition zone. Later, Yoshino et al. (2008a) distinguished hopping and proton conductions of their minerals and showed that contribution of proton conduction becomes smaller at higher temperatures. They demonstrated that three moderate conductivity jumps – at the 410 km depth for the olivine–wadsleyite transition, 520 km for the wadsleyite–ringwoodite transition, and 660 km for the post-spinel transition – are consistent with those obtained from electromagnetic observations, even if the minerals are dry.

Lower mantle

The lower mantle extends from 660 to 2,890 km depth (24–136 GPa), and is mainly composed of Mg-rich silicate perovksite (Mg-Pv) and ferropericlase (FP), with a small amount of Ca perovskite. Mg-Pv is thought to be the major phase in the Earth's lower mantle, and is assumed to store much of the aluminum and iron of the lower mantle. Mg-Pv is likely to control the bulk rock conductivity of the lower mantle (Yoshino et al., 2008b). Conductivity measurements of Mg-Pv at the conditions of the uppermost lower mantle show slightly higher conductivity and lower activation enthalpies than ringwoodite. If Mg-Pv contains significant amounts of Al_2O_3, the electrical conductivity is higher than that of aluminum-free perovskite (Xu et al., 1998b). The conduction mechanism in Mg-Pv changes from hopping conduction to ionic condution with increasing temperature. Last decade, two important discoveries in the lower mantle provided new insights to interpret lower mantle structure. One is the discovery of post-perovskite, and the other is the discovery of iron spin transition (Murakami et al., 2004; Badro et al., 2003). The transition pressure of $MgSiO_3$ perovskite to post-perovskite (125 GPa) is consistent with the depth of the velocity increase at the D″ seismic discontinuity (2,700 km depth). Only one conductivity measurement of silicate post-perovskite showed extremely high conductivity ($>10^2$ S/m) in comparison with that of silicate perovskite, as well as almost no temperature dependence under the conditions of the D″ layer (Ohta et al., 2008). Electrical conductivity of Mg-Pv and FP decreases with increasing pressure due to the high-spin to low-spin transition of iron in these minerals.

Electrical structure of the Earth's mantle

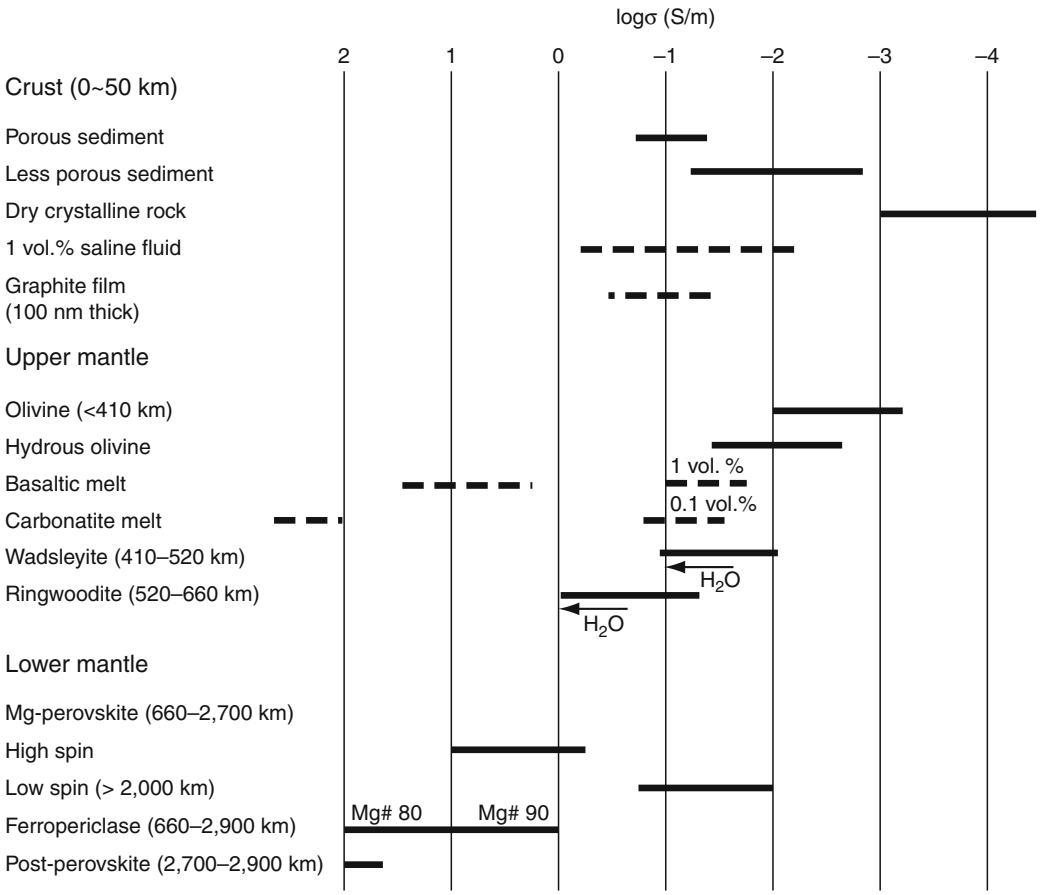
Electrical conductivity of mantle rocks tends to be similar to that of the most dominant phases, unless small amounts of conductive phases establish interconnection. The main mantle constituent minerals with higher-pressure stability fields generally have higher conductivity. A phase transformation affects electrical properties due to a change in atomic arrangement (crystal structure), as well as the nature of chemical bonding. Phase transitions would yield the conductivity jumps at certain depths where seismic velocity suddenly changes. Thus, electrical conductivity measurements of the main mantle minerals to depths of up to the core–mantle boundary indicate that the Earth's mantle is zoned electrically. This tendency can well explain the geophysically observed electrical conductivity structure in the mantle.

Conductivity generally increases with increasing depth up to the top of the lower mantle, from 10^{-3}–10^{-2} to 1 S/m. This increase is considered to represent both the

semiconductive behavior of the constituent minerals at high temperatures because mantle temperature adiabatically increases with depth, and also pressure-induced phase transformations of dominant mantle minerals. The high conductivity anomaly in the mantle would be caused by a thermal anomaly, relatively high concentrations of water and iron in mantle minerals, or presence of partial melt. Local variations of electrical conductivity are large at shallow depths, and become smaller with increasing depth. In the lower mantle, there is no phase transition down to the D″ seismic discontinuity. Because activation enthalpy for small polaron and ionic conductions is smaller for mantle minerals with higher-pressure stability fields, temperature dependence decreases as pressure increases. For this reason, a uniform electrical conductivity is expected to exist in the lower mantle. However, the high to low spin transition of iron in silicate perovskite, discovered recently would yield lower conductivity in the deeper part of the lower mantle. On the other hand, the D″ layer composed of post-perovskite might be an abnormal highly conductive layer. Thus, there should be huge conductivity jump at the D″ discontinuity.

Conclusions

The electrical conductivity of rocks within the Earth shows an ultimate variation that exceeds eight orders of magnitude (Figure 1). The electrical conductivity measurements of rocks as a function of temperature, pressure, oxygen fugacity, and impurity concentrations can constrain thermal and compositional structure in the solid Earth. At shallower depths, electrolytes filled with cracks or grain boundaries control the conductivity of rocks, whereas the conductivity of the mantle rocks is mainly controlled by semiconductive behavior of the constituent minerals. In the last 2 decades, development of measurement techniques for electrical properties of rocks and minerals under high pressure and high temperature conditions has improved our understanding of the electrical structure of the Earth's deep interior up to the core–mantle boundary. Although some discrepancies of data among laboratories have remained due to the difficulty of the electrical conductivity measurement at high pressures, it is not doubtful that the Earth's mantle is electrically zoned, and the conductivity generally increases with depth. Further improvement of experimental techniques will provide us



Electrical Properties of Rocks, Figure 1 Range of electrical conductivity of representative Earth's constituent materials at the corresponding temperature conditions.

accurate reference conductivity–depth model up to depth of the core–mantle boundary and more detailed knowledge of chemical composition and thermal distribution in the present Earth's interior by comparison with the results of the electromagnetic surveys.

Bibliography

Archie, G. E., 1942. The electrical resistivity log as an aid in determining some reservoir characteristics. *Transactions of the American Institute of Mining, Metallurgical and Petroleum Engineers*, **146**, 54–62.

Badro, J., Fiquet, G., Guyot, F., Rueff, J. P., Struzhkin, V. V., Vanko, G., and Monaco, G., 2003. Iron partitioning in Earth's mantle: toward a deep mantle discontinuity. *Science*, **300**, 789–791.

Cole, K. S., and Cole, R. H., 1941. Dispersion and absorption in dielectrics. I. alternating current field. *The Journal of Chemical Physics*, **9**, 341–352.

Duba, A. G., and Constable, S., 1993. The electrical conductivity of lherzolite. *Journal of Geophysical Research*, **98**, 11885–11899.

Duba, A. G., and Shankland, T. J., 1982. Free carbon and electrical-conductivity in the Earth's mantle. *Geophysical Research Letters*, **9**, 1271–1274.

Frost, B. R., Fyfe, W. S., Tazaki, K., and Chan, T., 1989. Grain-boundary graphite in rocks and implications for high electrical conductivity in the lower crust. *Nature*, **340**, 134–136.

Gaillard, F., Malki, M., Iacono-Marziano, G., Pichavant, M., and Scaillet, B., 2008. Carbonatite melts and electrical conductivity in the asthenosphere. *Science*, **322**, 1363–1365.

Hashin, Z., and Shtrikman, A., 1962. Avariational approach to the theory of the effective magnetic permeability of multiphase materials. *Journal of Applied Physics*, **33**, 3125–3131.

Huang, X., Xu, Y., and Karato, S., 2005. Water content in the transition zone from electrical conductivity of wadsleyite and ringwoodite. *Nature*, **434**, 746–749.

Hyndman, R. D., and Hyndman, D. W., 1968. Water saturation and high electrical conductivity in the lower continental crust. *Earth and Planetary Science Letters*, **4**, 427–432.

Karato, S., 1990. The role of hydrogen in the electrical conductivity of the upper mantle. *Nature*, **347**, 272–273.

Kelbert, A., Shultz, A., and Egbert, G., 2009. Grobal electromagnetic induction constraints on transition-zone water variation. *Nature*, **460**, 1003–1006.

Murakami, M., Hirose, K., Kawamura, K., Sata, N., and Ohishi, Y., 2004. Post-perovskite phase transition in $MgSiO_3$. *Science*, **304**, 855–858.

Ohta, K., Onoda, S., Hirose, K., Shinmyo, R., Shimizu, K., Sata, K., Ohishi, Y., and Yasuhara, A., 2008. The electrical conductivity of post-perovskite in earth's D″ layer. *Science*, **320**, 89–91.

Roberts, J. J., and Tyburczy, J. A., 1991. Frequency dependent electrical properties of polycrystalline olivine compacts. *Journal of Geophysical Research*, **96**, 16205–16222.

Shankland, T. J., and Ander, M. E., 1983. Electrical conductivity, temperatures and fluids in the lower crust. *Journal of Geophysical Research*, **88**, 9475–9484.

Shimizu, H., Koyama, T., Baba, K., and Utada, H., 2010. Revised I-D electrical conductivity structure beneath the North Pacific. *Geophysical Journal International*, **180**, 1030–1048.

Varotsos, P., and Alexopoulos, K., 1984. Physical properties of the variations of the electric field of the earth preceding earthquakes, I. *Tectonophysics*, **110**, 73–98.

Wang, D., Mookherjee, M., Xu, Y., and Karato, S., 2006. The effect of water on electrical conductivity of olivine. *Nature*, **443**, 977–980.

Xu, Y., Poe, B. T., Shankland, T. J., and Rubie, D. C., 1998a. Electrical conductivity of olivine, wadsleyite and ringwoodite under upper-mantle condition. *Science*, **280**, 1415–1418.

Xu, Y., McCammon, C., and Poe, B. T., 1998b. Effect of alumina on the electrical conductivity of silicate perovskite. *Science*, **282**, 922–924.

Yoshino, T., 2010. Laboratory electrical conductivity measurement of mantle minerals. *Surveys in Geophysics*, **31**, 163–206.

Yoshino, T., Matsuzaki, T., Yamashita, S., and Katsura, T., 2006. Hydrous olivine unable to account for conductivity anomaly at the top of the asthenosphere. *Nature*, **443**, 973–976.

Yoshino, T., Manthilake, G., Matsuzaki, T., and Katsura, T., 2008a. Dry mantle transition zone inferred from electrical conductivity of wadsleyite and ringwoodite. *Nature*, **451**, 326–329.

Yoshino, T., Yamazaki, D., Ito, E., and Katsura, T., 2008b. No interconnection of ferro-periclase in post-spinel phase inferred from conductivity measurement. *Geophysical Research Letters*, **35**, L22303, doi:10.1029/2008GL035932.

Cross-references

Earth's Structure, Lower Mantle
Earth's Structure, Upper Mantle
Earthquakes, Source Theory
Magnetotelluric Interpretation

ELECTRICAL RESISTIVITY SURVEYS AND DATA INTERPRETATION

Meng Heng Loke
Geotomo Software Sdn. Bhd., Gelugor, Penang, Malaysia

Definition

Electrical survey. Mapping the subsurface resistivity by injecting an electric current into the ground.

Electrode. Commonly a metal rod through which current is injected into the ground, or is used to measure the induced voltage on the ground surface.

Least-squares resistivity inversion. Finding the subsurface resistivity model that minimizes the sum of squares of the differences between the measured and calculated apparent resistivity values.

Introduction

Electrical resistivity surveys map the subsurface structure by making electrical measurements near the ground surface. An electric current is injected into the ground through two electrodes and the voltage difference is measured between two other electrodes (Figure 1a). The true subsurface resistivity can be estimated by making the measurements of potential difference at different positions of the current and potential electrodes, converting these values into apparent resistivity and then inverting the data set. The ground resistivity is related to various geological parameters such as the mineral and fluid content, porosity, and degree of water saturation in the rock (see *Electrical Properties of Rocks*). Over the past 20 years, the resistivity method has undergone rapid developments in the instrumentation, field survey

techniques, and data interpretation. It has now become one of the standard geophysical exploration techniques widely used in environmental, engineering, hydrogeological, and mining investigations.

The basic data from a resistivity survey are the positions of the current and potential electrodes, the current (I) injected into the ground and the resulting voltage difference (ΔV) between the potential electrodes (Figure 1a). The current and voltage measurements are then converted into apparent resistivity (ρ_a) by using the following formula

$$\rho_a = k\frac{\Delta V}{I}, \tag{1}$$

where k is the geometric factor that depends on the configuration of the current and potential electrodes (Koefoed, 1979). Over the years, various electrode configurations (or arrays) have been developed. Figure 1 shows the arrangements for some commonly used arrays. A discussion on the merits of the different arrays is given by Dahlin and Zhou (2004). Determining the true subsurface resistivity from the apparent resistivity values is the data inversion problem.

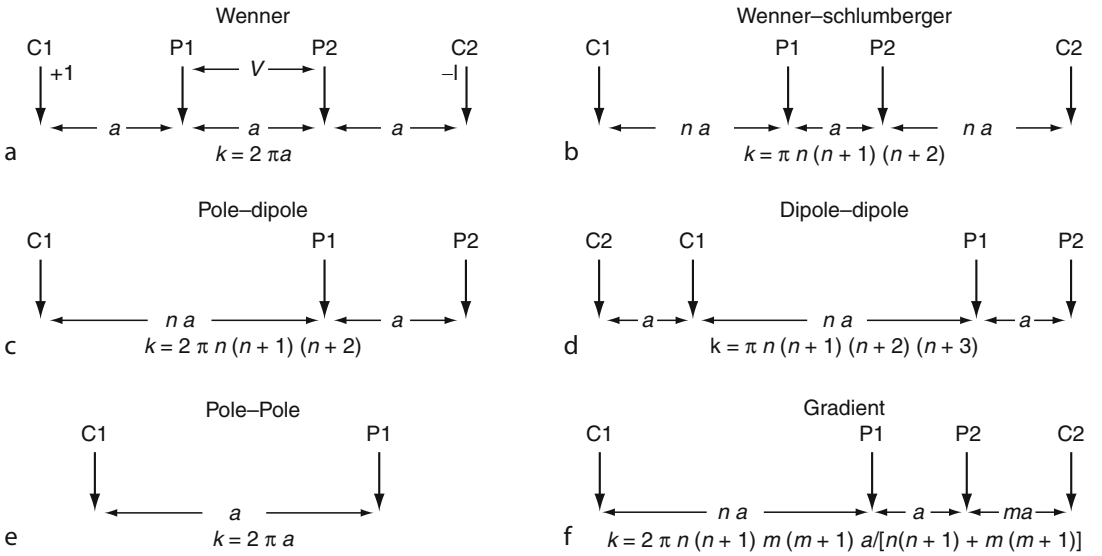
Traditional profiling and sounding surveys

The resistivity survey method has its origin in the 1920s due to the work of the Schlumberger brothers. Traditionally resistivity surveys were divided into profiling and sounding surveys. The distances between the electrodes are kept fixed in a profiling survey, such as in the Wenner survey (Figure 1a), and the four electrodes are moved along the survey line. The data interpretation for profiling surveys was mainly qualitative. The second type of survey is the vertical sounding method, such as the Schlumberger survey (Figure 1b), where the center point of the electrodes array remains fixed but the spacing between the electrodes is increased to obtain information about the deeper sections of the subsurface. The Schlumberger array is usually used for sounding surveys. Apparent resistivity plotted as a function of the current electrode spacing gives information about the subsurface resistivity variations. A sample sounding curve is shown in Figure 2.

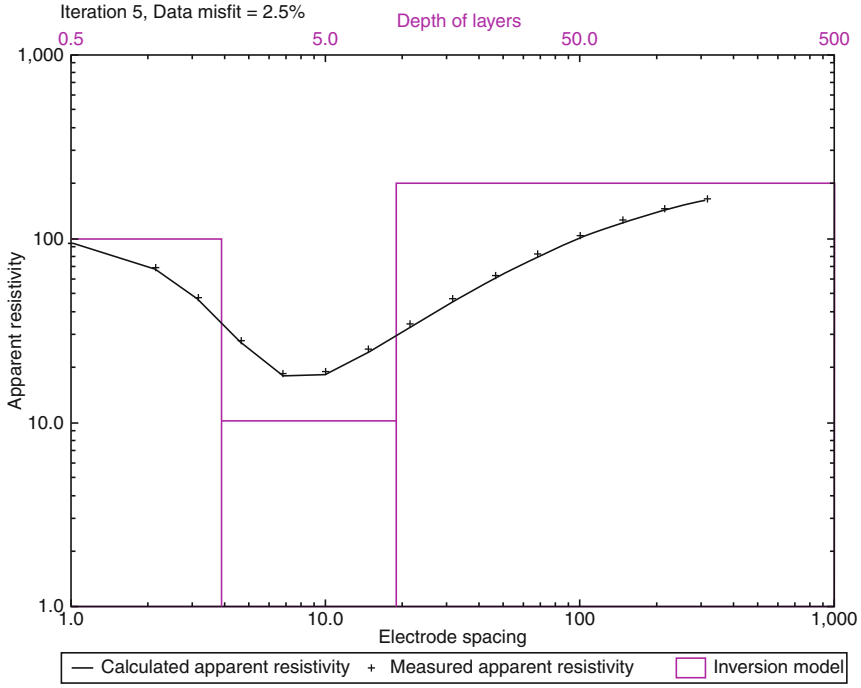
Quantitative data interpretation for sounding surveys uses a one-dimensional (1-D) earth model with a series of horizontal layers (Figure 3a). A major advance in automatic data modeling was made in the 1970s with the development of the linear filter method (Ghosh, 1971) that greatly reduced the computer time required to calculate the apparent resistivity values. This made it practical to carry out automatic inversion of resistivity-sounding data even on relatively slow early microcomputers (Koefoed, 1979). One commonly used method for 1-D data inversion is the damped least-squares method (see *Inverse Theory, Linear*) (Inman, 1975) that is based on the following equation

$$\left(\mathbf{J}^{\mathbf{T}}\mathbf{J} + \lambda\mathbf{I}\right)\Delta\mathbf{q} = \mathbf{J}^{\mathbf{T}}\Delta\mathbf{g}, \tag{2}$$

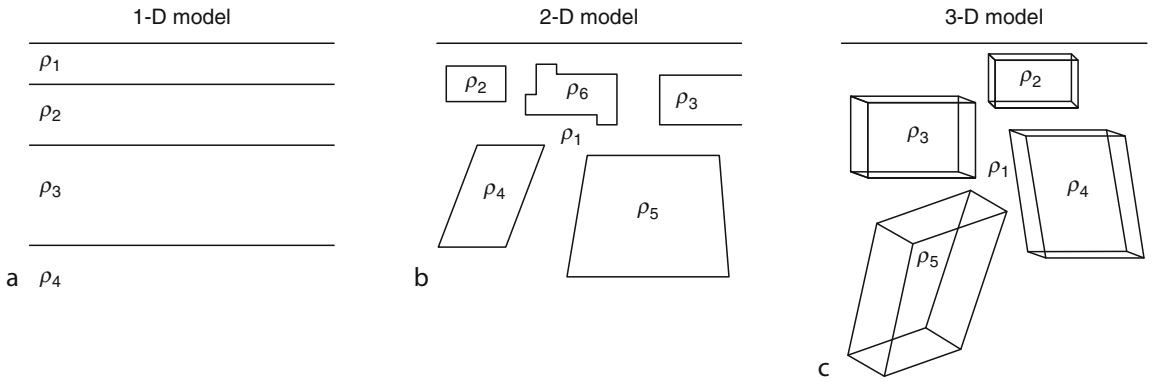
where the discrepancy vector $\Delta\mathbf{g}$ contains the difference between the logarithms of the measured and the calculated apparent resistivity values and $\Delta\mathbf{q}$ is a vector consisting of the deviation of the estimated model parameters from the true model. Here, the model parameters are the logarithms of the resistivity and thickness of the model layers. $\mathbf{J}$ is the Jacobian matrix of partial derivatives of apparent



Electrical Resistivity Surveys and Data Interpretation, Figure 1 Common arrays used in resistivity surveys and their geometric factors. The current is injected into the ground through the C1 and C2 electrodes while the voltage difference is measured between the P1 and P2 electrodes. For the arrays with less than four electrodes, the remaining electrodes are placed at sufficiently large distances so that they do not affect the measurements.



Electrical Resistivity Surveys and Data Interpretation, Figure 2 Example of a 1-D resistivity sounding curve and interpretation model with 3 layers.



Electrical Resistivity Surveys and Data Interpretation, Figure 3 The three different models used in the interpretation of resistivity measurements.

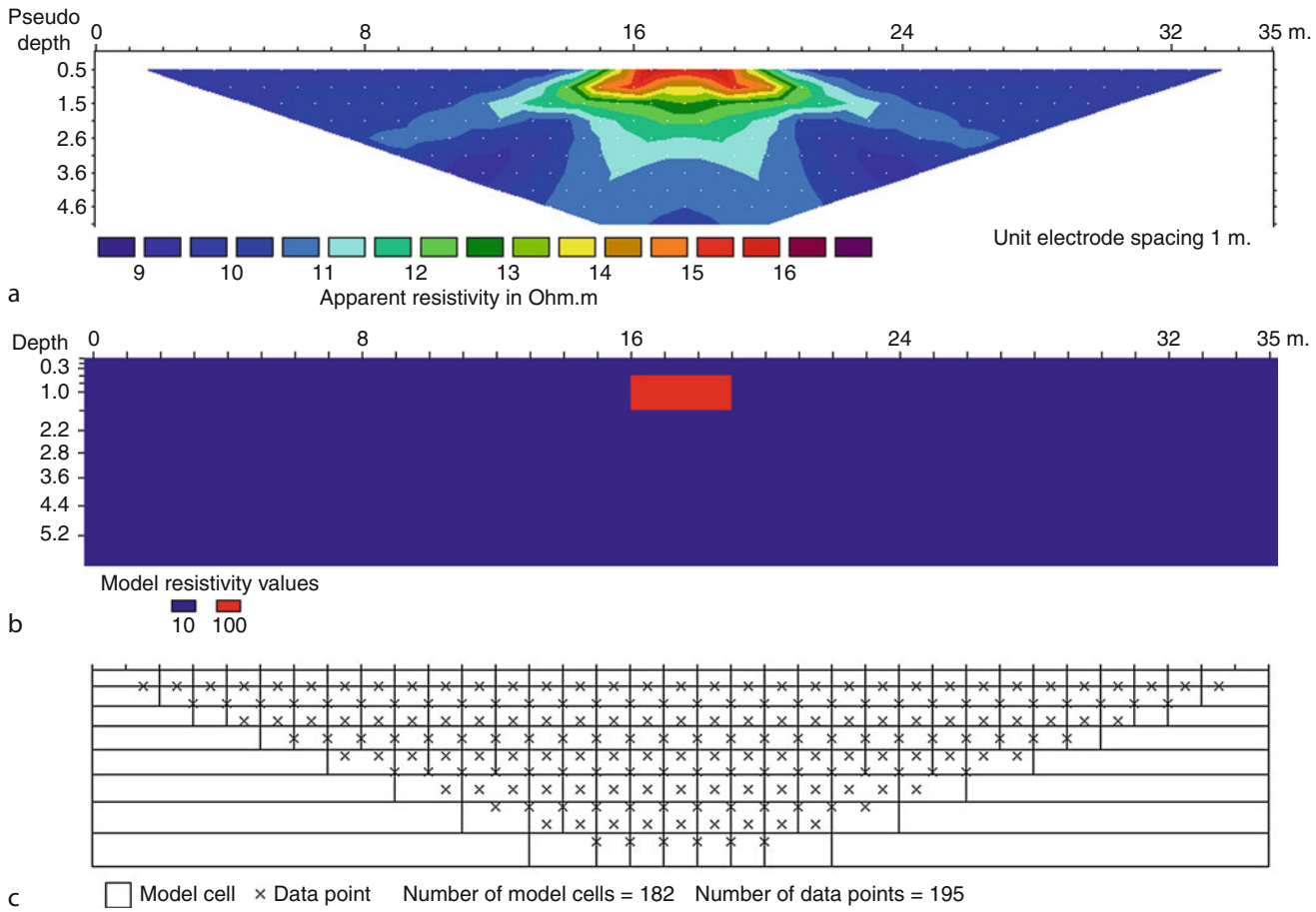
resistivity with respect to the model parameters. λ is a damping or regularization factor (Marquardt, 1963) that stabilizes the ill-condition Jacobian matrix usually encountered for geophysical problems (see *Inverse Theory, Singular Value Decomposition*). Starting from an initial model (such as a homogeneous earth model), this method iteratively refines the model so as to reduce the data misfit to a desired level (usually less than 5%). Figure 2 also shows an example of a 1-D interpretation model.

The resistivity sounding method has been used for many years, particularly in groundwater exploration. It gives useful results for geological situations (such as the water-table) where the 1-D model is approximately true. The greatest limitation of this method is that it does not take into account lateral changes in the subsurface resistivity. Such changes are probably the rule rather than the exception. The failure to include the effect of lateral changes can result in errors in the interpreted resistivity and thickness of the layers. The offset Wenner method was introduced by Barker (1978) to reduce the effect of lateral variations on the sounding data. However, to obtain a more accurate picture of the subsurface resistivity distribution, a two-dimensional (2-D) survey and interpretation model is required.

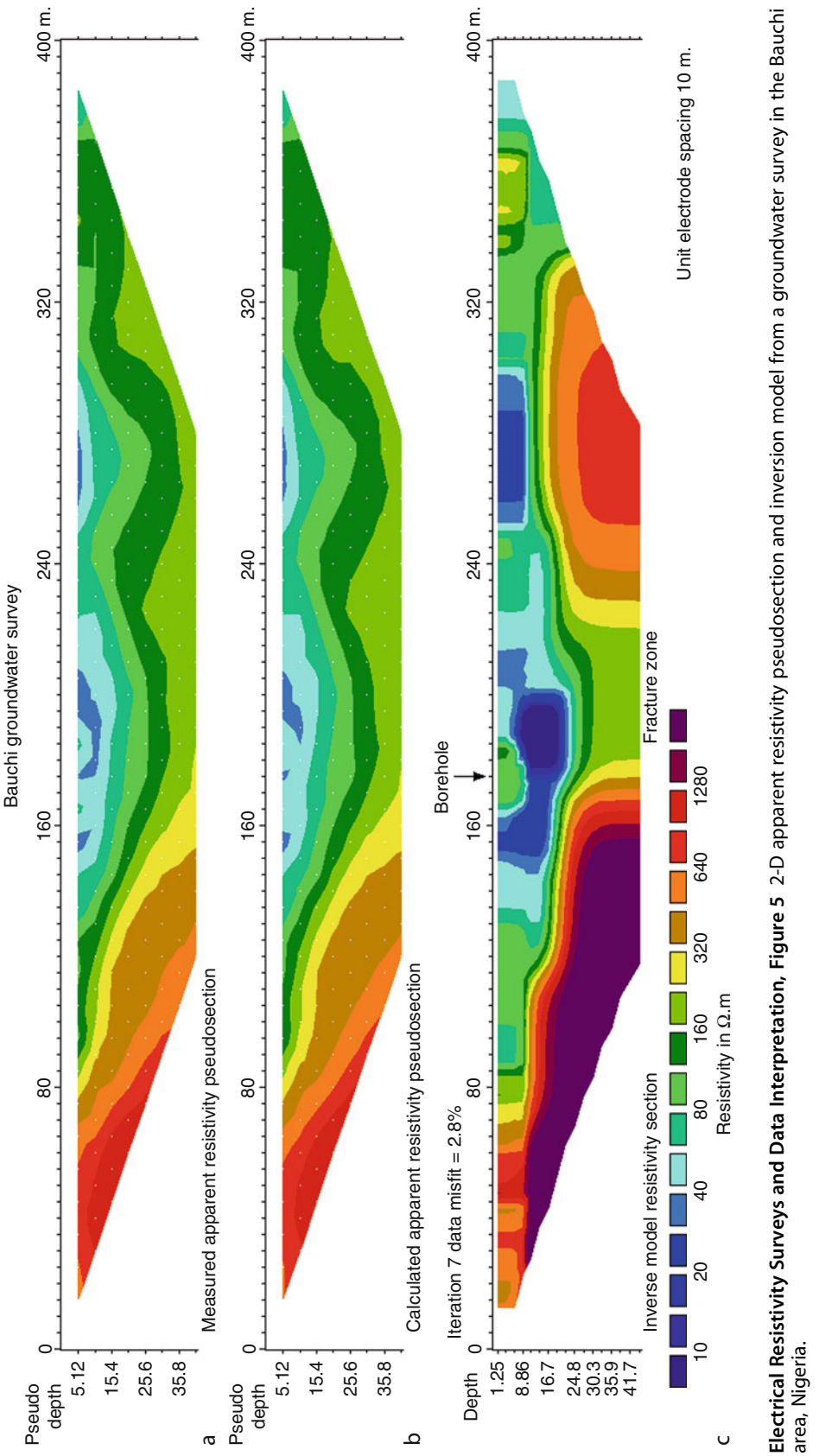
Two-dimensional resistivity imaging surveys

In recent years, development of the multielectrode resistivity meter system (see *Instrumentation, Electrical Resistivity*) has made 2-D surveys a practical tool for mapping moderately complex geological environments. The apparent resistivity measurements from the survey are commonly plotted in the form of a pseudosection (Figure 4a). The horizontal position of a data point is the average of the locations of the electrodes in the array used to make the measurement. The vertical position of the plotting point is commonly set at the median depth of investigation (Edwards, 1977) of the array. The pseudosection is a useful method to present the data in a pictorial form and as an initial guide for further quantitative interpretation. One useful practical application of the pseudosection is in identifying bad apparent resistivity measurements that usually appear as points with unusually high or low values. The pseudosection gives a distorted picture of the subsurface because the shapes of the contours depend on the type of array used as well as the true subsurface resistivity. Figure 4a shows an example of a Wenner-array pseudosection where the real structure is a rectangular block model (Figure 4b) in a homogenous medium. An inversion of the data set using a 2-D model is necessary to obtain an accurate picture of the subsurface resistivity.

A 2-D model that consists of a large number of rectangular cells is commonly used to interpret the data (Loke and Barker, 1996a). Figure 4c shows an example of a 2-D model where the distribution of the data points in the pseudosection is used as a guide in setting the arrangement of the model cells. The resistivity of the cells is allowed to vary in the vertical and the horizontal direction, but the size and position of the cells are fixed. An inversion scheme is used to determine the resistivity of the cells that will produce a response that agrees with the measured values. The finite-difference (see *Numerical Methods, Finite Difference*) or finite-element method (see *Numerical Methods, Finite Element*) is used to calculate the apparent resistivity values for the 2-D model. A non-linear optimization method is then used to automatically change the resistivity of the model cells to minimize the difference



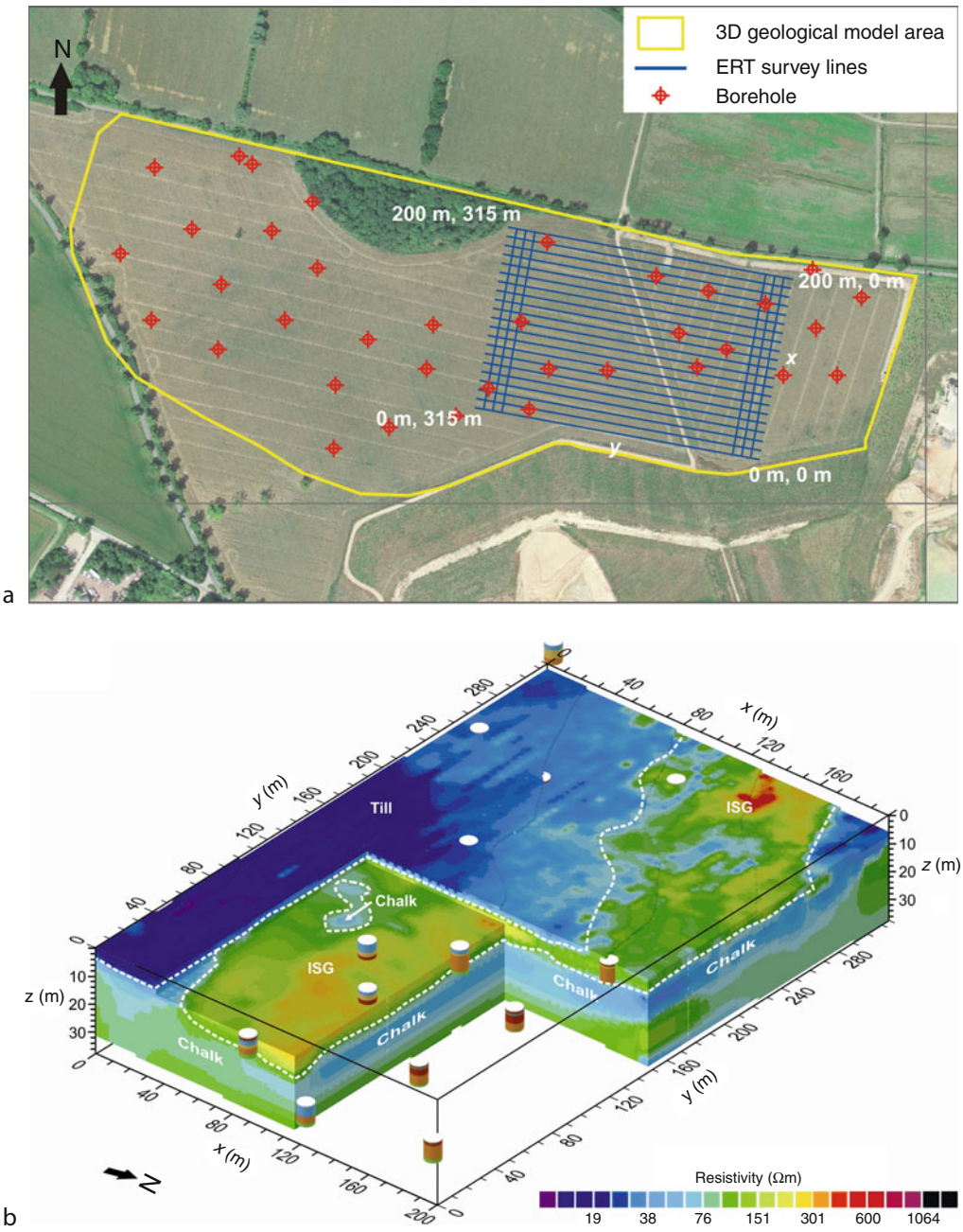
Electrical Resistivity Surveys and Data Interpretation, Figure 4 (**a**) Example of an apparent resistivity pseudosection (Wenner array). (**b**) The actual structure with a high resistivity block in a homogenous medium. (**c**) Example of a 2-D model showing the arrangement of the model cells.



Electrical Resistivity Surveys and Data Interpretation, Figure 5 2-D apparent resistivity pseudosection and inversion model from a groundwater survey in the Bauchi area, Nigeria.

between the measured and calculated apparent resistivity values. The inversion problem is frequently ill-posed due to incomplete, inconsistent and noisy data. Smoothness or other constraints (e.g., in Equation 2) are usually incorporated to stabilize the inversion procedure such that numerical artifacts are avoided. As an example, the following equation includes a model smoothness constraint to the least-squares optimization method,

$$\left(\mathbf{J}^{\mathrm{T}}\mathbf{J}+\lambda\mathbf{F}\right)\Delta\mathbf{q}_k=\mathbf{J}^{\mathrm{T}}\Delta\mathbf{g}_k-\lambda\mathbf{F}\mathbf{q}_{k-1},\ \text{where}\quad \mathbf{F}=\alpha_x\mathbf{C}_x^{\mathrm{T}}\mathbf{C}_x+\alpha_z\mathbf{C}_z^{\mathrm{T}}\mathbf{C}_z. \tag{3}$$



Electrical Resistivity Surveys and Data Interpretation, Figure 6 (**a**) Map showing the survey grid of the BGS survey at Ingham, Suffolk (Aerial Photography © UKP/Getmapping License No. UKP 2010/01). (**b**) 3-D resistivity inversion model with cut-outs, annotation and borehole logs. The target of the survey is the Ingham sand and gravel (ISG) deposits. Details on the boreholes are available in Chambers et al. (2007). Reproduced with permission of the British Geological Survey and Tarmac Ltd.

Here, $\mathbf{C}_x$ and $\mathbf{C}_z$ are the roughness filter matrices in the horizontal (x) and vertical (z) directions, and α_x and α_z are the respective relative weights of the roughness filters. k represents the iteration number. One common form of the roughness filter is the first-order difference matrix (deGroot-Hedlin and Constable, 1990), but the elements of the matrices can be modified to introduce other desired characteristics into the inversion model (Pellerin and Wannamaker, 2005; Farquharson, 2008). The method can also be modified to produce "blocky" models for regions that are piecewise constant and separated by sharp boundaries (Loke et al., 2003). The laterally constrained inversion method (Auken and Christiansen, 2004) that includes sharp boundaries has been successfully used in areas with a layered subsurface structure. A number of microcomputer based software is now available that can automatically carry out the inversion of a 2-D survey data set in seconds.

Figure 5a shows an example of a measured apparent resistivity pseudosection from a groundwater survey (Acworth, 1987) using the Wenner array. The resistivity model after inversion and the calculated apparent resistivity pseudosection for this model are shown in Figure 5c and 5b, respectively. The model obtained after inversion has a prominent low resistivity fracture zone in the bedrock (approximately between the 180 and 230 m marks and below a depth of 20 m). A borehole that was placed at the 175 m mark had yields that were lower than expected, possibly because it lies at the edge of the fracture zone. The placement of the well was partly based on EM profiling data. The shallow low resistivity anomaly with values of less than 40 Ω·m in the overburden had probably shifted the EM anomaly to the left of the fracture zone.

Three-dimensional resistivity imaging surveys

Since all geological structures are three-dimensional (3-D) in nature, a 3-D resistivity survey and interpretation model (Figure 3c) should give the most accurate results. Although it has not reached the same level of usage as 2-D surveys, it is now more widely used in very complex areas such as in many environmental (Dahlin et al., 2002; Chambers et al., 2006; Rucker et al., 2010) and mineral (White et al., 2001; Bingham et al., 2006) exploration surveys. Much of the early 3-D surveys used the pole-pole array over relatively small grids (up to about 20 by 20 electrodes) with measurements in different directions (Park and Van, 1991; Li and Oldenburg, 1992). The use of other arrays, such as the pole-dipole and Wenner–Schlumberger, is now becoming more common in 3-D surveys that involve thousands of electrode positions (White et al., 2001; Chambers et al., 2006). A true 3-D survey requires placing of electrodes in the form of a 2-D grid. However, a cost-effective strategy is usually followed wherein 3-D data sets are collated from independent parallel 2-D survey lines with possibly a few crosslines. This strategy greatly reduces the cost of a 3-D survey and also allows for 2-D inversions of each individual line. The data interpretation techniques used for 2-D surveys can be extended to 3-D surveys (Loke and Barker, 1996b). Fast computer software that takes minutes to hours to invert a 3-D data set (depending on the size of the data set) on multicore PCs is now available (Rucker et al., 2010). Other non-linear optimization methods such as neural networks, simulated annealing and conjugate gradient techniques have also been used for resistivity data inversion (Pellerin and Wannamaker, 2005).

Figure 6 shows an example of such a 3-D survey at Ingham, Suffolk by the British Geological Survey (Chambers et al., 2007) to map sand and gravel deposits. Most of the survey lines are in an approximately east–west direction with a few crosslines near the edges (Figure 6a). The area has a fairly complex geology with Chalk bedrock overlain by the fluvial Ingham sand and gravel, glacial sand, and clayey till (Figure 6b). The till has the lowest resistivity values of below 50 Ω·m while the sand and gravel is generally between 100 to 1,000 Ω·m. The Chalk shows a distinct increase of resistivity with depth due to weathering.

The data inversion methods have been adapted for other types of surveys such as induced polarization (IP) surveys (White et al., 2001), measurements across boreholes (Wilkinson et al., 2008), surveys in water covered areas (Loke and Lane, 2005), and for cylindrical geometries (Chambers et al., 2003; al Hagrey et al., 2004). The least-squares optimization method has also been modified for time-lapse surveys so as to minimize temporal changes in the model resistivity values (Loke, 2001; Kim et al., 2009).

Summary

The electrical resistivity survey method has undergone tremendous changes over the past 20 years. While traditional resistivity profiling and sounding surveys are still widely used, 2-D imaging survey is now the method of choice for most areas as this can accurately map moderately complex structures. The field equipment and computer interpretation software are widely available. 3-D surveys now play an increasingly important role in very complex areas. In many cases the 3-D data set is collated from a series of parallel 2-D survey lines to reduce the survey cost.

Acknowledgments

I would like to thank the British Geological Survey and Tarmac Ltd for permission to use the figures for the Ingham survey.

Bibliography

Acworth, R. I., 1987. The development of crystalline basement aquifers in a tropical environment. *Quarterly Journal of Engineering Geology*, **20**, 265–272.

al Hagrey, S. A., Meissner, R., Werban, U., Rabbel, W., and Ismaeil, A., 2004. Hydro-, bio-geophysics. *The Leading Edge*, **23**, 670–674.

Auken, E., and Christiansen, A. V., 2004. Layered and laterally constrained 2D inversion of resistivity data. *Geophysics*, **69**, 752–761.
Barker, R. D., 1978. The offset system of electrical resistivity sounding and its use with a multicore cable. *Geophysical Prospecting*, **29**, 128–143.
Bingham, D., Nimeck, G., Wood, G., and Mathieson, T., 2006. 3D resistivity in the Athabasca basin with the pole-pole array. In *Geophysical methods and techniques applied to uranium exploration workshop proceedings*. SEG Annual General Meeting 2006, New Orleans.
Chambers, J. E., Loke, M. H., Ogilvy, R. D., and Meldrum, P. I., 2003. Non-invasive monitoring of DNAPL migration through a saturated porous medium using electrical impedance tomography. *Journal of Contaminant Hydrology*, **68**, 1–22.
Chambers, J. E., Kuras, O., Meldrum, P. I., Ogilvy, R. D., and Hollands, J., 2006. Electrical resistivity tomography applied to geologic, hydrogeologic, and engineering investigations at a former waste-disposal site. *Geophysics*, **71**, B231–B239.
Chambers, J. E., Weller, A. L., Wilkinson, P. B., Burke, H. F., Ogilvy, R. D., Aumonier, J., Penn, S., Kuras, O., and Meldrum, P. I., 2007. *The Development of Electrical Resistivity Tomography (ERT) for Sand and Gravel Resource Visualisation: Case Histories.* Nottingham: British Geological Survey Commissioned Report, CR/07/175.
Dahlin, T., and Zhou, B., 2004. A numerical comparison of 2D resistivity imaging with ten electrode arrays. *Geophysical Prospecting*, **52**, 379–398.
Dahlin, T., Bernstone, C., and Loke, M. H., 2002. A 3D resistivity investigation of a contaminated site at Lernacken in Sweden. *Geophysics*, **60**, 1682–1690.
deGroot-Hedlin, C., and Constable, S., 1990. Occam's inversion to generate smooth, two-dimensional models from magnetotelluric data. *Geophysics*, **55**, 1613–1624.
Edwards, L. S., 1977. A modified pseudosection for resistivity and induced-polarization. *Geophysics*, **42**, 1020–1036.
Farquharson, C. G., 2008. Constructing piecewise-constant models in multidimensional minimum-structure inversions. *Geophysics*, **73**, K1–K9.
Ghosh, D. P., 1971. The application of linear filter theory to the direct interpretation of geoelectrical resistivity sounding measurements. *Geophysical Prospecting*, **19**, 192–217.
Inman, J. R., 1975. Resistivity inversion with ridge regression. *Geophysics*, **40**, 798–817.
Kim, J. H., Yi, M. J., Park, S. G., and Kim, J. G., 2009. 4D inversion of DC resistivity monitoring data acquired over a dynamically changing earth model. *Journal of Applied Geophysics*, **68**, 522–532.
Koefoed, O., 1979. *Geosounding Principles 1: Resistivity Sounding Measurements*. Amsterdam: Elsevier Science Publishing Company.
Li, Y., and Oldenburg, D. W., 1992. Approximate inverse mappings in DC resistivity problems. *Geophysical Journal International*, **109**, 343–362.
Loke, M. H., 2001. Constrained time-lapse resistivity imaging inversion. In *Proceedings of the 2001 Symposium on the Application of Geophysics to Engineering and Environmental Problems (SAGEEP)*. EEGS, Denver, Colorado, March 4–7.
Loke, M. H., and Barker, R. D., 1996a. Rapid least-squares inversion of apparent resistivity pseudosections using a quasi-Newton method. *Geophysical Prospecting*, **44**, 131–152.
Loke, M. H., and Barker, R. D., 1996b. Practical techniques for 3D resistivity surveys and data inversion. *Geophysical Prospecting*, **44**, 499–523.
Loke, M. H., and Lane, J. W., Jr., 2005. Inversion of data from electrical resistivity imaging surveys in water-covered areas. *Exploration Geophysics*, **35**, 266–271.
Loke, M. H., Acworth, I., and Dahlin, T., 2003. A comparison of smooth and blocky inversion methods in 2D electrical imaging surveys. *Exploration Geophysics*, **34**, 182–187.
Marquardt, D. W., 1963. An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, **11**, 431–441.
Park, S. K., and Van, G. P., 1991. Inversion of pole-pole data for 3-D resistivity structures beneath arrays of electrodes. *Geophysics*, **56**, 951–960.
Pellerin, L., and Wannamaker, P. E., 2005. Multi-dimensional electromagnetic modeling and inversion with application to near-surface earth investigations. *Computers and Electronics in Agriculture*, **46**, 71–102.
Rucker, D., Loke, M. H., Levitt, M. T., and Noonan, G. E., 2010. Electrical resistivity characterization of an industrial site using long electrodes. *Geophysics*, **75**, WA95–WA104.
White, R. M. S., Collins, S., Denne, R., Hee, R., and Brown, P., 2001. A new survey design for 3D IP modelling at Copper hill. *Exploration Geophysics*, **32**, 152–155.
Wilkinson, P. B., Chambers, J. E., Lelliott, M., Wealthall, P., and Ogilvy, R. D., 2008. Extreme sensitivity of crosshole electrical resistivity tomography measurements to geometric errors. *Geophysical Journal International*, **173**, 49–62.

Cross-references

Electrical Properties of Rocks
Instrumentation, Electrical Resistivity
Inverse Theory, Linear
Inverse Theory, Singular Value Decomposition
Numerical Methods, Finite Difference
Numerical Methods, Finite Element

ELECTRONIC GEOPHYSICAL YEAR

William K. Peterson[1], Daniel N. Baker[1], C. E. Barton[2], Peter Fox[3], M. A. Parsons[4], Emily A. CoBabe-Ammann[5]
[1]Laboratory for Atmospheric and Space Physics, University of Colorado, Boulder, CO, USA
[2]Australian National University, Canberra, Australia
[3]Rensslear Polytechnic Institute, Troy, NY, USA
[4]National Snow and Ice Data Center, University of Colorado, Boulder, CO, USA
[5]Emily A. CoBabe and Associates, Boulder, CO, USA

Definition

eGY. Electronic Geophysical Year. An international science year (2007–2008) celebrating the IGY+50 anniversary that focused on increasing access to scientific data and information for all people.

Introduction and concept

Sharing scientific data and information for the common good has long been at the heart of good scientific practice. For the geosciences, this culminated in the outstandingly successful International Geophysical Year, 1957–1958 (IGY). During IGY, in addition to the initiation of much new research, scientists and member countries worked together to emplace a large number of geophysical observatories around the globe, and establish a network of

World Data Centers for archiving data and for making data available without restrictions.

The Electronic Geophysical Year, 2007–2008 (*e*GY), was conceived in the conviction that the benefits of this "Information Commons" are even greater today than they were 50 years ago. Modern information and communications technologies endow us with capabilities for sharing data and information that are unprecedented, provide impressive and timely opportunities for Earth and space science, and require international coordination to avoid replication of effort and a multiplicity of standards.

> The eGY VISION. We can achieve a major step forward in Earth and space science capability, knowledge, and usage throughout the world for the benefit of humanity by accelerating the adoption of modern, visionary practices for managing and sharing data and information.

History

In 1999, the International Union of Geodesy and Geophysics (IUGG) called on its Scientific Associations to propose activities to mark the 50-year anniversary of IGY. The International Association of Geomagnetism and Aeronomy (IAGA) responded through a resolution passed at the IUGG General Assembly in Sapporo in 2003 to lead an Electronic Geophysical Year.

*e*GY began on July 1, 2007 and ended on December 31, 2008, exactly 50 years after the start and end of IGY. Support for *e*GY came from IAGA, IUGG, NASA, the United States National Science Foundation, United States Geological Survey, and the Laboratory for Atmospheric and Space Physics (LASP) of the University of Colorado. In-kind contributions came from the American Geophysical Union (AGU), the National Centre for Atmospheric Research in Boulder, and the volunteer labor of *e*GY participants.

Scope

*e*GY focused the international science community to achieve a step increase in making past, present, and future geoscientific data (including information and services) rapidly, conveniently, and openly available. The themes of eGY were electronic data location and access, data release and permission, data preservation and rescue, data integration and knowledge discovery, capacity building in developing countries (mainly improving Internet connectivity), and education and outreach. Promoting the development of virtual observatories (VOs) and similar user-community systems for providing open access to data and services was a central feature of *e*GY.

Operations

Working groups were established to focus effort on each of the eGY theme areas. National eGY groups were particularly active in Canada, China, India, Japan, Russia, and the UK. A Web site (www.egy.org) served not only as a means of networking among eGY participants, but was developed as, and remains, an information resource for data and information practice and policy.

Presentations, theme sessions, Town Hall meetings, workshops, and demonstrations of VO technology and working systems were organized at international conferences run by AGU, The Asia Oceania Geophysics Society (AOGS), ICSU's Committees on Data for Science and Technology (CODATA), the Committee on Space Research (COSPAR), the European Geosciences Union (EGU), IAGA, and IUGG. Virtual Observatories in Geosciences (VOiG) Conference was held in Denver in June 2007, with selected papers published in the April 2008 issue of Earth Science Informatics (http://www.springerlink.com/content/1865–0473).

In 2008, an Earth and Space Science Informatics Summit was held in Rome to establish a better understanding and cooperation among the various geoinformatics and space science informatics initiatives and programs that are growing up rapidly worldwide. A report is available at http://egy.org/files/ESI_Summit_Report_final.pdf.

One of the first actions of eGY was to establish a declaration for an Earth and Space Science Information Commons. It states

> *The Electronic Geophysical Year (eGY) joins with the International Council for Science, the World Summit on the Information Society, and many other bodies in recognizing that knowledge is the common wealth of humanity. We have a shared responsibility to create and implement strategies to realize the full potential of digital information for present and future generations. In the 21st century and beyond, access to digital information and new technologies for information integration and knowledge discovery will influence the free and productive development of societies around the world. Providing ready and open access to the vast and growing collections of cross-disciplinary digital information is the key to understanding and responding to complex Earth system phenomena that influence human survival.*

and includes eight articles addressing the major aspects of an Earth and space science information commons (mirroring the eGY themes).

Achievements

Principal legacies of eGY are (1) stronger awareness of the role that informatics plays in modern research, (2) adoption of virtual observatories and similar systems for accessing data, information, and services, and (3) an expanding infrastructure at the international and national levels. As with the IGY, the mission of eGY is being carried forward through existing and newly formed national and international organizations.

Advances in which the eGY effort played a role include:

- Establishment of the Earth & Space Science Informatics Division in EGU
- Establishment of a COSPAR's Panel for Informatics
- Establishment of an IUGG Union Commission for Data & Information
- Strengthening of AGU's Earth & Space Science Informatics Special Focus Group

- IUGG readmission as a member of CODATA and support for a CODATA delegate
- Drafting of implementation guidelines for GEOSS's (GEOSS = the Global Earth Observing System of Systems) data and information policy
- Formation of an International Polar Year (IPY) Data Policy and Management Subcommittee under ICSU, WMO, and CODATA as a CODATA Task Group
- Contributions to development of the IPY Data and Information Service (IPYDIS, http://ipydis.org) and the Polar Information Commons (http://polarcommons.org)
- Formation of a CODATA Task Group: eGY Earth & Space Science Data Interoperability
- ICSU/SCID and SCCID: eGY membership of SCID to review data policy and representation on ICSU's Strategic Coordinating Committee for Information and Data
- eGY-Africa as a contribution to IUGG's Geoscience in Africa initiative

Bibliography

Baker, D. N., and Barton, C. E., 2009. EGY: progress in global earth and space science informatics. *Data Science Journal*, **8**, 226, doi:10.2481/dsj.008-008.

Barton, C., Smith, R., and Weaver, R., 2010. Data practices, policy, and rewards in the Information Era demand a new paradigm. *Data Science Journal*, **9**, IGY95-IGY99, doi:10.2481/dsj.SS_IGY-003.

CoBabe-Ammann, E., Peterson, W. K., Baker, D., Fox, P., Barton, C., 2007. "The Electronic Geophysical Year (2007, 2008): eScience for the 21st Century", *The Leading Edge*, **26**(10), 1294–1295.

Fox, P., 2008. Virtual observatories in geosciences. *Earth Science Informatics*, **1**, 3, doi:10.1007/s12145-008-0001-8.

Peterson, W. K., 2009. Open access to digital information: opportunities and challenges identified during the electronic geophysical year. *Data Science Journal*, **8**, S108-S112, doi:10.2481/dsj.IGY-002.

Cross-references

International Geophysical Year
International Polar Year 2007–2008
International Year of Planet Earth

ENERGY BUDGET OF THE EARTH

Jean-Claude Mareschal[1], Claude Jaupart[2]
[1]Centre GEOTOP-UQAM, University of Québec at Montréal, Montréal, QC, Canada
[2]Institut de Physique du Globe, Paris, France

Definition

The Earth is a physical system with an energy budget where the outgoing energy must be balanced against all the sources of energy. The planet is not in equilibrium: the difference between the energy flowing out of the Earth and the internal heat production is accounted for by the secular cooling of the mantle and core.

Total heat loss of the earth

The main constraint on the total heat loss of the Earth comes from many heat flux measurements that have been made both on continents and oceans. Because the continents and the oceans have different thermal structures and because the heat flux data are affected by different biases in each region, it is necessary to analyze separately the continental and the oceanic heat losses (Jaupart et al., 2007).

Continental heat flow

There are more than 18,000 reported heat flux measurements on the continents and their margins. This number includes many additions by individual researchers to the data in the compilation by Pollack et al. (1993). The measurements are very unevenly distributed geographically with the majority of the data coming from Eurasia and North America. Two large continents, Antarctica and Greenland, are almost unsampled. The raw average of all continental heat flux values is 79.7 mW m^{-2}. This value is biased because the data set includes many data from the USA that were collected for geothermal exploration in anomalously hot regions. The bias is made evident when data from the USA and the rest of the world are analyzed separately: the average is 112.4 mW m^{-2} for the USA versus 65.7 mW m^{-2} for the rest of the world. One way to remove the bias is by area-weighting, that is, estimating the average over sufficiently large windows (1° × 1°), and then taking the average of all the windows. It yields an average heat flux for all the continents of 63 mW m^{-2} (Table 1). Multiplying by the continental surface area of 210×10^6 km^2 gives 13.5 TW for the heat loss through the continents. An alternative method to calculate continental heat loss is to bin the data by "age," determine the average heat flux for each age group, and integrate the

Energy Budget of the Earth, Table 1 Continental heat flux statistics[a]

	$\mu(Q)$ mW m^{-2}	$\sigma(Q)$ mW m^{-2}	$N(Q)$
World			
All values	79.7	162	14,123
Averages 1° × 1°	65.3	82.4	3,024
Averages 2° × 2°	64.0	57.5	1,562
Averages 3° × 3°	63.3	35.2	979
USA			
All values	112.4	288	4,243
Averages 1° × 1°	84	183	532
Averages 2° × 2°	78.3	131.0	221
Averages 3° × 3°	73.5	51.7	128
Without USA			
All values	65.7	40.4	9,880
Averages 1° × 1°	61.1	30.6	2,516
Averages 2° × 2°	61.6	31.6	1,359
Averages 3° × 3°	61.3	31.3	889

[a]μ is the mean, σ is the standard deviation, and N is the number of values

product of heat flux times the areal age distribution for continental crust (Pollack et al., 1993). It gives about the same value for the heat loss as the area weighted average.

Oceanic heat flow

There are more than 20,000 heat flux measurements reported in the oceans but these data are not directly used to calculate the total heat loss through the oceans. The oceanic heat loss is calculated from a model of cooling of the seafloor. The measurements are used to determine the parameters of this model. The cooling model of the seafloor is a direct consequence of the seafloor spreading hypothesis. It assumes that seafloor forms at mid-oceanic ridges and cools as it moves away from the ridge. The surface heat flux is calculated by solving the heat conduction equation with fixed surface temperature. For a cooling half-space model, heat flux q decreases with seafloor age τ as (Lister, 1977):

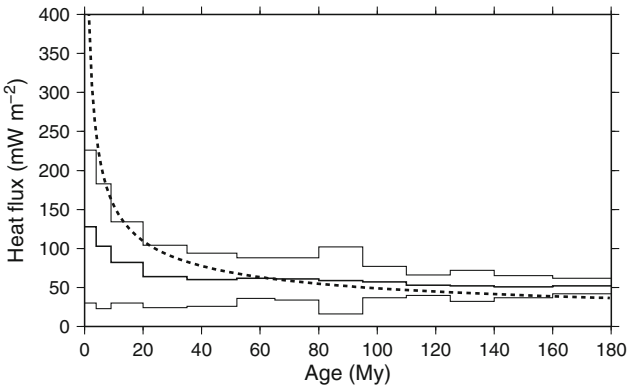
$$q(\tau) = C_Q \tau^{-1/2}$$

where C_Q depends on the temperature of the magma ascending at the seafloor, and on the thermal properties (thermal diffusivity and conductivity) of the cooling lithosphere.

In order to compare the oceanic heat flux measurements with the model, data have been binned by age. Figure 1 shows that the observed heat flux is systematically less than predicted for seafloor ages <60 My, and that heat flux is approximately constant and higher than predicted for ages >80 My.

Young seafloor

Only the heat that is conducted through the seafloor can be measured. Near the mid-oceanic ridges, where there are large open fractures in the shallow crust, much heat is transported by the movement of hydrothermal fluids. This convective component of the heat flux accounts for the difference between the observations and the predictions of the cooling model. At depth, fractures are sealed by the confining pressure and cooling is by conduction only. In regions of the seafloor that are well sealed by sediments, the heat flux observations do fit the cooling model and can be used to determine the constant C_Q. With the additional constraint being $q(\tau \to \infty) \to 0$ (Harris and Chapman, 2004), the value of C_Q ranges between 470 and 510 mW m^{-2} My$^{1/2}$.



Energy Budget of the Earth, Figure 1 Observed and predicted oceanic heat flux as a function of seafloor age. The *dotted line* is the predicted heat flux for the half-space cooling model. The average observed heat flux (±1 SD) within different age groups is shown by *continuous (thin) lines*. Data from Stein and Stein (1994).

Old seafloor

For ages greater than 80 My, the heat flux levels off and is higher than calculated for a half-space cooling model. The common interpretation is that cooling of the mantle triggers the onset of small-scale convection, which maintains a constant temperature or heat flux at fixed depth below the seafloor. Experiments designated to measure the heat flux on seafloor older than 100 My have shown that it is almost constant and equal to 48 mW m^{-2}.

Bathymetry

One consequence of the cooling of the seafloor is that the average density of a rock column increases because of thermal contraction:

$$\Delta\rho_m = -\alpha\rho_m \Delta T$$

where α is the thermal expansion coefficient, ρ_m is mantle density, and ΔT is the temperature difference at time τ. The change in bathymetry can be calculated from the isostatic balance condition:

$$\begin{aligned}\Delta h &= \frac{1}{(\rho_m - \rho_w)} \int_0^d \Delta\rho_m(z,\tau)dz \\ &= \frac{-\alpha\rho_m}{(\rho_m - \rho_w)} \int_0^d \Delta T(z,\tau)dz \\ &= \frac{\alpha}{C_p(\rho_m - \rho_w)} \int_0^\tau [q(0,\tau) - q(d,\tau)]d\tau\end{aligned}$$

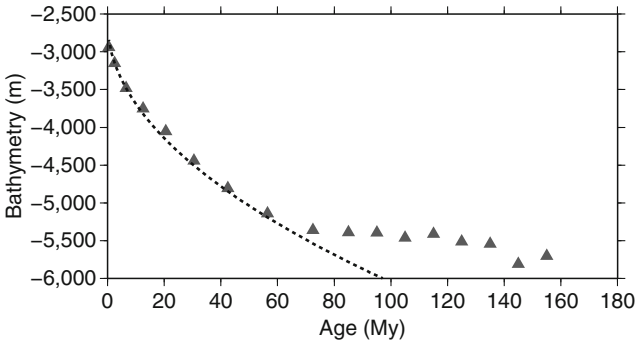
where C_p is the specific heat of the lithospheric rocks and ρ_w density of seawater. For a half-space, $d \to \infty$ and $q(d, \tau) \to 0$, we have:

$$\Delta h = \frac{\alpha}{C_p(\rho_m - \rho_w)} \int_0^\tau q(\tau)d\tau$$

It gives:

$$h = h_0 + \frac{2\alpha}{C_p(\rho_m - \rho_w)} C_Q \tau^{1/2} = h_0 + C_H \tau^{1/2}$$

The bathymetry of the seafloor fits very well the half-space cooling model for ages less than 80 My. For greater ages, the bathymetry becomes flat, confirming that heat supplied at the base balances the heat loss at the surface of the plate (Figure 2). The constant C_H is related to C_Q and, with standard values for the physical properties



Energy Budget of the Earth, Figure 2 Worldwide average seafloor bathymetry as a function of age (*triangle*), and predicted bathymetry for a cooling half-space (*dotted line*). Bathymetry data from Crosby and McKenzie (2009).

of the mantle ($\alpha = 3.1 \times 10^{-5}$ K^{-1}, $C_p = 1{,}170$ J kg^{-1} K^{-1}, $\rho_m = 3{,}330$ kg m^{-3}, and $\rho_w = 1{,}000$ kg m^{-3}), we obtain $C_H/C_Q = 704$ m^3 W^{-1} My^{-1}. Measurements of the depth to the basement on old seafloor have been used to determine a value of 345 m My$^{-1/2}$ for C_H, which translates as 480 mW m^{-2} My$^{1/2}$ for C_Q (Carlson and Johnson, 1994). This is within the range obtained from the heat flux data set.

Age distribution of the seafloor

In order to determine the total oceanic heat loss, we have to integrate the heat flux times the areal distribution of seafloor ages. The distribution of seafloor ages has been very well determined from studies of the marine magnetic anomalies (Müller et al., 2008). There are few ages higher than 180 My and the areal distribution appears to decrease linearly with age τ, such that the areal distribution can be approximated by:

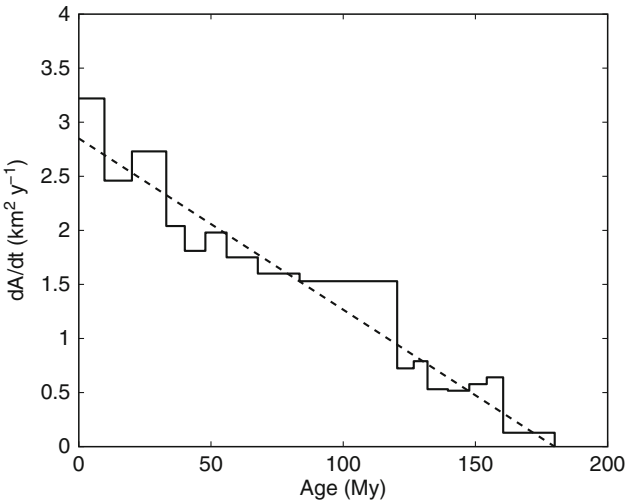
$$C_A(1 - \tau/180) \tag{1}$$

where τ is age in My (Figure 3). In order to account for the total area covered by the oceans including the marginal basins (300×10^6 km^2), the accretion rate $C_A = 3.4$ km^2 y^{-1}.

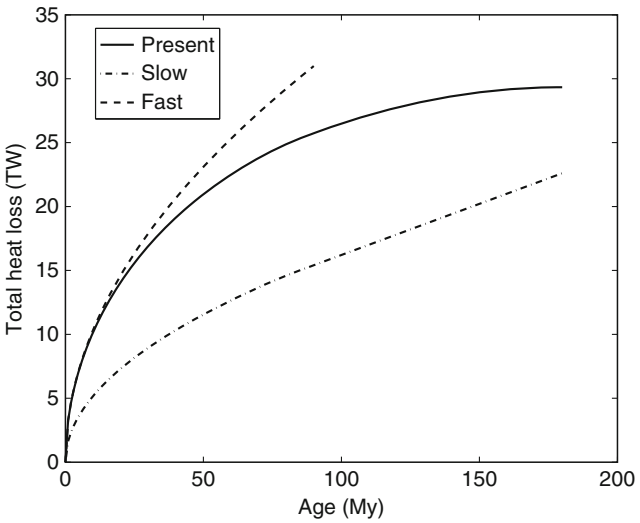
Integrating separately seafloor younger and older than 80 My gives:

$$\begin{aligned} Q_{80-} &= \int_0^{80} C_Q \tau^{-1/2} C_A(1 - \tau/180) d\tau \\ &= 24.3 \text{ TW} \\ Q_{80+} &= q_{80} \int_{80}^{180} C_A(1 - \tau/180) d\tau \\ &= 4.4 \text{ TW} Q_{\text{oceans}} = 29 \pm 1 \text{ TW} \end{aligned} \tag{2}$$

The cumulative heat loss as a function of age depends strongly on the age distribution (Figure 4). With a lower accretion rate and a rectangular age distribution, the oceanic heat loss could be as low as 22 TW.



Energy Budget of the Earth, Figure 3 Age distribution of the seafloor, not including the marginal basins: distribution from isochrons (*solid line*); best fit to the distribution (*dashed line*). Data from Cogné and Humler (2004).



Energy Budget of the Earth, Figure 4 Cumulative total heat loss of the seafloor for three different distributions of the seafloor age: the present triangular distribution, triangular distribution with twice the present spreading rate and oldest seafloor 90 My, rectangular distribution with oldest seafloor 180 My and spreading rate half of present. The total oceanic area is 300×10^6 km^2, and $C_Q = 490$ mW m^{-2} My$^{1/2}$; for ages >80 My, heat flux is 48 mW m^{-2}.

Hot spots

Hot spots bring additional heat to the oceanic plates. This heat is not accounted for by the plate cooling model and should be added to the oceanic heat loss. The extra heat flux is barely detectable, but the heat that has been put in the plate can be estimated from the volume of the swell of the seafloor: it amounts to 2–4 TW. The resulting value

for the total energy loss at the Earth's surface, 46 TW, is not much different from the first estimates by Williams and von Herzen (1974) (Table 2).

Energy budget of the mantle

The total heat loss at the Earth surface includes continental heat losses 14 TW, cooling of the sea floor, 29 TW, and heat from the hot spots, 3 TW. Part of this heat comes from radioactive heat production in the continental crust, which has been estimated by multiple and extensive sampling (Rudnick and Fountain, 1995; Rudnick and Gao, 2003; Jaupart and Mareschal, 2003). It is estimated to be 7 TW. This leaves 39 TW for the mantle, either by cooling of the seafloor or by conduction through the continental lithosphere. Three main components balance the mantle energy loss: radiogenic heat production in the mantle, heat flux from the core, and secular cooling of the mantle. Other sources (tidal dissipation, gravitational settling, changes in gravitational potential energy due to thermal contraction) contribute less than 1 TW.

Energy Budget of the Earth, Table 2 Estimates of the continental and oceanic heat flux and global heat loss

	Continental (mW m^{-2})	Oceanic (mW m^{-2})	Total (TW)
Williams and von Herzen (1974)	61	93	43
Davies (1980)	55	95	41
Sclater et al. (1980)	57	99	42
Pollack et al. (1993)	65	101	44
Jaupart et al.[a] (2007)	65	94	46

[a]The average oceanic heat flux does not include the contribution of hot spots. The total heat loss estimate does include 3 TW from oceanic hot spots

Radiogenic heat production

Different approaches have been used to estimate the composition of the Earth's mantle and its radiogenic heat production. Earth was formed by accretion of planetesimals in the Solar nebula that have the same origin as the chondritic meteorites. However, the accretion process and the giant impact that formed the Moon have modified Earth's composition relative to the meteorites. Estimates of mantle composition rely on comparing elemental ratios on Earth and meteorites, to infer the composition of bulk silicate Earth (BSE), that is, the mantle and crust. Another method assumes that compositional trends in meteorites and mantle rocks record processes in the solar nebula and the mantle, and that they will intersect at the composition of their common source. Both methods yield a heat production of ≈20 TW for BSE. Accounting for the radioelements that have been stored in the continental lithosphere leaves 13 TW in the convecting mantle. It is useful to compare the mantle heat production to its heat loss: this ratio is referred to as the *Urey ratio* of the mantle. It estimated to be in the range 0.21–0.49 with a preferred value of 0.33.

A lower bound for mantle heat production can be obtained from the composition of the rocks that are extracted from the mantle to form the seafloor. This only gives a lower bound because it does not take into account a possible hidden reservoir that has not been tapped by the extraction of the MORBs. This estimate is 11 TW for the mantle; accounting for the heat production in the continental lithosphere gives a total heat production higher than 18 TW for BSE. These various estimates are summarized in Table 3.

Energy Budget of the Earth, Table 3 Radioelement concentration and heat production (A*) in meteorites, in the Bulk Silicate Earth, in Earth mantle, and crust

	U (ppm)	Th (ppm)	K (ppm)	A* (pW kg^{-1})
CI Chondrites				
Palme and O'Neill (2003)	0.0080	0.030	544	3.5
McDonough and Sun (1995)	0.0070	0.029	550	3.4
Bulk Silicate Earth				
From CI Chondrites				
Javoy (1999)	0.020	0.069	270	4.6
From EH Chondrites				
Javoy (1999)	0.014	0.042	385	3.7
From Chondrites and Lherzolites Trends				
Hart and Zindler (1986)	0.021	0.079	264	4.9
From Elemental Ratios and Refractory Lithophile Elements Abundances				
McDonough and Sun (1995)	0.020 ± 20%	0.079 ± 15%	240 ± 20%	4.8 ± 0.8
Palme and O'Neill (2003)	0.022 ± 15%	0.083 ± 15%	261 ± 15%	5.1 ± 0.8
Average MORB Mantle Source				
Su (2000)	0.013	0.040	160	2.8
Continental Crust				
Rudnick and Gao (2003)	1.3	5.6	1.5×10^4	330
Jaupart and Mareschal (2003)	/	/	/	293 – 352

Heat flux from the core

Some authors have assumed that the heat flux from the core is equal to the heat flux transported by the hot spots. There is no reason that it should be so. The heat carried away by hot spots originating at the core mantle boundary must be less than the core's output. This lower bound is 4 TW. A different bound can be obtained by calculating the minimum heat flux that would be conducted along an adiabat in the core. The core is thought to have a very high thermal conductivity and the heat flux conducted along the adiabat is 40 mW m^{-2}, that is, 5 TW. How much energy is required to power the geodynamo has been much debated. Present estimates vary within a wide range (5–14 TW) with an average value of 9 TW (Nimmo, 2007; Buffett, 2007).

Secular cooling

Table 4 compares different ways of breaking down the energy budget of the Earth. Stacey and Davis (2008) add the radioactivity in the continental crust to that of BSE and thus overestimate the total heat production. Both Stacey and Davis (2008) and Davies (1999) assumed that all the heat from the core is carried away by the hot spots, although the hot spots provide only a weak lower bound on core heat flow.

The difference between the total mantle energy loss and the inputs from the core, radioactivity, and other sources, $39 - 1 - 13 - 9 = 16$ TW, must be accounted for by the secular cooling of the mantle. This gives a mantle cooling rate of $\approx$100 K Gy^{-1}. Mantle temperatures can be calculated from the composition of the Mid Oceanic Ridge Basalts (MORB). Petrological studies on Archean (3.5–2.5 Gy old) MORB like rocks suggest a more modest cooling rate of 50 K Gy^{-1} (Figure 5). The range of temperatures found in these studies is too wide to rule out that the secular cooling rate is higher than 50 K Gy^{-1}, but the difference between the present and secular rates of cooling raises interesting questions about the present energy budget.

Energy Budget of the Earth, Table 4 Various estimates of the breakdown of the global budget in TW. Stacey and Davis (2008) have added heat production in the continental crust to that of bulk silicate earth. They also assume that gravitational potential energy released by thermal contraction produces heat rather than strain energy. Both Stacey and Davis (2008) and Davies (1999) assume that hot spots carry all the heat from the core

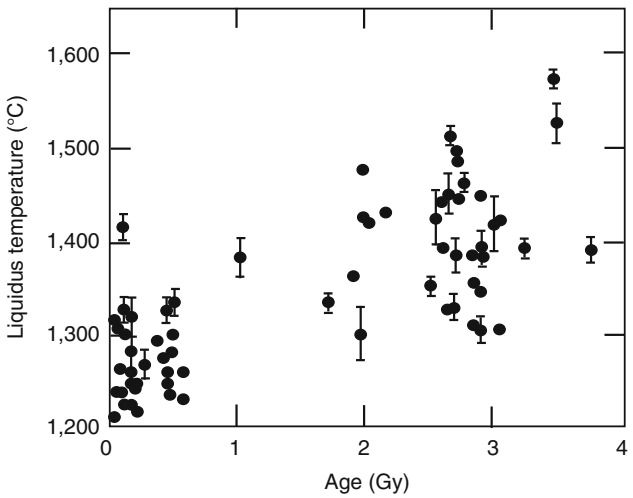
	Stacey and Davis (2008)	Davies (1999)	Jaupart et al. (2007) range (preferred value)
Total heat loss	*42*	*41*	*46*
Continental heat production	8	5	6–8 (7)
Upper mantle		1.3	
Lower mantle		11–27	
Mantle heat production	19	12–28[b]	11–15 (13)
Latent heat (Core differentiation)	1.2	<1	
Mantle differentiation	0.6	0.3	0.3
Gravitational (Thermal contraction)	2.1		
Tidal dissipation		0.1	0.1
Core heat loss	3	5	6–14 (8)
Mantle cooling	10	9[a]	9–23 (18)[c]

[a]Mantle cooling is fixed
[b]Lower mantle heat production is variable and calculated to fit the mantle cooling rate
[c]Mantle cooling is adjusted to fit the other terms in the budget

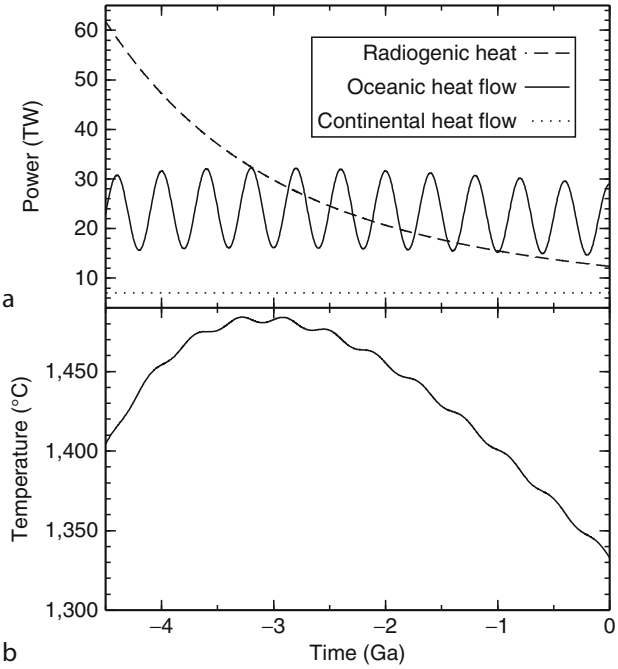
Present and long-term cooling rates

Is the present rate of energy loss equal to the long-term rate? The energy lost through the seafloor depends on the age distribution of the seafloor. It is likely that the energy lost through the seafloor fluctuates with the super continent cycle, and that the present energy loss is different from the long-term mean. The second point is that radioactive heat production in the mantle was higher in the past than now for two reasons: one reason is the run-down of radioelements and the other is the extraction of continental crust that removed radioelements from the mantle. Finally, the heat loss from the mantle depends on the total continental area as much more heat escapes through the seafloor than through the continental lithosphere. The energy flow from the core may also oscillate and have been different in the past, when the frequency of geomagnetic reversals was lower than in recent times.

The relative importance of these effects remains difficult to assess because we do not know exactly how the volume of continental crust changed with time, but Figure 6 presents one plausible thermal history of the mantle based on the assumption that the mantle heat loss oscillates by $\pm$7 TW over a period of 400 My corresponding to the supercontinent cycle.



Energy Budget of the Earth, Figure 5 Liquidus temperatures of basalts from ophiolites and greenstone belts with MORB characteristics as a function of age. Adapted from Abbott et al. (1994).



Energy Budget of the Earth, Figure 6 Effect of fluctuations of oceanic heat loss over a supercontinent timescale of τ_W = 400 My on the secular thermal evolution of the mantle, from Labrosse and Jaupart (2007). The model calculations rely on a simple parametrization of seafloor spreading that leads to a weak secular variation of oceanic heat loss. Heat loss fluctuations due to changes of seafloor spreading are superimposed on the secular trend. These changes occur on a timescale that is short compared to that of the secular trend and leave almost no trace on mantle temperatures. *Top panel* (a): the various contributions to the Earth's energy budget that are used in the calculation. *Bottom panel* (b): predicted evolution of the average mantle temperature through time.

Perspectives

The most promising avenue for improving our understanding of the energy budget is the recent development of neutrino observatories that can detect geo-neutrinos, that is, neutrinos from the decay of U and Th in the Earth (Fiorentini et al., 2005). With sufficient observations of geo-neutrinos, we might be able to determine more precisely the total heat production due to mantle radioactivity.

Summary

The Earth loses energy as heat flows out through its surface. The total energy loss of the Earth has been estimated at 46 $\pm$ 2 TW, of which 14 TW comes through the continents and 32 TW comes from the seafloor. After removing the heat production of radioactive elements in the continental lithosphere, the energy loss from the convecting mantle is 39 TW. Three main sources balance the mantle energy loss: Heat flow from the core, radiogenic heat production in the mantle, and the secular cooling of the mantle. The uncertainty on the exact contribution of each component is much larger than that on the total heat loss. The best estimate of the present cooling rate of the mantle (110 K Gy^{-1}) is higher than the secular cooling rate estimated from petrology. The present cooling rate may be different from its long-term average because of variations in seafloor spreading rates and higher mantle heat production in the past.

Bibliography

Abbott, D., Burgess, L., and Longhi, J., 1994. An empirical thermal history of the Earth's upper mantle. *Journal of Geophysical Research*, **99**, 13835–13850.

Buffett, B. A., 2007. Core-mantle interactions. In Olson, P. (Ed.), *Treatise of Geophysics*. New York: Elsevier, Vol. 8, Chap. 12, pp. 345–358.

Carlson, R. L., and Johnson, H. P., 1994. On modeling the thermal evolution of the oceanic upper mantle: an assessment of the cooling plate model. *Journal of Geophysical Research*, **99**, 3201–3214.

Cogné, J.-P., and Humler, E., 2004. Temporal variation of oceanic spreading and crustal production rates during the last 180 My. *Earth and Planetary Science Letters*, **227**, 427–439.

Crosby, A. G., and McKenzie, D., 2009. An analysis of young ocean depth, gravity and global residual topography. *Geophysical Journal International*, **178**, 1198–1219.

Davies, G. F., 1980. Review of oceanic and global heat flow estimates. *Reviews of Geophysics and Space Physics*, **18**, 718–722.

Davies, G., 1999. *Dynamic Earth: Plates, Plumes, and Mantle Convection*. New York: Cambridge University Press.

Fiorentini, G., Lissia, M., Mantovani, F., and Vannucci, R., 2005. Geo-neutrinos: a new probe of Earth's interior. *Earth and Planetary Science Letters*, **238**, 235–247 (rapid communication).

Harris, R. N., and Chapman, D. S., 2004. Deep-seated oceanic heat flow, heat deficits, and hydrothermal circulation. In *Hydrogeology of the Oceanic Lithosphere*. Cambridge: Cambridge University Press, pp. 311–336.

Hart, S. R., and Zindler, A., 1986. In search of a bulk-Earth composition. *Chemical Geology*, **57**, 247–267.

Jaupart, C., and Mareschal, J. C., 2003. Constraints on crustal heat production from heat flow data. In Rudnick, R. L. (ed.), *The Crust*. New York: Permagon. Treatise on Geochemistry, Vol. 3, pp. 65–84.

Jaupart, C., Labrosse, S., and Mareschal, J. C., 2007. Temperatures, heat and energy in the mantle of the earth. In Bercovici, D. (ed.), *The Mantle*. New York: Elsevier. Treatise on Geophysics, Vol. 7, pp. 253–303.

Javoy, M., 1999. Chemical earth models. *Comptes rendus de l'academie des sciences de Paris*, **329**, 537–555.

Labrosse, S., and Jaupart, C., 2007. Thermal evolution of the Earth: secular changes and fluctuations of plate characteristics. *Earth and Planetary Science Letters*, **260**, 465–481.

Lister, C., 1977. Estimators for heat flow and deep rock properties based on boundary layer theory. *Tectonophysics*, **41**, 157–171.

McDonough, W. F., and Sun, S. S., 1995. The composition of the earth. *Chemical Geology*, **120**, 223–253.

Müller, R. D., Sdrolias, M., Gaina, C., and Roest, W. R., 2008. Age, spreading rates, and spreading asymmetry of the world's ocean crust. *Geochemistry, Geophysics, Geosystems*, **9**, 4006.

Nimmo, F., 2007. Energetics of the core. In Olson, P. (Ed.), *Treatise of Geophysics*. New York: Elsevier, Vol. 8, Chap. 2, pp. 31–65.

Palme, H., and O'Neill, H. S. C., 2003. Cosmochemical estimates of mantle composition. In Carlson, R. W. (ed.), *Mantle and Core*. Oxford: Elsevier. Treatise on Geochemistry, Vol. 2.

Pollack, H. N., Hurter, S. J., and Johnston, J. R., 1993. Heat flow from the earth's interior: analysis of the global data set. *Reviews of Geophysics*, **31**, 267–280.
Rudnick, R. L., and Fountain, D. M., 1995. Nature and composition of the continental crust: a lower crustal perspective. *Reviews of Geophysics*, **33**, 267–310.
Rudnick, R. L., and Gao, S., 2003. Composition of the continental crust. In Rudnick, R. L. (ed.), *The Crust*. New York: Pergamon. Treatise on Geochemistry, Vol. 3, pp. 1–64.
Sclater, J. G., Jaupart, C., and Galson, D., 1980. The heat flow through oceanic and continental crust and the heat loss from the earth. *Reviews of Geophysics*, **18**, 269–311.
Stacey, F. D., and Davis, P. M., 2008. *Physics of the Earth*, 4th edn. New York: Cambridge University Press.
Stein, C. A., and Stein, S., 1994. Constraints on hydrothermal heat flux through the oceanic lithosphere from global heat flow. *Journal of Geophysical Research*, **99**, 3081–3095.
Su, Y. J., 2000. *Mid-Ocean Ridge Basalt Trace Element Systematics: Constraints from Database Management, ICPMS Analysis, Global Data Compilation and Petrologic Modeling*. Unpublished PhD thesis, New York 1, Columbia University, 569 pp.
Williams, D. L., and von Herzen, R. P., 1974. Heat loss from the earth: new estimate. *Geology*, **2**, 327–330.

Cross-references

Core Dynamo
Core-Mantle Coupling
Geodynamics
Heat Flow Measurements, Continental
Heat Flow, Seafloor: Methods and Observations
Lithosphere, Continental: Thermal Structure
Lithosphere, Oceanic: Thermal Structure
Mantle Convection
Mantle Plumes
Radiogenic Heat Production of Rocks

ENERGY PARTITIONING OF SEISMIC WAVES

Kalachand Sain
National Geophysical Research Institute, Council of Scientific & Industrial Research (CSIR), Hyderabad, India

Synonyms

Partitioning – conversion

Definition

When an incoming P wave strikes an interface between two isotropic homogeneous elastic media at some angle other than the vertical, some of the energy is converted into S wave which gets reflected and transmitted in the same way as the P wave. This is known as the energy partitioning of seismic wave.

Snell's Law: The angles of incidence, reflection and transmission as shown in Figure 1 are related by Snell's law, which says that the ratio of sines of angles of incidence, reflection, and transmission to velocities of respective media across the interface is a constant. In seismology, this constant is defined as the ray parameter, R_P as

$$\frac{\sin\theta_1}{V_{P1}} = \frac{\sin\theta_2}{V_{P2}} = \frac{\sin\varphi_1}{V_{S1}} = \frac{\sin\varphi_2}{V_{S2}} = R_P \quad (1)$$

where V_P, V_S and ρ are the P-wave velocity, S-wave velocity and density, and subscripts 1 and 2 represent the first and second layers respectively.

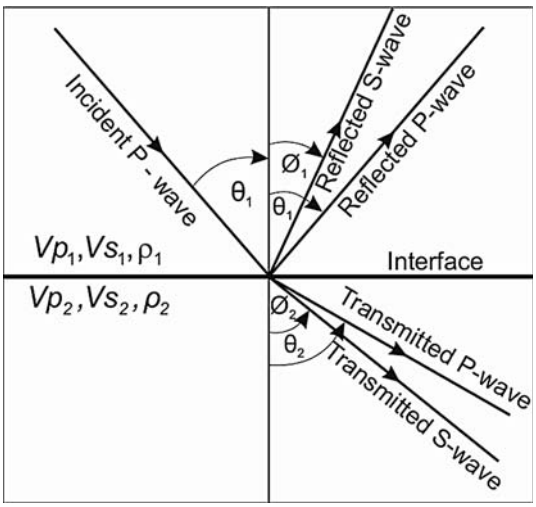
Reflection and transmission coefficients: A reflection coefficient describes either the amplitude or the energy of a reflected wave relative to an incident wave and similarly a transmission coefficient describes either the amplitude or the energy of a transmitted wave relative to an incident wave. When a plane wave strikes a horizontal interface vertically (i.e., at normal incidence), the reflection coefficient, R_C, and the transmission coefficient, T_C, are defined as

$$R_C = \frac{Z_2 - Z_1}{Z_2 + Z_1} \quad (2)$$

$$T_C = \frac{2Z_1}{Z_2 + Z_1} \quad (3)$$

Where Z_1 and Z_2 are the acoustic impedance (P-wave velocity multiplied by density) of the first and second media respectively.

Zoeppritz equation: When the seismic ray obliquely strikes the interface (Figure 1), separated by three elastic parameters (V_P, V_S, and ρ) on either side, the R_C and T_C as expressed by Equations 2 and 3 are not valid. The reflection and transmission coefficients or amplitudes of both P-P and mode converted P-S waves as a function of angle can be derived using the Zoeppritz equations (Zoeppritz, 1919) that can be easily expressed through a matrix (Waters, 1987) as



Energy Partitioning of Seismic Waves, Figure 1 Ray diagram for the partition of a plane wave energy at a horizontal interface.

$$Q = P^{-1}R \tag{4}$$

Where P is a 4 × 4 square matrix and, R and Q are 4 × 1 column matrices whose elements are shown as

$$\mathbf{P} = \begin{bmatrix} \sin\theta_1 & \cos\varphi_1 & -\sin\theta_2 & \cos\varphi_2 \\ -\cos\theta_1 & \sin\varphi_1 & -\cos\theta_2 & -\sin\varphi_2 \\ \sin 2\theta_1 & \frac{V_{P1}}{V_{S1}}\cos 2\varphi_1 & \frac{\rho_2}{\rho_1}\frac{V_{S2}^2}{V_{S1}^2} 2\sin\theta_1\cos\theta_2 & -\frac{\rho_2}{\rho_1}\frac{V_{P1}V_{S2}}{V_{S1}^2}\cos 2\varphi_2 \\ \cos 2\varphi_1 & -\frac{V_{S1}}{V_{P1}}\sin 2\varphi_1 & -\frac{\rho_2}{\rho_1}\frac{V_{P2}}{V_{P1}}\cos 2\varphi_2 & -\frac{\rho_2}{\rho_1}\frac{V_{S2}}{V_{P1}}\sin 2\varphi_2 \end{bmatrix} \tag{5}$$

$$Q = [\mathrm{R_{PPR}}\ \mathrm{R_{PSVR}}\ \mathrm{R_{PPT}}\ \mathrm{R_{PSVT}}]^{\mathrm{T}} \text{ and}$$
$$\mathrm{R} = [-\sin\theta_1 \ -\cos\theta_1 \sin 2\theta_1 \ -\cos 2\phi_1]^{\mathrm{T}} \tag{6}$$

$\mathrm{R_{PPR}}$ and $\mathrm{R_{PSVR}}$ are the reflection coefficients and, $\mathrm{R_{PPT}}$ and $\mathrm{R_{PSVT}}$ are the transmission coefficients of P-P and mode converted P-S waves respectively.

Figure 2a displays the coefficients of reflections and transmissions for P-P and converted P-S waves as a function of incident angles (θ) as per Equation 4 for the model as shown in the figure itself. At zero angle or offset, the R_C for P-P wave is mainly governed by the P-wave velocity but at large offsets its value depends on both the P and S wave velocities. The nature and amplitudes of curves vary with the elastic parameters across the interface (Figure 2b). These elastic parameters again depend on the lithology and fluids present within.

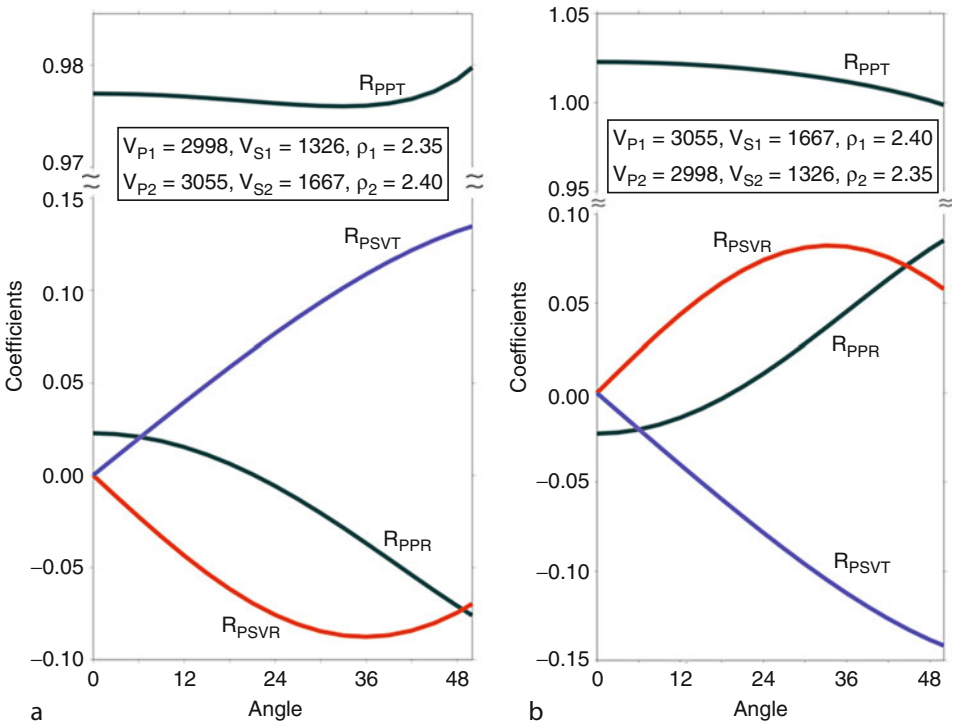
Approximation to the Zoeppritz equation led to an elegant method of reducing these curves to just two parameters. According to Shuey (1985), the P-wave reflection coefficient, $R_C(\theta)$ for an incidence angle (θ) up to ~30° can be expressed as

$$R_C(\theta) \approx A + B\ sin^2\theta \tag{7}$$

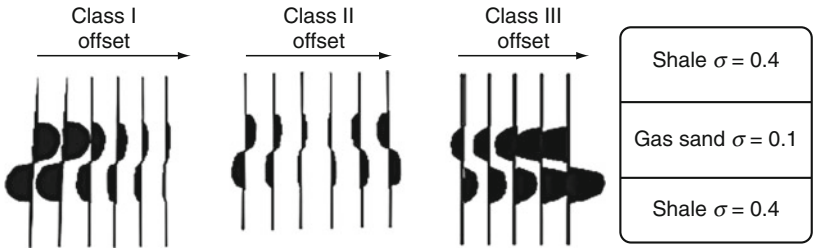
where,

$$A = \frac{1}{2}\left(\frac{\Delta V_P}{V_P} + \frac{\Delta\rho}{\rho}\right),\ B = AA_0 + \frac{\Delta\sigma}{(1-\sigma)^2} \tag{8}$$

$$A_0 = B_0 - 2(1+B_0)\frac{1-2\sigma}{1-\sigma}, B_0 = \frac{\Delta V_P/V_P}{\Delta V_P/V_P + \Delta\rho/\rho},$$
$$V_P = \frac{(V_{P2}+V_{P1})}{2}, V_S = \frac{(V_{S2}+V_{S1})}{2}, \rho = \frac{(\rho_2+\rho_1)}{2},$$
$$\Delta V_P = (V_{P2}-V_{P1}), \Delta V_S = (V_{S2}-V_{S1}),$$
$$\Delta\rho = (\rho_2-\rho_1), \Delta\sigma = (\sigma_2-\sigma_1), \sigma = \frac{(\sigma_2+\sigma_1)}{2},$$
$$\sigma_i = \frac{(.5V_{Pi}^2 - V_{Si}^2)}{(V_{Pi}^2 - V_{Si}^2)}$$



Energy Partitioning of Seismic Waves, Figure 2 (**a**) Reflection and transmission coefficients as a function of incidence angles for the model shown inside. (**b**) reflection and transmission coefficients as a function of incidence angles for the model shown inside in which the layer parameters are interchanged from the model in (**a**).



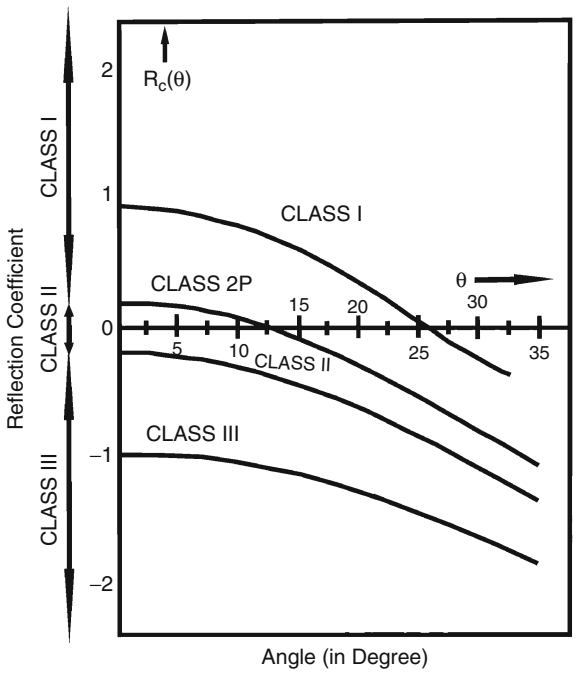
Energy Partitioning of Seismic Waves, Figure 3 Seismogram showing AVO response for three classes of gas sand anomalies, and a typical reservoir showing the Poisson ratio.

Equations 7 and 8 are valid up to high A for a negative impedance contrast across an interface, and to about $A = 0.2$ for a positive impedance contrast. Since 30° angle corresponds to the maximum angle that we generally encounter in normal seismic data acquisition, this approximation is widely used to determine the amplitude versus offset (AVO) attributes such as the intercept, A and the gradient, B for the interpretation of seismic data.

AVO response for a realistic earth model: The reflections from various subsurface interfaces exhibit a wide range of AVO characteristics that depend upon the rock types and fluid contents. From the field data, the amplitudes and offsets of reflected phases can be converted into reflection coefficients and angles, and by modeling using the Zoeppritz equation we can derive various elastic parameters (Ojha and Sain, 2007, 2008) or Poisson ratio (σ) that can be used as a direct indicator for hydrocarbons, as inclusion of oil or gas in the rock lowers the σ. Any overpressure in the trapped gas can also be known from AVO anomaly (Tinivella, 2002) that helps in avoiding drilling hazard. In oil industries, the AVO attributes have gained considerable popularity for lithology identification and reservoir characterizations (Castagna and Backus, 1993; Castagna et al., 1998) due to the fact that seismic amplitudes at interfaces are affected by the variations of physical properties just above and below the interfaces. In a recent paper, Ojha et al. (2010) have demonstrated through a field example that a plot of A versus B coupled with a rock physics can be used as a rapid tool for detecting free-gas below a bottom simulating reflector, and quantifying the amount of gas-hydrate and free-gas along a seismic reflection profile.

The gas-sand detection is the most promising application of AVO analysis. The characteristically low V_P/V_S ratio or σ of gas sand is expected to be differentiated from other low-impedance layers, such as coals and porous brine or oil sands. Rutherford and Williams (1989) have defined three distinct classes of gas-sand AVO anomalies (Figures 3 and 4) mainly based on the intercept, A (normal incidence reflection coefficient) and gradient, B attributes.

The Class I is a high impedance sand anomaly and occurs when the normal-incidence P-wave reflection coefficient is strongly positive and shows a strong amplitude decrease with offset and a possible phase change at far offset. The Class II anomaly occurs when the gas sand has similar impedance to the neighboring shale. For the gas



Energy Partitioning of Seismic Waves, Figure 4 AVO curves for three classes of gas sand anomalies.

sand scenario, the normal incidence reflection coefficient is close to zero and becomes negative at larger offsets. The Class III is the classic Gulf of Mexico AVO anomaly where gas sand has low impedance compared to shale. The normal incidence reflection coefficient is negative and becomes more negative as offset increases. The "bright spot" is associated with the Class III AVO anomaly and can be used as a direct indicator for hydrocarbons.

If the rock overlying the reservoir has a velocity appreciably lower than that of the reservoir itself (e.g., carbonate reservoir capped by shale), the effect of hydrocarbon is to decrease the impedance contrast and reduce the reflection coefficient producing a "dim spot." Class I and Class II AVO anomalies can produce the reverse effect. When the rock overlying the reservoir has a velocity slightly smaller than that of the reservoir rock, lowering the impedance of reservoir rock by hydrocarbons may invert the sign of the reflection, producing a polarity

reversal. The sand types and presence of gas can be identified from the classical A-B crossplot.

Summary

Because of energy partitioning, we can record both P- and S waves using multi-component sensors, and derive both P- and S-wave seismic velocities and their ratio (V_P/V_S), study of which provides more accurate information on lithology and fluid content than that obtained either by P- or S-wave velocity alone.

Acknowledgments

I am grateful to the Director, NGRI for his permission. The Ministry of Earth Sciences, Government of India is thankfully acknowledged for sponsoring the gas-hydrate research at NGRI.

Bibliography

Castagna, J. P., and Backus, M. M., 1993. Offset dependent reflectivity: Theory and Practice of AVO analysis. *Society of Exploration Geophysics*, **8**, 303–313.

Castagna, J. P., Swan, H. W., and Foster, D. J., 1998. Framework for AVO gradient and intercept interpretation. *Geophysics*, **63**, 948–956.

Ojha, M., and Sain, K., 2007. Seismic velocities and quantification of gas hydrates from AVA modeling in the western continental margin of India. *Marine Geophysical Researches*, **28**, 101–107.

Ojha, M., and Sain, K., 2008. Appraisal of gas hydrates/free-gas from Vp/Vs ratio in the Makran accretionary prism. *Marine and Petroleum Geology*, **25**, 637–644.

Ojha, M., Sain, K., and Minshull, T. A., 2010. Assessment of gas hydrates saturation in the Makran accretionary prism using the offset dependence of seismic amplitudes. *Geophysics*, **75**(2), C1–C6.

Rutherford, S. R., and Williams, R. H., 1989. Amplitude-versus-offset variations in gas sands. *Geophysics*, **54**, 680–688.

Shuey, R. T., 1985. A simplification of Zoeppritz equations. *Geophysics*, **50**, 609–614.

Tinivella, U., 2002. The seismic response to overpressure versus gas hydrate and free gas concentration. *Journal of Seismic Exploration*, **11**, 283–305.

Waters, K. H., (ed.), 1987. Reflection Seismology: A tool for energy resource exploration, A Wiley Interscience Publication, New York, pp. 6–42.

Zoeppritz, K., 1919. On the reflection and propagation of seismic waves at discontinuities. In Goettinger Nachrichten, I. (ed.), *Erdbebenwellen VII B*, pp. 66–84.

Cross-references

Seismic Data Acquisition and Processing
Seismic, Reflectivity Method
Seismic, Waveform Modeling and Tomography
Single and Multichannel Seismics

EQUATORIAL ELECTROJET

Archana Bhattacharyya
Indian Institute of Geomagnetism, Navi Mumbai, India

The equatorial electrojet (EEJ) is a narrow band of intense electric current that flows along the dip equator in the day side ionospheric E region with peak current density at an altitude of around 105 km.

Cause

Electrical conductivity of the partially but significantly ionized upper atmosphere of the Earth is anisotropic due to the presence of Earth's magnetic field. In the ionospheric E region, movement of electrons in response to an electric field is influenced by the geomagnetic field, whereas for the heavier ions, collisions with neutral particles have much greater influence. This difference in the behavior of electrons and ions gives rise to a large Hall conductivity in the E region. Atmospheric tides, driven mainly by solar heating, cause the neutral atmosphere to undergo global-scale daily oscillations, which set the electrically conducting fluid in the ionosphere into motion thereby generating an electric field in the E region due to dynamo action in the presence of the geomagnetic field. At the dip equator, where the geomagnetic field is horizontal and northward directed, an eastward electric field in the magnetic east-west direction drives a downward Hall current, which is restricted due to the limited vertical extent of the E region. This causes charges to accumulate producing an upward polarization electric field much larger than the original eastward electric field in a region where the Hall conductivity is much larger than the Pedersen conductivity, and the upward electric field now drives an eastward Hall current that greatly exceeds the original eastward Pedersen current. Together these eastward currents constitute the normal equatorial electrojet (EEJ) (Forbes, 1981).

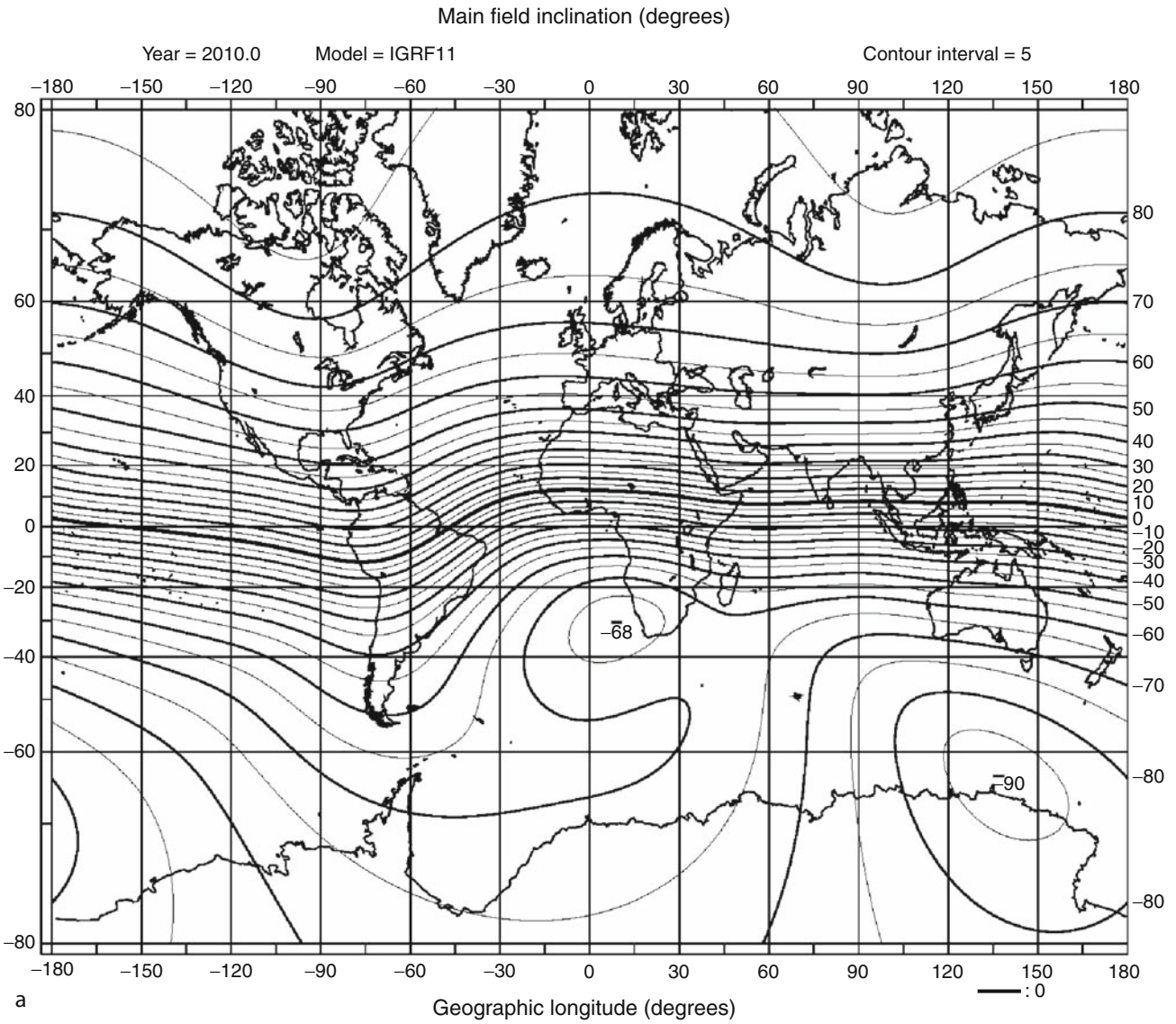
Observations

Comparison of geomagnetic field variations recorded at observatories located near the dip equator with data from low latitude observatories outside the influence of the EEJ but in the same longitude zone, show a typical enhancement of about 40–100 nT in the horizontal component of the geomagnetic field during daytime at the equatorial location, due to the EEJ. Figure 1a shows the location of the dip equator at different geographic longitudes computed from International Geomagnetic Reference Field model IGRF11. Figure 1b gives the locations of a chain of magnetic observatories in the Indian longitude region located at different latitudes, which extend northward from the dip equator to the focus of the solar quiet (Sq) current system that also flows in the day side E region of the ionosphere driven by E region dynamo electric fields generated by atmospheric tides, while Figure 1c depicts the average pattern of daily variation of the horizontal component of the geomagnetic field measured at these locations on magnetically quiet days of November 2001. The excess current that flows in the dip equatorial region of the ionspheric E layer as a consequence of the geometry of Earth's main magnetic field is the EEJ. In situ measurements by rockets have provided information about the current density and electric field profiles in the EEJ (Pfaff et al., 1997). Satellite

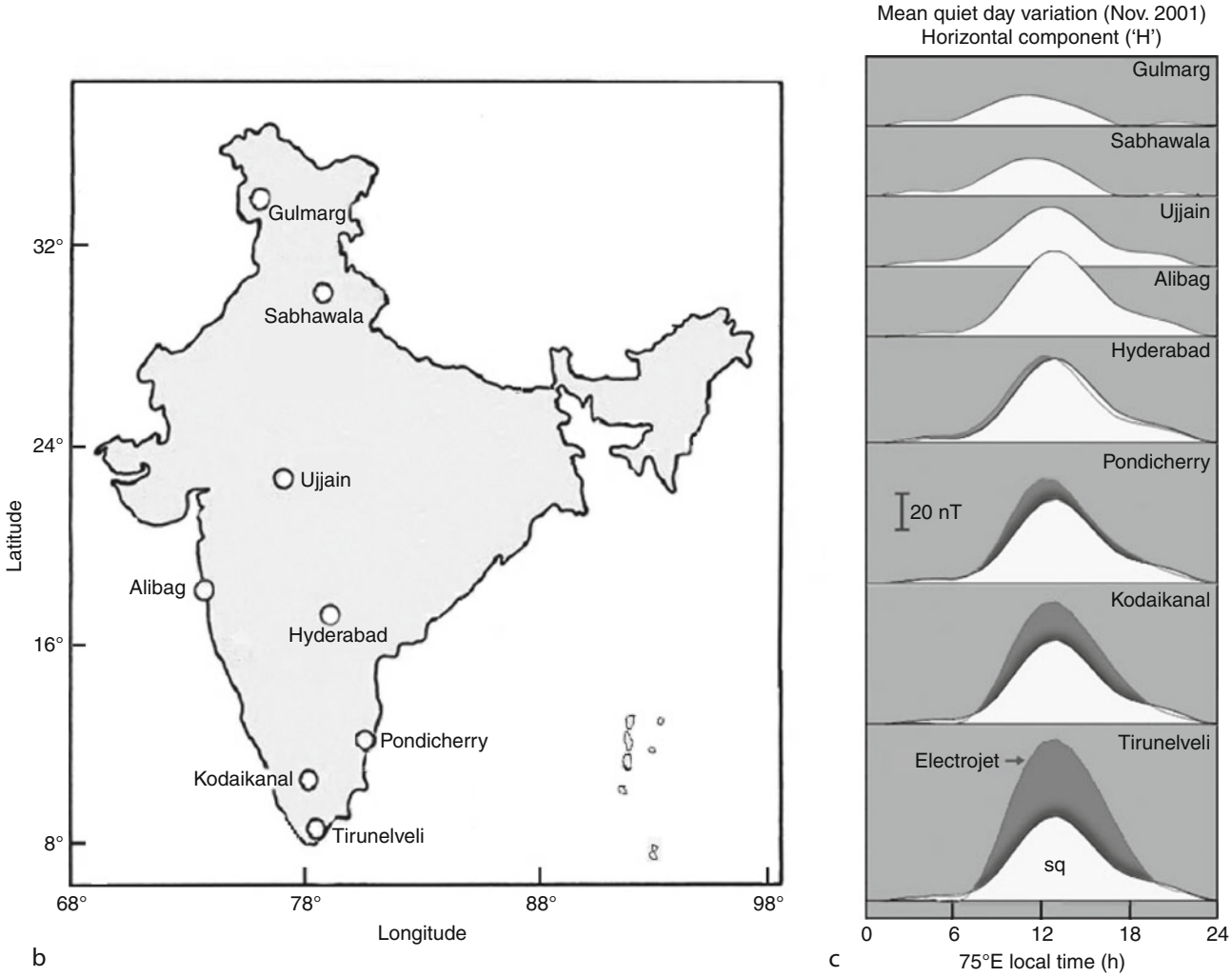
observations of magnetic field produced by the EEJ have yielded information about longitudinal variations in the spatial features and strength of the EEJ. For a normal EEJ, the same pattern of steep rise in current strength in the morning hours reaching a maximum between 10:30 and 12:00 LT, and then a gradual decrease to insignificant values around 18:00 LT is seen at all longitudes. Assuming the EEJ to flow in a thin sheet, on the average the current density estimated from CHAMP satellite observations is 0.19 A/m, while the latitudinal extent of the EEJ at noontime has been estimated to be about ±2 degrees around the dip equator, where the current peaks (Lühr et al., 2004). It is yet to be established how return currents close the EEJ. At a fixed longitude, the EEJ shows considerable day-to-day variation, and also seasonal and solar cycle dependence. On some days, the magnetic signatures recorded at observatories or by satellites indicate a daytime current flowing in the westward direction. This is called the counter electrojet, a phenomenon not yet understood (Rastogi, 1989).

Uses of EEJ observations

The electric field in the equatorial ionospheric E region plays a major role in determining the spatial distribution of ionospheric plasma throughout the low latitude ionosphere, which affects the operation of satellite-based communication and navigation systems such as the Global Positioning System (GPS). The EEJ is an indicator of the



Equatorial Electrojet, Figure 1 (Continued)



Equatorial Electrojet, Figure 1 (**a**) Contour plots of the main field inclination (I) computed for the year 2010.0 using model IGRF11 shows the location of the dip equator (I = 0) at different longitudes (http://wdc.kugi.kyoto-u.ac.jp); (**b**) Locations of some magnetic observatories in the Indian longitude region, extending northward from the dip equator (near Tirunelveli) to the focus (near Gulmarg) of the solar quiet (Sq) current system; (**c**) Mean quiet day variation of the horizontal component (H) of the geomagnetic field measured at these observatories during November 2001 shows the enhancement in H near the dip equator due to the equatorial electrojet.

state of this electric field during daytime. In the processing of data collected in magnetic surveys as a part of geophysical exploration for natural resources, it is necessary to remove the diurnal variation in the geomagnetic field due to the EEJ. The EEJ is also a natural source for electromagnetic induction studies. However, its latitudinal extent sets a limit on the depth within the Earth to which electrical conductivity can be determined using temporal variations of the EEJ.

Bibiliography

Forbes, J. M., 1981. The equatorial electrojet. *Reviews of Geophysics and Space Physics*, **19**, 469–504.

Lühr, H., Maus, S., and Rother, M., 2004. Noon-time equatorial electrojet: Its spatial features as determined by the CHAMP satellite. *Journal of Geophysical Research*, **109**, A01306, doi:10.1029/2002 JA 009656.

Pfaff, R. F., Acuña, M. H., Marionni, P. A., and Trivedi, N. B., 1997. DC polarization electric field, current density, and plasma density measurements in the daytime equatorial electrojet. *Geophysical Research Letters*, **24**, 1667–1670.

Rastogi, R. G., 1989. The equatorial electrojet. In Jacobs, J. (ed.), *Geomagnetism*. San Diego: Academic, Vol. 3, pp. 461–525.

Cross-references

Magnetic Storms and Electromagnetic Pulsations

F

FRACTALS AND CHAOS

Vijay P. Dimri[1], Ravi P. Srivastava[2], Nimisha Vedanti[2]
[1]Research, Innovation & Incubation Center, Gujarat Energy Research and Management Institute, Gandhinagar, Gujarat, India
[2]Fractals in Geophysics, National Geophysical Research Institute, Hyderabad, India

Synonyms

Long-range dependence and disorder; Power law; Randomness; Scaling; Self-affine; Self-similar

Definition

Fractals. The objects having fractional dimension are termed as fractal. Also, such objects are self-similar in nature, thus their shape is independent of the scale and remains same under dilatation or contraction. The concept is extended to the time series analysis, and those time series in which power spectrum follows power law are termed as fractal time series.

Chaos. It is a state of extreme disorder. Behavior of chaotic systems is highly sensitive to initial conditions. Small change in initial conditions, even due to rounding off errors in computation can yield widely diverging outcomes for chaotic systems, rendering long-term prediction impossible in general. The chaos can be quantified by Lyapunov exponent if it is positive.

Introduction

The origin of fractals dates back to the eighteenth century in the study of non-differentiable functions by Karl Weierstras, Georg Cantor, and Felix Hausdorff, however, the term fractal was coined by Mandelbrot (1967, 1983) in a paper published in science "How long is the coast of Britain?" and later in his book *Geometry of Nature*. Fractals and Chaos are relatively new mathematical innovations, which find their applications in several branches of pure and applied sciences. Applications of fractals and chaos have been widely explored in solid earth geophysics and its successful applications are demonstrated in many areas of geophysics, namely: gravity, magnetic modeling (Turcotte, 1997; Dimri, 2005), seismic inversion, earthquake frequency–magnitude relationship (Ouchi and Uekawa, 1986; Dimri et al., 2005), analysis of reflectivity sequence (Painter et.al., 1995; Toverud et al., 2001), and fractal representation of soil structures to characterize soil water properties in terms of porosity and particle size distribution (Rieu and Sposito, 1991). Another important study is simulation of fractures using fractal method (Borodich, 2000; Tanaka et al., 2001). A new technique is proposed by Srivastava and Dimri (2007) that uses modified Voronoi tessellation to model the complex geological structures. Further, fractal or scaling behavior of time series has been widely used in analysis and modeling of geophysical time series (Todoeschuck and Jensen, 1988, 1989). Chaos has been used in studies related to atmosphere, climate, and earthquake studies (Grassberger and Procaccia, 1983; Srivastava, 2005).

Basic concepts

Power law (scaling behavior)

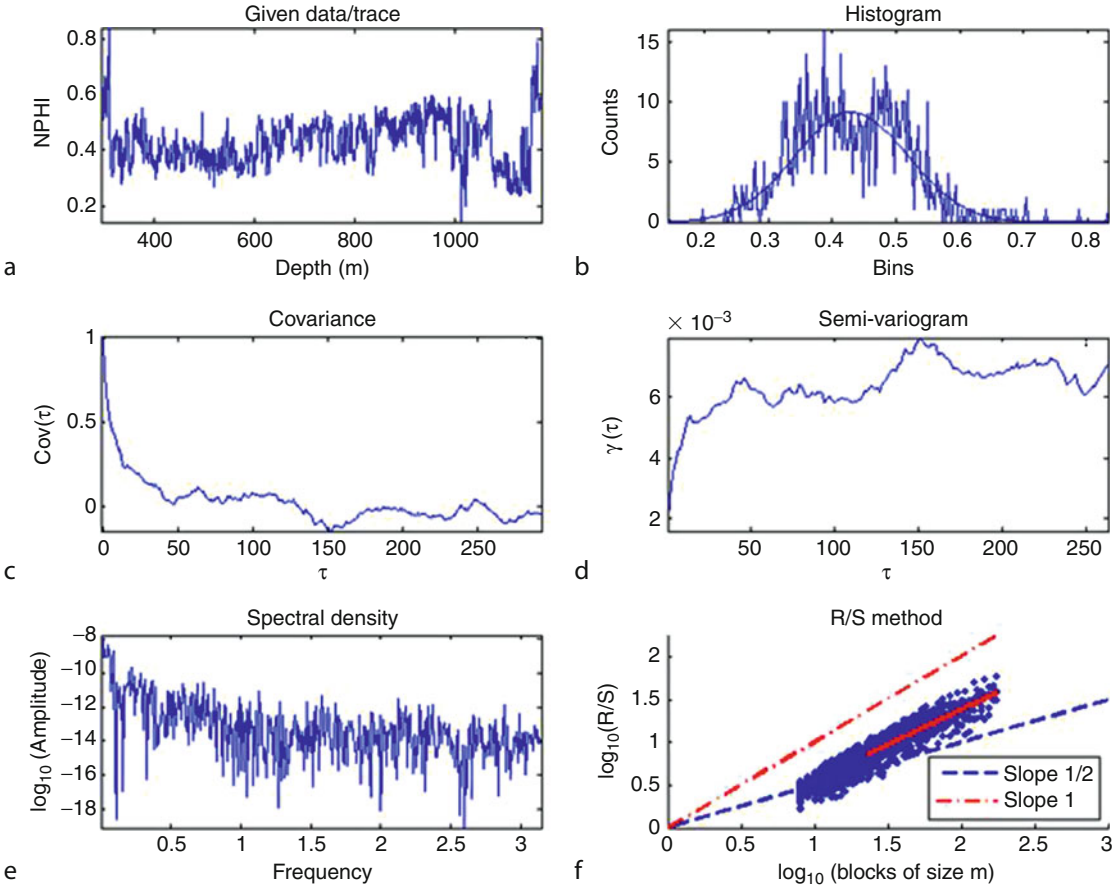
Spectral analysis of fractal time series shows that power (P) is proportional to frequency (f) raise to power a real exponent (β), termed as scaling exponent (Figure 1a, e). In mathematical form it is represented as:

$$p(f) \propto f^{-\beta} \tag{1}$$

This behavior is also known as scaling law.

The fractal analysis of time series is fundamentally based on the concept of fractional Brownian motion and fractional Gaussian noise. The fractal time series, also

Harsh K. Gupta (ed.), *Encyclopedia of Solid Earth Geophysics*, DOI 10.1007/978-90-481-8702-7,



Fractals and Chaos, Figure 1 Test of fractal behavior of the time series: (**a**) given Neutron porosity log, (**b**) histogram of the well log, (**c**) covariance showing power law behavior, (**d**) semi-variogram showing power law behavior, (**e**) log power spectrum showing power law, and (**f**) Hurst coefficient ($H = 0.80$) showing fractional Gaussian noise (fGn) characteristic of the well log.

known as time series having long memory are characterized by fractional Gaussian noise (fGn). A time series is said to follow fGn characteristic if (1) its histogram is fat tailed Gaussian, (2) covariance, semi-variogram, and power spectrum follow power law behavior, i.e., values of log power spectrum either decays or increases with lag and covariance behaves as a mirror image of semi-variogram, and (3) Hurst coefficient lies within (0.5–1) as shown in Figure 1. A popular method of Hurst coefficient estimation is Re-Scaled range analysis (R/S analysis). Below we discuss the definition and basic concepts of fractional Brownian motion and fractional Gaussian noise.

Fractional Brownian motion (fBm)

Fractional Brownian motion is generalization of classical Brownian motion (cBm). The fractal nature of fBm can be understood from the following mathematical definition. Assume that $f(t)$ is a single valued continuous function of time/space (t). Let the increments (or difference of adjacent values) $\delta f = f(t_2) - f(t_1)$ follow Gaussian distribution (Molz and Boman, 1993) with the following properties:

$$\begin{aligned} &\mathrm{E} < f(t_2) - f(t_1) \geq 0, \text{ and} \\ &\mathrm{E}[f(t_2) - f(t_1)]^2 \geq \alpha |t_2 - t_1|^{2H}, \end{aligned} \tag{2}$$

where, $\mathrm{E}<X>$ is the mean or expected value of X, α is constant of proportionality, and H is an exponent known as Hurst coefficient (Mandelbrot and Van Ness, 1968; Voss, 1985a, b) lying between 0 and 1 ($0 \leq H \leq 1$). The expected values are obtained by averaging over many increments of size $|t_2 - t_1|$. The Equation 2 holds for any value of time increment $\delta t = |t_2 - t_1|$. It is obvious from Equation 2 that for $H = 0.5$, $\mathrm{E}<[f(t_2) - f(t_1)]^2 \geq \alpha |t_2 - t_1|$, so the fractional Brownian motion reduces to classical Brownian motion (cBm).

Fractional Gaussian noise (fGn)

The concept of fractional Gaussian noise is intertwined with the fractional Brownian motion. The detailed discussion and criteria to distinguish between them are explained

in the paper by Hewett (1986) and Hardy and Beier (1994). Here we describe in brief the relationship between fBm and fGn.

Let, $x_1, x_2, x_3, \ldots x_n$ be the sequence of numbers separated by time interval δt that represents a fBm process, then one can generate a corresponding fGn or time derivative process by computing the sequence $[(x_2 - x_1), (x_3 - x_2), (x_4 - x_3), \ldots (x_n - x_{n-1})]/\ \delta t$. Conversely, if one starts with a fGn sequence of numbers given by $y_1, y_2, y_3, \ldots y_n$, then, numerical integration will yield the corresponding fBm sequence given by:

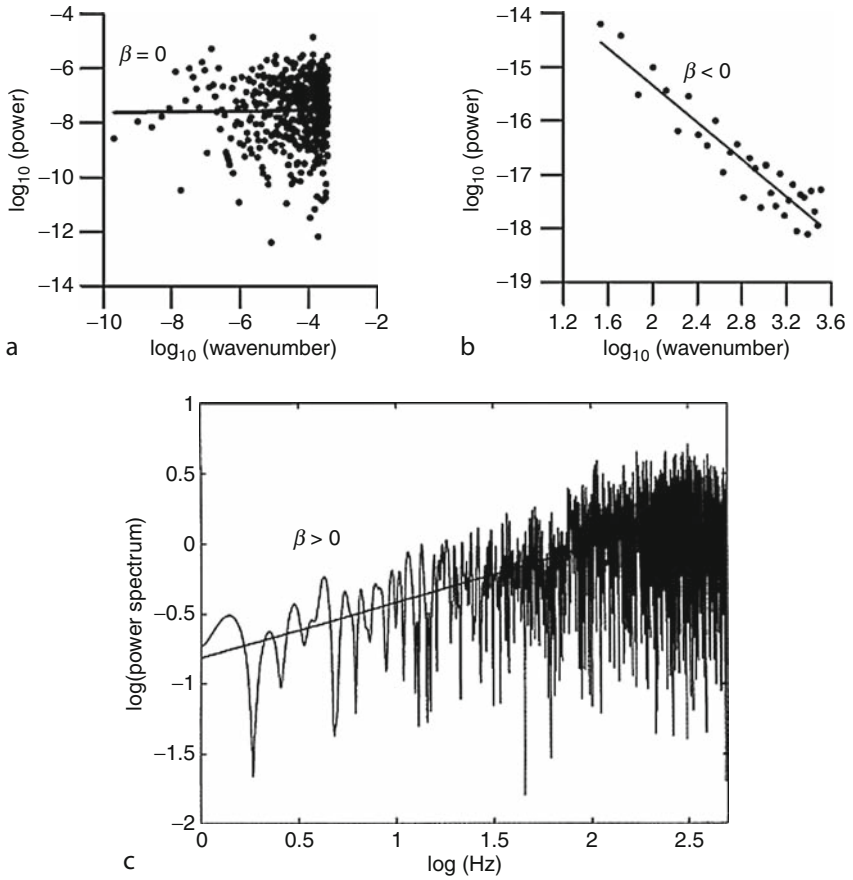
$$[y_1, (y_1 + y_2), (y_1 + y_2 + y_3), \ldots (y_1 + y_2 + y_3 \ldots + y_n)].$$

Fractal signal analysis

Geophysical signals are often in the form of time/space series that can be characterized by combination of stochastic component, trend component, and periodic component (Malamud and Turcotte, 1999). To quantify the stochastic component of the time series it is necessary to specify the statistical distribution of values and persistence. There exist a variety of techniques to quantify the strength of persistence, but the most commonly used is the spectral analysis, where the log-log plot of power spectrum of time series is plotted against frequency (or wave number in case of space series) and the value of slope, known as scaling exponent (say β), gives an estimation of persistence. Persistence study is useful to understand the nature of the phenomenon. Depending upon the value of scaling exponent, persistence can be characterized as weak or strong. Scaling exponent controls the balance of high and low frequencies and determines the degree of smoothness or correlation of the time series.

Analyzing geophysical time series

Persistence of the time series is a measure of correlation between adjacent values of the terms in time series. The uncorrelated or random data with zero persistence is the white noise. The case $\beta = 0$ (Figure 2a) has a power independent of frequency and is the familiar case of white noise. Values of such a time series are uncorrelated and completely independent of one another. The time series is anti-persistent if adjacent values are anti-correlated. When $\beta < 0$ (Figure 2b), the series is anti-correlated and each successive value tends to have the opposite sign. Noise with $\beta = -1$ is sometimes called flicker noise, while that with $\beta = -2$ is Brownian noise. The time series is persistent if adjacent values are positively correlated to



Fractals and Chaos, Figure 2 Power spectrum of (**a**) random data, (**b**) anti-persistent data, and (**c**) persistent data (after Dimri, 2005).

each other. For $\beta > 0$ (Figure 2c), the series is positively correlated, that is, persistent.

Fractal behavior of various physical properties

The Earth's inherent complexity makes it difficult to infer the location, distribution, and structure of rock types, grain size distribution, material strength, porosity, permeability, etc., from the geophysical observations; these characteristics are often inferred from the distribution of fundamental physical properties such as density, electrical conductivity, acoustic impedance, and many others. For geophysical study, the physical response of the Earth can be approximated by convolution model (Dimri, 1992). Extraction of useful information from the observations made over the surface needs advanced processing and interpretation techniques. Since physical properties of the earth system follow fractal (scaling) rather than hitherto assumed random distribution, the interpretation techniques of geophysical exploration are reformulated for fractal (scaling) geology (Dimri, 2005).

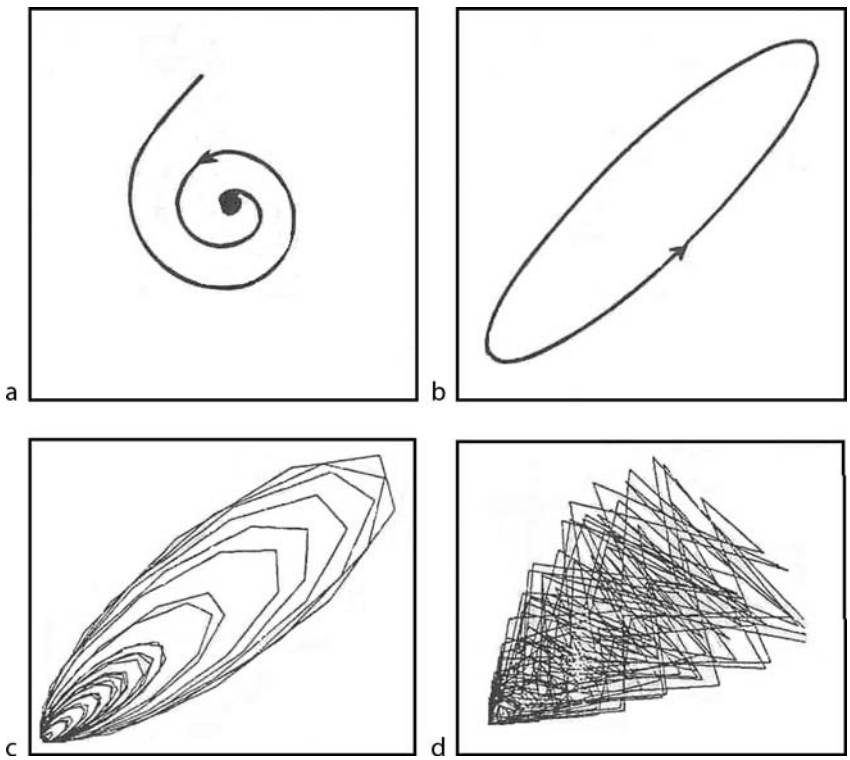
Fractal dimension and seismicity distribution

The fractal theory has led to the development of a wide variety of physical models of seismogenesis including nonlinear dynamics. Fractal concepts can be efficiently used to characterize the seismicity pattern of a region. The fractal nature of the spatial distribution of earthquakes was first demonstrated by Kagan and Knopoff (1980), Hirata and Imoto (1991), and Hirabayashi et al. (1992). The hypocenter distribution data suggests that the changes in fractal dimension could be a good precursor parameter for earthquakes as it is a measure of the degree of clustering of seismic events. A change in fractal dimension corresponds to the dynamic evolution of the states of the system. Generally the values of b-value (half of fractal dimension) cluster around 1 when the system is relatively stable and decreases to lower values around 0.3 prior to the failure. The decrease in value of fractal dimension before a big earthquake is observed by several authors (Ouchi and Uekawa, 1986, De Rubies et al., 1993). Hirata et al. (1987) showed that the change in fractal dimension with time is largely dependent on the crustal conditions of the study region.

Chaos

'And now a bubble bursts, and now a world'
An essay on Man, Alexander Pope

The concept of chaos was proposed by Lorenz, who advocated that a small perturbation in the coupled system can amplify to a great deal in some other part of the system. The chaotic systems are known to have finite fractal dimension (Moharir, 2000). The chaotic behavior of the system can be assessed by the quantitative estimate of



Fractals and Chaos, Figure 3 Trajectory of the types of attractors (**a**) point attractor (**b**) periodic limit cycle (**c**) quasi-periodic, and (**d**) chaotic attractor (after Tiwari, 2000).

Lyapunov exponent. The system is said to be chaotic, if Lyapunov exponent is positive (>0).

Various evolutionary fluctuations exhibit the features of a dissipative dynamical system with possible chaos/attractors in phase space (Tiwari, 2000). If all the trajectories of evolutionary processes converge towards a subset of phase space irrespective of their initial conditions, then the subset is called an "attractor" (Tsonis and Elsner, 1988). The nature of attractor is characterized by its fractal dimension. A "strange attractor" will have a non-integer (fractal) dimension and can be represented by a more complex form in a phase space. Also, such an attractor is characterized by the chaotic trajectories and a broadening of the frequency spectrum. Evidence of chaos/strange attractor has been reported in many global geophysical and planetary system dynamics. The shape of the attractor depends on the nature of the physical process. The shape of most common four kinds of attractors, namely, point attractor, periodic limit cycle, quasi-periodic, and chaotic attractors is shown in Figure 3. The fractal dimension of a chaotic system (strange attractor) indicates the number of ordinary differential equations needed to model the process, or in other words, it is indicative of the number of variables required for almost complete description of the dynamic system (Tsonis and Elsner, 1988).

Summary

Geophysical interpretation techniques have been reformulated after the findings from the analysis of physical properties recorded in different deep bore wells worldwide, which shows that physical properties of the earth system follow fractal instead of random distribution (Dimri, 2005). Recent research in solid earth geophysics employing fractals and chaos includes modeling of the earth's magnetic and gravity fields, where it has been shown that rocks have scaling stratification (Lovejoy et al., 2008; Lovejoy and Schertzer, 2007). Another, significant research includes simulation of fGn series, which helps to improve the inversion results by way of providing a suitable initial model (Srivastava and Sen, 2009, 2010). In atmospheric sciences, the interesting research is done to model the atmospheric dynamics, which repeats scales after scales and behaves like a cascade (Lovejoy et al., 2009). There are several interesting applications of chaos in climate and rainfall modeling (Hosking, 1984; Sivakumar, 2000).

Bibliography

Bansal, A. R., and Dimri, V. P., 2001. Depth estimation from the scaling power spectral density of nonstationary gravity profile. *Pure and Applied Geophysics*, **158**, 799–812.

Beltrami, H., and Mareschal, J. C., 1993. Strange seismic attractor? *Pure and Applied Geophysics*, **141**, 71–81.

Borodich, F. M., 2000. Self-similar models and size effect of multiple fractures. *Fractals*, **9**(1), 17–30.

De Rubies, V., Dimitriu, P., Papa, D. E., and Tosi, P., 1993. Recurrent patterns in the spatial behavior of Italian seismicity revealed by the fractal approach. *Geophysical Research Letters*, **20**, 1911–1914.

Dimri, V. P., 1992. *Deconvolution and inverse theory*. Amsterdam/London/New Yor/Tokyo: Elsevier Science, p. 230.

Dimri, V. P., 1998. Fractal behavior and detectibility limits of geophysical surveys. *Geophysics*, **63**, 1943–1946.

Dimri, V. P., 2000. Fractal dimension analysis of soil for flow studies. In Dimri, V. P. (ed.), *Application of Fractals in Earth Sciences*. Rotterdam: A. A. Balkema, pp. 189–193.

Dimri, V. P., 2005. Fractals in geophysics and seismology: an introduction. In Dimri, V. P. (ed.), *Fractal Behaviour of the Earth System*. Berlin: Springer, pp. 1–18.

Dimri, V. P., Nimisha, V., and Chattopadhyay, S., 2005. Fractal analysis of aftershock sequence of Bhuj earthquake - a wavelet based approach. *Current Science*, **88**(10), 1617–1620.

Fedi, M., Quarta, T., and Santis, A. D., 1997. Inherent power-law behavior of magnetic field power spectra from a Spector and Grant ensemble. *Geophysics*, **62**, 1143–1150.

Grassberger, P., and Procaccia, I., 1983. Characteristics of strange attractors. *Physical Review Letters*, **50**, 346–349.

Hardy, H. H., and Beier, R. A., 1994. Fractals in Reservoir Engineering. Singapore: World Scientific, pp. XIV + 359.

Hewett, T. A., 1986. Fractal distribution of reservoir heterogeneity and their influence on fluid transport. SPE Paper 15386, doi: 10.2118/15386-MS

Hirabayashi, T., Ito, K., and Yoshi, 1992. Multifractal analysis of earthquakes. *Pure and Applied Geophysics*, **138**, 591–610.

Hirata, T., and Imoto, M., 1991. Multifractal analysis of spatial distribution of micro earthquakes in the Kanto region. *Geophysical Journal International*, **107**, 155–162.

Hirata, T., Sato, T., and Ito, K., 1987. Fractal structure of spatial distribution of microfracturing in rock. *Geophysical Journal of the Royal Astronomical Society*, **90**, 369–374.

Hosking, J. R. M., 1984. Modeling persistence in hydrological time series using fractional differencing. *Water Resources Research*, **20**(12), 1898–1908.

Kagan, Y. Y., and Knopoff, L., 1980. Spatial distribution of earthquakes, the two point correlation function. *Geophysical Journal of the Royal Astronomical Society*, **62**, 303–320.

Lovejoy, S., and Schertzer, D., 2007. Scaling and multifractal fields in the solid earth and topography. *Nonlinear Processes in Geophysics*, **14**, 465–502.

Lovejoy, S., Gaonac'h, H., and Schertzer, D., 2008. Anisotropic scaling models of rock density and the earth's surface gravity field. *Mathematical Geosciences*, **40**, 533–573, doi:10.1007/s11004-008-9171-7.

Lovejoy, S., Schertzer, D., Allaire, V., Bourgeois, T., King, S., Pinel, J., and Stolle, J., 2009. Atmospheric complexity or scale by scale simplicity? *Geophysical Research Letters*, **36**, L01801, doi:10.1029/2008GL035863.

Malamud, B. D., and Turcotte, D. L., 1999. Self affine time series I: generation and analysis. In Dmowska, R., and Saltzman, B. (eds.), *Advances in Geophysics*. San Diego: Academic. Long Range Persistence in Geophysical Time Series, Vol. 40, pp. 1–87.

Mandelbrot, B. B., 1967. How long is the coastline of Britain? Statistical self-similarity and fractional dimension. *Science*, **156**, 636–638.

Mandelbrot, B. B., 1983. *The Fractal Geometry of Nature*. New York: W. H. Freeman, p. 468.

Mandelbrot, B. B., and Van Ness, J. W., 1968. Fractional Brownian motions, fractional noises, and applications. *SIAM Review*, **10**, 422–437.

Maus, S., and Dimri, V. P., 1994. Fractal properties of potential fields caused by fractal sources. *Geophysical Research Letters*, **21**, 891–894.

Maus, S., and Dimri, V. P., 1995. Potential field power spectrum inversion for scaling geology. *Journal of Geophysical Research*, **100**, 12605–12616.

Maus, S., and Dimri, V. P., 1996. Depth estimation from the scaling power spectrum of potential fields? *Geophysical Journal International*, **124**, 113–120.
Moharir, P. S., 2000. Concepts similar to self-similarity in science. In Dimri, V. P. (ed.), *Application of Fractals in Earth Sciences*. Rotterdam: A. A. Balkema, pp. 33–43.
Molz, F. J., and Boman, G. K., 1993. A fractal-based stochastic interpolation scheme in subsurface hydrology. *Water Resources Research*, **29**(11), 3769–3774.
Ouchi, T., and Uekawa, T., 1986. Statistical analysis of the spatial distribution of earthquakes before and after large earthquakes. *Physics of the Earth and Planetary Interiors*, **44**, 211–225.
Painter, S., Beresford, G., and Paterson, L., 1995. On the distribution of seismic reflection coefficients and seismic amplitudes. *Geophysics*, **60**(4), 1187–1194.
Rieu, M., and Sposito, G., 1991. Fractal fragmentation, soil porosity and soil water properties: I. theory. *Soil Science Society of America Journal*, **55**, 1231–1238.
Saggaf, M., and Toksoz, M., 1999. An analysis of deconvolution: modeling reflectivity by fractionally integrated noise. *Geophysics*, **64**, 1093–1107.
Sivakumar, B., 2000. Fractal analysis of rainfall observed in two different climatic regions. *Hydrological Sciences Journal-des Sciences Hydrologiques*, **45**(5), 727–738.
Srivastava, H. N., 2005. Earth system modeling through chaos. In Dimri, V. P. (ed.), *Fractal Behaviour of the Earth System*. Berlin: Springer, pp. 179–203.
Srivastava, R. P., and Sen, M. K., 2009. Fractal based stochastic inversion of poststack seismic data using very fast simulated annealing. *Journal of Geophysics and Engineering*, **6**, 412–425.
Srivastava, R. P., and Sen, M. K., 2010. Stochastic inversion of prestack seismic data using fractal based prior. *Geophysics*, **75**(3), R47–R59, doi:10.1190/1.3379322.
Srivastava, R. P., and Dimri, V. P., 2007. Generation of three dimensional fractal subsurface structure by Voronoi tessellation and computation of gravity response of such fractal structure. US patent no. US 7,307,629 B2.
Tanaka, M., Kato, R., and Kayama, A., 2001. Growth and linkage of cracks and formation of creep fracture pattern simulated by a multi-crack growth model. *Fractals*, **9**(2), 223–230.
Tiwari, R. K., 2000. Detecting chaos from geophysical time series. In Dimri, V. P. (ed.), *Application of Fractals in Earth Sciences*. Rotterdam: A. A. Balkema, pp. 195–213.
Todoeschuck, J. P., and Jensen, O. G., 1988. Joseph geology and scaling deconvolution. *Geophysics*, **53**, 1410–1411.
Todoeschuck, J. P., and Jensen, O. G., 1989. Scaling geology and seismic deconvolution. *Pure and Applied Geophysics*, **131**, 273–288.
Toverud, T., Dimri, V. P., and Ursin, B., 2001. Comparison of deconvolution methods for scaling reflectivity. *Journal of Geophysics*, **22**, 117–123.
Tsonis, A. A., and Elsner, J. B., 1988. The weather attractor over very short time scale. *Nature*, **333**, 545–547.
Turcotte, D. L., 1997. Fractals and Chaos in Geology and Geophysics. Cambridge: Cambridge University Press, 398 pp.
Voss, R. F., 1985a. *Random fractals: Characterization and measurement, paper presented at NATO Advanced Study Institute on Scaling Properties of Disordered Media*. Norway: Geilo, pp. 1–11.
Voss, R. F., 1985b. Random fractal forgeries. In Earnshaw, R. A. (ed.), *Fundamental Algorithms for Computer Graphics*. Yorkshire: Springer, p. 8.

Cross-references

Earth, Density Distribution
Earthquake, Magnitude
Earthquakes, Intensity
Gravity, Data to Anomalies
Seismic Properties of Rocks
Statistical Seismology
Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat
Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity

FREE OSCILLATIONS OF THE EARTH

Sarva Jit Singh[1], Sunita Rani[2]
[1]Indian National Science Academy, New Delhi, India
[2]Department of Mathematics, Guru Jambheshwar University of Science and Technology, Hisar, India

Introduction

The seismological data can be analyzed and interpreted in terms of three basic concepts. The first is based on the geometrical ray theory of optics. Most of our knowledge about the interior of the Earth has come from the application of the ray-theoretic methods to seismological data. Ray theory is useful for periods shorter than a minute. Travel-times of seismic phases constitute an important component of the ray theory concept.

The second concept is that of surface waves. The study of the dispersion of surface waves helps us in determining the regional structure, e.g., low-velocity zones, oil-bearing structures, etc. The surface wave theory is useful in the period range 10–300 s.

The third concept is that of the normal-mode theory or free oscillations of the Earth. In here, the field is viewed as a sum of standing waves instead of traveling waves. The standing waves can be seen as peaks in the amplitude spectrum of a seismogram. Theoretical seismograms can be computed for spherically symmetric, nonrotating, elastic, isotropic (SNREI) Earth models by mode-summation techniques at low frequencies. At high frequencies, the number of modes required to represent the field becomes so large that the mode-summation technique is not practicable.

The Earth reacts to an earthquake or an explosion by vibrating as a whole in the same manner as does a bell when it is struck. The study of the Earth's free oscillations is based on the theory of vibrations of an elastic sphere. The problem increases in complexity as we proceed from an isotropic homogeneous sphere to models of the Earth which take into account its anelastcity, anisotropy, radial and lateral heterogeneity, rotation, gravity, and prestress.

The problem of the oscillations of an elastic sphere is an old one. It attracted many classical physicists of the nineteenth century, including S.D. Poisson, Lord Kelvin, G.H. Darwin, and H. Lamb. In a classic paper, Lamb (1882) discussed the various modes of oscillations of a uniform elastic sphere and calculated the important roots of the frequency equation. He classified the oscillations of a sphere

as of the "first class" and the "second class." In an oscillation of the first class, the dilatation and the radial component of the displacement vanish everywhere, so that there is no change of shape. In the case of the oscillations of the second class, the radial component of the vorticity vanishes. The oscillations of the first class are now known as toroidal or torsional oscillations and the oscillations of the second class as spheroidal or poloidal oscillations.

The Earth differs from a homogeneous elastic sphere in two important respects: self-gravitation and radial heterogeneity. Gravity does not affect the toroidal oscillations because these are divergence-free and have no radial component of the displacement, so that the density distribution remains unchanged. But gravity plays an important role in the case of the spheroidal oscillations. Bromwich (1898) investigated the effect of self-gravitation, simplifying his work by dealing only with an incompressible sphere. Love (1911) showed that a self-gravitating sphere of the size and mass of the Earth and of rigidity that of steel has a free period of oscillation of almost exactly 1 h for the mode ${}_0S_2$ (the fundamental mode of the spheroidal oscillation of the second order). Obviously, if this mode were to be observed, instruments would be needed that had sensitivity at periods very much longer than the periods of seismic motion then being routinely recorded.

Shida (1925) designed an electromagnetic seismograph with pendulum period 180 s and galvanometer period 1,200 s for observing Earth's free oscillations. However, the technology of that time was incapable of achieving the precision and sensitivity that were demanded. It was Hugo Benioff, who first constructed an electromagnetic strainmeter which was capable of recording free vibrations as long as 1 h. However, he had to wait for a big earthquake to make an appropriate use of the strainmeter developed by him, since the Earth will produce its deeper modes only when a large amount of kinetic energy is released by an earthquake or an explosion.

Benioff was able to record a ground motion of 57 min period for the Kamchatkan earthquake of November 4, 1952, on his strainmeter. He (Benioff, 1958) attributed it to the mode ${}_0S_2$ of the oscillations of the Earth. This was the first time that the natural oscillations of the whole Earth had been observed. It was but natural that several seismologists started investigating the problem of the free oscillations of the Earth theoretically and calculating the periods of the normal modes of real Earth models. Jobert (1956) applied Rayleigh's principle in calculating the period for the mode ${}_0T_2$ (the fundamental mode of the toroidal oscillation of the second order) as 43.54 min for the Bullen B model of the Earth. In the following year, Jobert (1957) applied the same method in obtaining the period for the mode ${}_0S_2$ as about 53 min. Pekeris and Jarosch (1958) applied variational methods in calculating the period of the mode ${}_0S_2$ as 53 min. Soon after, Alterman et al. (1959) integrated the equations of motion numerically and calculated the periods of various normal modes of the Earth. In particular, they showed that the period of the mode ${}_0S_2$ should be 53.7 min for Bullen B model of the Earth. A wide range of normal modes of the whole Earth were observed by several seismological groups in the USA for the great Chilean earthquake of May 22, 1960. The agreement between the observed periods and the periods calculated by Alterman et al. (1959) was excellent. This was the beginning of a new branch of seismology – the low frequency seismology. Impetus to research in this new field was provided by several factors: (1) availability of instruments capable of recording long-period oscillations of the Earth, (2) availability of high-speed electronic computers for Fourier analysis of the recorded data, and (3) availability of computer codes for numerical solution of differential equations efficiently. A network of low-noise accelerometers deployed all over the globe in 1970s permitted routine detection of normal-mode spectral peaks following earthquakes of relatively modest magnitude.

Observations of the 1960 Chilean earthquake indicated that the mode ${}_0S_2$ did not have a single spectral peak, but, instead, was composed of at least two lines with periods 53.1 and 54.7 min. Indeed, Backus and Gilbert (1961) and Pekeris et al. (1961) showed theoretically that for the mode ${}_0S_2$, five spectral lines exist if the rotation of the Earth is taken.

The work of Alterman et al. (1959) and their contemporaries did not include the effects of lateral heterogeneity, ellipticity, prestress, anisotropy, and anelasticity. Considerable progress has since been made in the direction of taking one or more of these features into account. Madariaga (1972) studied toroidal oscillations of a laterally heterogeneous Earth. Kanamori and Anderson (1977) and Woodhouse (1980) discussed the effect of the departure from perfect elasticity. Woodhouse and Dahlen (1978) investigated the effect of a general aspherical perturbation on the free oscillations of the Earth. Excellent reviews on the free oscillations of the Earth have been given by Stoneley (1961), Alterman et al. (1974), Lapwood and Usami (1981), Buland (1981), Dahlen and Tromp (1998), and others.

Toroidal oscillations of a uniform elastic sphere

On applying the method of separation of variables, the equation of motion for the toroidal oscillations of a homogeneous isotropic (uniform) elastic sphere of density ρ and Lamé's parameters λ and μ leads us to the following equation for the radial eigenfunction $W(r)$:

$$\frac{\mathrm{d}^2W}{\mathrm{d}r^2}+\frac{2}{r}\frac{\mathrm{d}W}{\mathrm{d}r}+\left[\frac{\rho\omega^2}{\mu}-\frac{l(l+1)}{r^2}\right]W=0, \tag{1}$$

where ω is the angular frequency and $l = 0, 1, 2, 3, \ldots$ is a separation variable. Equation 1 can be solved in terms of the spherical Bessel functions. Solution of Equation 1, which remains bounded as $r \to 0$, is of the form

$$W=A_0 j_l\left(\frac{\omega r}{\beta}\right), \tag{2}$$

where $j_l(\chi)$ is the spherical Bessel function of the first kind and $\beta = (\mu/\rho)^{1/2}$ is the shear wave velocity. The boundary condition that the surface $r = a$ of the sphere is traction-free yields the frequency equation:

$$\chi j_{l+1}(\chi) = (l-1) j_l(\chi), \tag{3}$$

where $\chi = \omega a/\beta$.

Since the frequency equation (3) is a transcendental equation, for each l, it possesses an infinite number of roots in χ. Let these roots be numbered in ascending order of magnitude and denoted by ${}_n\chi_l$ $(n = 0, 1, 2, \ldots)$. The corresponding frequencies and periods can be found through the relations.

$${}_n\omega_l = \left(\frac{\beta}{a}\right){}_n\chi_l, \quad {}_nT_l = \frac{2\pi}{{}_n\omega_l}. \tag{4}$$

Then ${}_n\omega_l$ are the eigenfrequencies and ${}_nT_l$ are the eigenperiods. Further, $n = 0$ is associated with the fundamental mode and $n = 1, 2$, etc., are associated with the first overtone, the second overtone, etc. For a given $l > 0$, there exists a family of n concentric nodal surfaces, i.e., interior surfaces at which the toroidal eigenfunction W vanishes. These nodal surfaces are given by

$$j_l\left(\frac{r}{a}\,{}_n\chi_l\right) = 0. \tag{5}$$

There are no toroidal oscillations associated with $l = 0$, since the corresponding displacement is identically zero. The value $l = 1$ yields a rigid rotation. The accompanying toroidal oscillations are known as rotatory vibrations. For $l = 1$, the frequency equation (3) becomes $j_2(\chi) = 0$, which may be expressed in the form

$$\frac{\tan\chi}{\chi} = \frac{3}{3-\chi^2}. \tag{6}$$

The lowest roots of Equation 6 are given, approximately, by

$$\chi = 5.763,\ 9.095,\ 12.322,\ 15.514\ldots \tag{7}$$

Equation 7 lists the first four overtones. For $l = 1$, the fundamental mode does not exist. For this mode $(l = 1,\ n = 0)$, every part of the sphere shows rotation around some axis in one direction. Such a motion could not be excited without the application of an external torque and so would not be a free oscillation. Consequently, the frequency equation has no root corresponding to this mode (Lapwood and Usami, 1981; p. 35).

For rotatory vibrations, the nodal spheres are given by

$$j_1\left(\frac{r}{a}\,{}_n\chi_1\right) = 0$$

which yields

$$\tan\left(\frac{r}{a}\,{}_n\chi_1\right) = \frac{r}{a}\,{}_n\chi_1. \tag{8}$$

Table 1 lists the dimensionless frequencies, periods, and nodal spheres for rotatory vibrations of a homogeneous sphere for the first four overtones.

From Equation 3, the frequency equation for the toroidal oscillations of the second order $(l = 2)$ is given by

$$\chi j_3(\chi) = j_2(\chi),$$

or, equivalently, by

$$\frac{\tan\chi}{\chi} = \frac{\chi^2 - 12}{5\chi^2 - 12}. \tag{9}$$

The corresponding dimensionless frequencies, periods, and nodal surfaces for the fundamental mode and the first four overtones are listed in Table 2.

Spheroidal oscillations of a uniform elastic sphere

Spheroidal modes corresponding to $l = 0$ are known as radial modes because the associated particle motion is purely radial. The frequency equation for the radial vibrations is (Ben-Menahem and Singh, 1981; p. 346)

$$\zeta\cot\zeta = 1 - \frac{1}{4}\left(\frac{\lambda + 2\mu}{\mu}\right)\zeta^2, \tag{10}$$

where

$$\zeta = \frac{\omega a}{\alpha}, \quad \alpha = \left[\frac{(\lambda + 2\mu)}{\rho}\right]^{1/2}. \tag{11}$$

Free Oscillations of the Earth, Table 1 Dimensionless frequency (χ), period (T in seconds), and nodal spheres (r_n) for the rotatory vibrations ($l = 1$) of a homogeneous sphere

n	χ	T[a]	r_1/a	r_2/a	r_3/a	r_4/a
1	5.763	1113	0.780			
2	9.095	705	0.494	0.849		
3	12.322	521	0.365	0.627	0.885	
4	15.514	414	0.290	0.498	0.703	0.907

[a]Assuming $a = 6371$ km, $\beta = 6.24$ km/s

Free Oscillations of the Earth, Table 2 Dimensionless frequency (χ), period (T in seconds) and nodal spheres (r_n) for the toroidal oscillations of the second order ($l = 2$) of a homogeneous sphere

n	χ	T[a]	r_1/a	r_2/a	r_3/a	r_4/a
0	2.501	2565				
1	7.136	899	0.808			
2	10.514	610	0.548	0.865		
3	13.771	466	0.419	0.660	0.895	
4	16.983	378	0.339	0.535	0.725	0.914

[a]Assuming $a = 6371$ km, $\beta = 6.24$ km/s

It may be noted that the frequency equation for the spheroidal oscillations is a function of the ratio λ/μ. For the Poissonian case ($\lambda = \mu$), the lowest roots of Equation 10 are given, approximately, by

$$\frac{\zeta}{\pi} = 0.816, 1.929, 2.936, 3.966, \ldots. \tag{12}$$

Assuming $a = 6371$ km, $\alpha = 9$km/s, the corresponding periods (in seconds) are found to be

$$T = 1735, 734, 482, 357, \ldots. \tag{13}$$

The fundamental radial mode ${}_0S_0$ corresponds to an expansion and contraction of the whole sphere without its form changing. For the nth overtone ${}_nS_0$, there are n concentric nodal surfaces inside the sphere at which the displacement vanishes identically.

For $l = 1$, the fundamental mode ${}_0S_1$ does not exist, since it represents a rigid-body translation. Even though the geometric center of the sphere is displaced, its center of mass does not move (Lapwood and Usami, 1981; p. 37). Thus the movement of the center does not contradict the physical requirement that the center of mass stays unmoved. The roots of the frequency equation representing the first four overtones ($l = 1$, $n = 1, 2, 3, 4$) are given, approximately, by

$$\chi = 3.424, 6.771, 7.444, 10.695, \tag{14}$$

assuming $\lambda = \mu$.

In the case of the spheroidal oscillations of the second order ($l = 2$), the oscillating sphere assumes alternatively the forms of a prolate and an oblate spheroid. Therefore, this mode of oscillation is known as the football mode.

Oscillations of an SNREI Earth model

In the last two sections, we have discussed oscillations of a homogeneous, isotropic, elastic, non-gravitating, spherical Earth model. This relieves the free oscillation problem of its mathematical intricacies while retaining its important features.

The equations of motion of a spherically symmetric, nonrotating, elastic, isotropic (SNREI) Earth model may be written in the form (Aki and Richards, 1980; Ben-Menahem and Singh, 1981)

$$\text{div}\ \vec{T} + \rho[\text{grad}(\Psi - gu_r) + \boldsymbol{e}_r g\ \text{div}\ \boldsymbol{u}] = \frac{\rho\,\partial^2 \boldsymbol{u}}{\partial t^2}, \tag{15}$$

$$\nabla^2\Psi = 4\pi G\,\text{div}(\rho\boldsymbol{u}), \tag{16}$$

where

$\boldsymbol{u}$ = displacement vector
$\vec{T} = \lambda\vec{I}\text{div}\boldsymbol{u} + \boldsymbol{\mu}(\nabla\boldsymbol{u} + \boldsymbol{u}\nabla)$ = stress tensor
g = acceleration due to gravity
G = gravitational constant
Ψ = perturbation in the gravitational potential

Equations 15 and 16 are to be solved subject to suitable boundary conditions. Thus, $\boldsymbol{u}$, $\boldsymbol{e}_r \cdot \vec{T}$, Ψ and $\partial\Psi/\partial r - 4\pi G\rho u_r$ must be continuous everywhere. However, at a boundary between a solid and a fluid or between two fluids, u_r rather than $\boldsymbol{u}$ must be continuous.

In the absence of external body force, the motion of the Earth consists of a superposition of free simple harmonic oscillations of the form

$$\boldsymbol{u}(\boldsymbol{r}, t) = \boldsymbol{u}(\boldsymbol{r})\exp(i\omega t).$$

Therefore, the equation which governs the free oscillations of the Earth is obtained from Equation 15 by making the substitution $\partial/\partial t \to i\omega$. This yields

$$\mathcal{L}\boldsymbol{u} + \rho\omega^2\boldsymbol{u} = 0, \tag{17}$$

where $\mathcal{L}$ is a self-adjoint linear operator. Equations 16 and 17 together with boundary conditions define an eigenvalue problem. Every model of the Earth will possess an infinite number of eigenfrequencies ω_i and a corresponding infinite number of eigenfunctions $\boldsymbol{u}_i(\boldsymbol{r})$.

As mentioned in the Introduction, the oscillations of the Earth are of two types: toroidal and spheroidal. At high frequencies, the toroidal oscillations correspond to the Love surface waves and *SH* body waves. Toroidal eigenfunctions are of the form

$$\boldsymbol{u}(\boldsymbol{r}) = W(r)[-\boldsymbol{e}_r \times \text{grad}\ Y_{ml}], \tag{18}$$

where

$$Y_{ml} = P_l^m(\cos\theta)\exp(im\varphi)$$

is a surface spherical harmonic. For every value of l except zero, there is a fundamental toroidal mode ${}_0T_l$ and an infinite number of overtones ${}_nT_l (n = 1, 2, \ldots)$. From Equation 18 it is seen that, for toroidal oscillations, $u_r = \text{div}\ \boldsymbol{u} = 0$.

At high frequencies, the spheroidal oscillations correspond to the Rayleigh surface waves and $P - SV$ body waves. The spheroidal eigenfunctions are of the form

$$\boldsymbol{u}(\boldsymbol{r}) = \boldsymbol{e}_r U(r) Y_{ml} + V(r)\text{grad}\ Y_{ml}. \tag{19}$$

For every value of l, there is a fundamental spheroidal mode ${}_0S_l$ and an infinite number of overtones ${}_nS_l (n = 1, 2, \ldots)$. From Equation 19, it can be verified that $(\text{curl}\ \boldsymbol{u})_r = 0$. On putting $\Psi(\boldsymbol{r}) = P(r)\,Y_{ml}$, and using Equations 18 and 19, Equations 16 and 17 can be transformed into a set of four ordinary differential equations of the second order in U, V, W, and P. One of these is an equation in W alone and corresponds to the toroidal oscillations. The remaining three equations are coupled in U, V, and P, and correspond to the spheroidal oscillations. These equations are awkward for numerical integration, because one needs to evaluate the derivatives of empirically determined quantities $\lambda(r), \mu(r)$, and $\rho(r)$ in order to obtain the coefficients. This can be formally obviated by converting these equations into an equivalent

set of eight linear differential equations of the first order which are free from the derivatives of λ, μ, and ρ. Out of these eight equations, two correspond to the toroidal oscillations and the remaining six to the spheroidal oscillations. We can thus express the equations governing the free oscillations of an SNREI Earth model in the form (Singh, 1992)

$$\frac{dY}{dr} = A(r)Y, \tag{20}$$

where Y is an $(N \times 1)$ column matrix and $A(r)$ is an $(N \times N)$ square matrix. The number N depends upon the kind of oscillations and the nature of the medium (Table 3).

From Equations 17 and 18, we have the following radial equation for the toroidal oscillations

$$\mu\left(\frac{d^2W}{dr^2} + \frac{2}{r}\frac{dW}{dr}\right) + \frac{d\mu}{dr}\left(\frac{dW}{dr} - \frac{W}{r}\right) + \left\{\rho\omega^2 - \frac{l(l+1)}{r^2}\mu\right\}W(r) = 0. \tag{21}$$

This equation involves $d\mu/dr$. Numerical evaluation of the derivative of the empirically determined $\mu(r)$ can be avoided by defining $y_1 = W, \quad y_2 = \mu\left(\frac{d}{dr} - \frac{1}{r}\right)y_1$.

The differential equation of the second order (21) is equivalent to two differential equations of the first order:

$$\begin{aligned}\frac{dy_1}{dr} &= \frac{y_1}{r} + \frac{y_2}{\mu},\\ \frac{dy_2}{dr} &= \left[\left(l^2 + l - 2\right)\frac{\mu}{r^2} - \rho\omega^2\right]y_1 - \frac{3}{r}y_2.\end{aligned} \tag{22}$$

For spheroidal oscillations, the set of six linear differential equations of the first order is obtained by defining

$$\begin{aligned}&y_1 = U, \quad y_2 = (\lambda + 2\mu)\frac{dU}{dr} + \frac{2\lambda}{r}U - \lambda\frac{l(l+1)}{r}V,\\ &y_3 = V, \quad y_4 = \mu\left[\frac{1}{r}(U - V) + \frac{dV}{dr}\right],\\ &y_5 = P, \quad y_6 = \frac{dP}{dr} - 4\pi G\rho U.\end{aligned} \tag{23}$$

In the case of the liquid core, $\mu = y_4 = 0$ and

$$y_3 = \frac{1}{r\omega^2}\left(gy_1 - \frac{y_2}{\rho} - y_5\right). \tag{24}$$

Explicit expressions for the elements of the matrix $A(r)$ have been given by Singh (1992).

Equation 20 subject to an appropriate set of homogeneous boundary conditions can be integrated by the Runge–Kutta method. For a given SNREI Earth model $\lambda(r), \mu(r)$ and $\rho(r)$ and given l, there is a discrete set of values of ω for which the boundary conditions are satisfied. We denote this set by ω_i (eigenfrequencies) and the corresponding displacements by $\boldsymbol{u}_i$. The index i signifies the normal mode. Each i stands for a triplet (l, m, n) of numbers; l being the colatitudinal mode number, also known as the angular order ($l \geq 0$); m, the azimuthal mode number ($-l \leq m \leq l$); and n, the radial mode number ($n \geq 0$). The numbers l, m, n describe the manner in which the displacement field depends upon the colatitude, the azimuth, and the radial distance, respectively. However, because of the spherical symmetry, the eigenfrequencies are degenerate in m (i.e., ω_i are independent of m). This degeneracy is removed when the rotation of the Earth, its ellipticity or lateral heterogeneity is taken into account.

The spherical harmonic

$$P_l^m(\cos\theta)\,(\cos m\phi,\ \sin m\phi)$$

from which the eigenfunctions are derived has (l-$|m|$) nodal latitude lines and $2|m|$ nodal longitude lines on the surface of the Earth. The value of m depends upon the source exciting the oscillations.

Table 4 compares the eigenfrequencies computed for PREM (preliminary reference Earth model (Dziewonski

Free Oscillations of the Earth, Table 3 Value of the number N indicating the order of the differential system

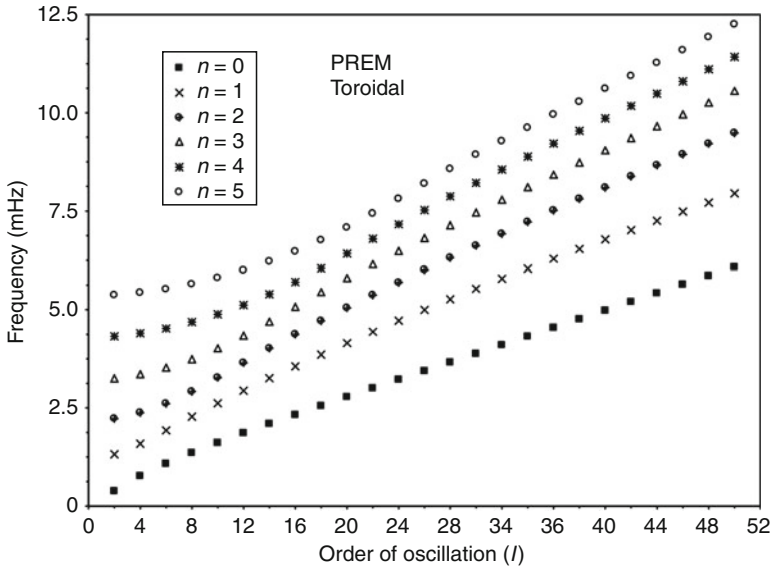
Oscillations	Medium	Gravity	N
Toroidal	Solid	No effect	2
Spheroidal	Solid	Yes	6
Spheroidal	Liquid	Yes	4
Spheroidal	Solid	No	4
Spheroidal	Liquid	No	2
Radial	Solid	Yes	2

Free Oscillations of the Earth, Table 4 Fundamental mode (n = 0) eigenfrequencies ($f = 1/T = \omega/2\pi$), in mHz (10^{-3} cycles per second), for PREM (preliminary reference Earth model). The corresponding mean observed frequencies are also shown

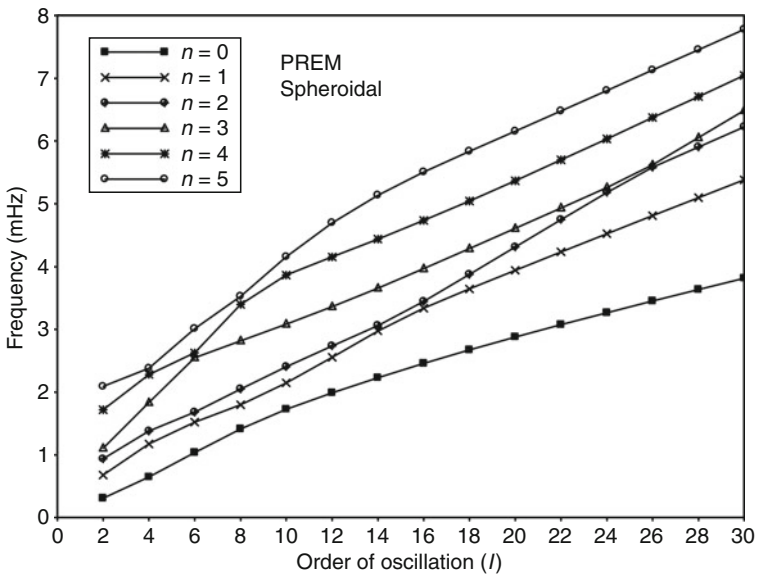
	Toroidal		Spheroidal	
l	Computed frequencies	Observed frequencies	Computed frequencies	Observed frequencies
2	0.37917	0.37970	0.30928	0.30945
3	0.58616	0.58630	0.46856	0.46850
4	0.76566	0.76593	0.64707	0.64682
5	0.92824	0.92845	0.84042	0.84002
6	1.07883	1.07913	1.03822	1.03755
7	1.22070	1.22115	1.23179	1.23105
8	1.35611	1.35660	1.41351	1.41285
9	1.48661	1.48707	1.57828	1.57755
10	1.61326	1.61367	1.72647	1.72560
15	2.21034	2.21116	2.34638	2.34495
20	2.77667	2.77759	2.87837	2.87795
30	3.88175	3.88290	3.81552	3.81480
40	4.97942	4.98070	4.71010	4.70840
50	6.07699	6.07900	5.61073	5.61030
100	11.56366	11.58500	10.24285	10.28900
150	17.01277	17.03500	14.99560	15.11000

and Anderson, 1981)) with mean observed frequencies. The computed eigenfrequencies are obtained from the website (http://stuplots.geophys.uni-stuttgart.de/~plots/Modes/modes.html). In this website, the eigenfrequencies and eigenfunctions can be computed for PREM and three other Earth models for $0 \leq n < 50$, $0 \leq l < 200$, using the FORTRAN program MINOS written by Drs. Freeman Gilbert and Guy Masters. The mean observed frequencies are taken from the website (http://igppweb.ucsd.edu/~gabi/rem.dir/surface/rem.surf.html). The agreement between the computed and observed values is excellent.

The eigenfrequencies of the toroidal oscillations computed for PREM are shown in Figure 1 as functions of the angular order l for the fundamental mode $n = 0$ and the first five overtones $n = 1$–5. The corresponding eigenfrequencies of the spheroidal oscillations are shown in Figure 2.



Free Oscillations of the Earth, Figure 1 Eigenfrequency $f = 1/T = \omega/2\pi$ of the toroidal oscillations for PREM.



Free Oscillations of the Earth, Figure 2 Eigenfrequency of the spheroidal oscillations for PREM. Only points corresponding to even values of $l = 2, 4, \ldots, 30$ have been plotted. For the sake of clarity, points representing the eigenfrequencies for even values of l have been joined by continuous curves. However, the intervening curves do not represent any physical reality.

Effect of the rotation of the Earth

As mentioned earlier, the eigenfrequencies of a spherically symmetric, nonrotating, elastic, isotropic Earth model are degenerate in the azimuthal mode number m. The introduction of the rotation removes this degeneracy. The effect of the diurnal rotation of the Earth can be calculated by carrying out a first-order perturbation calculation. Let $\mathbf{\Omega}$ denote the uniform angular velocity of the Earth about its center. Assume that the observer is referred to a non-inertial frame, which, for all times, maintains a state of uniform rotation with angular velocity $\mathbf{\Omega}$. Let (x_1, x_2, x_3) be a Cartesian coordinate system in this uniformly rotating frame of reference, let the origin of the system coincide with the center of the Earth, and let $\boldsymbol{e}_3$ be aligned along the axis of rotation, so that $\mathbf{\Omega} = \Omega \boldsymbol{e}_3$.

The rotation of the Earth introduces two body forces. One of them is the centrifugal force $r \sin\theta\, \Omega^2$ perpendicular to and away from the axis of rotation. The other force introduced by the rotation is the Coriolis force, which, per unit mass, is equal to $2\Omega(\partial \boldsymbol{u}/\partial t) \times \boldsymbol{e}_3$.

Assuming $\Omega/{}_n\omega_l << 1$ (the highest value of $\Omega/{}_n\omega_l \simeq 1/27$) and carrying out a first-order perturbation calculation, it can be shown that (Ben-Menahem and Singh, 1981; p. 390)

$$ {}_n\omega_l^m = {}_n\omega_l + m\tau\,\Omega, -l \le m \le l, \tag{25} $$

where

${}_n\omega_l$ = eigenfrequencies of a nonrotating Earth model
${}_n\omega_l^m$= eigenfrequencies of the corresponding rotating Earth model.

The splitting parameter

$$ \tau = \frac{1}{l(l+1)} \tag{26} $$

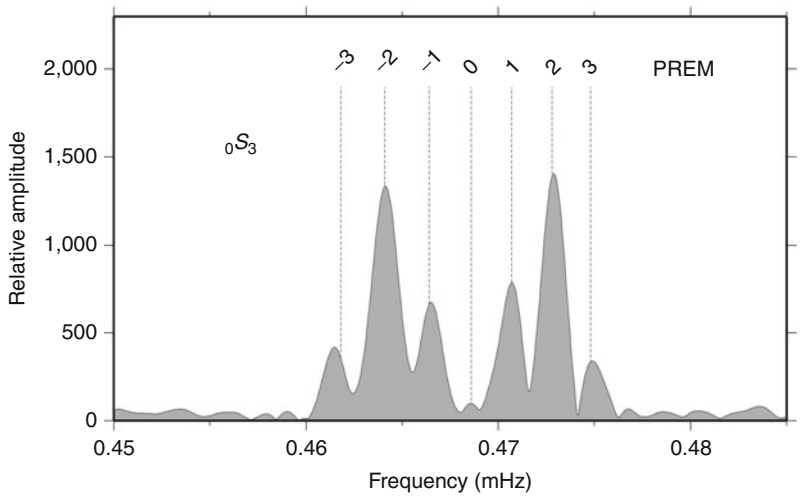
for the toroidal oscillations, and

$$ \tau = \frac{\int_0^a (2U + V) V \rho r^2 \mathrm{d}r}{\int_0^a [U^2 + l(l+1)V^2] \rho r^2 \mathrm{d}r} \tag{27} $$

for the spheroidal oscillations, where a is the radius of the Earth; U and V are defined in Equation 19.

Therefore, for both toroidal and spheroidal oscillations, the degenerate eigenfrequency ${}_n\omega_l$ is resolved by a slow rotation of the Earth into $(2l + 1)$ frequencies ${}_n\omega_l^m (-l \le m \le l)$. The set of $(2l + 1)$ spectral lines for a given (l, n) is called a mode multiplet and each member of this set is called a singlet. For the toroidal oscillations the splitting parameter $[l(l+1)]^{-1}$ does not depend upon the Earth model. However, in the case of the spheroidal oscillations, τ depends upon the Earth model. The splitting of the terrestrial spectral lines is the elastodynamic analogy of the splitting of atomic spectral lines by a magnetic field discovered by Zeeman in 1896.

Figure 3 shows the Zeeman splitting of the fundamental spheroidal mode of the third order ${}_0S_3$ observed in the Fourier spectrum of the gravimeter record of the February 27, 2010, Maule (Chile) earthquake. The time series is 400-h long and the effect of local barometric fluctuations has been removed. The rotation of the Earth removes the degeneracy of the multiplet; six out of seven singlets



Free Oscillations of the Earth, Figure 3 Fourier spectrum of the February 27, 2010, Maule (Chile) earthquake (Mw = 8.8) record provided by the superconducting gravimeter at the Black Forest Observatory (BFO). The rotation removes the degeneracy and splits the ${}_0S_3$ multiplet into seven singlets. The vertical dashed lines indicate the theoretically predicted frequencies of the singlets for the spherically symmetric Earth model PREM, including the effect of the rotation of the Earth and hydrostatic flattening. (Figure made available by Dr. Rudolf Widmer-Schnidrig of BFO who holds the copyright.)

can be clearly identified. The singlet corresponding to the azimuthal mode number $m = 0$ is missing because either the source or the receiver is located on a nodal line of the mode.

The ellipticity of the Earth also removes the degeneracy giving (Dahlen, 1968)

$$ {}_n\omega_l^m = {}_n\omega_t\left[1 + e\left(b + m^2 c\right)\right] \tag{28} $$

to first order in ellipticity e, where $b = b\,(n, l)$, $c = c\,(n, l)$. Thus the ellipticity splits the degenerate eigenfrequency ${}_n\omega_l$ into $(l + 1)$ lines ${}_n\omega_l^m (0 \le m \le l)$.

Other departures from symmetry, e.g., lateral inhomogeneity, anisotropy and initial stress, have the same effect of removing the degeneracy.

Jeans' formula

The complete seismic field induced by a point source in an SNREI Earth model can be expressed as an infinite sum of normal modes. However, we know from seismogram analysis that most of the recorded Earth motion can be explained in terms of propagating waves. Therefore, there must exist a relationship between these two seemingly different aspects of seismic wave motion.

Consider a general term in the normal-mode solution expressed by Equations 18 and 19. The factor of this term which depends upon the time and the colatitude is derivable from

$$ P_l^m(\cos\,\theta)\exp(i\omega t), \tag{29} $$

where $\omega = {}_n\omega_l$ is an eigenfrequency. Replacing the Legendre function by its asymptotic approximation for large values of l, this factor becomes

$$ (-l)^m\left(\frac{1}{2\pi l \sin\,\theta}\right)^{1/2}\left[\exp\left\{i\left[\omega t - (l + 1/2)\theta + \frac{\pi}{4} - \frac{m\pi}{2}\right]\right\}\right. \\ \left. + \exp\left\{i\left[\omega t + (l + 1/2)\theta - \frac{\pi}{4} + \frac{m\pi}{2}\right]\right\}\right]. \tag{30} $$

The first term describes a wave motion of frequency ω diverging from the pole $\theta = 0$, whereas the second term describes a wave motion diverging from the antipode $\theta = \pi$.

In cylindrical coordinates, a wave diverging from the axis may be expressed in the form

$$ \exp[i(\omega t - k\Delta)], \tag{31} $$

where k is the wave number. Putting $\Delta = a\theta$ and comparing Equations 30 and 31, we get Jeans' formula (Jeans, 1923):

$$ l + \frac{1}{2} = ka \quad (l >> 1). $$

This formula tells us that, if l is large, every mode of oscillation can be interpreted as a traveling wave whose parameters are functions of l and n and are given by

$$ T = \frac{2\pi}{\omega}, c = \frac{a\omega}{l + \frac{1}{2}}, \Lambda = \frac{2\pi a}{l + \frac{1}{2}}, \tag{32} $$

where T is the period of the oscillation, c is the phase velocity, and Λ is the associated wavelength. These relations are found to yield a good approximation for $l \ge 7$. The connection between the free oscillations and traveling waves (mode-ray duality) has been studied by Jeans (1923), Ben-Menahem (1964), Brune (1964), and others.

Jeans' formula can be used to determine the eigenfrequencies from the observed dispersion data, and then the normal-mode formalism can be used to find the elastic properties of the Earth.

Conclusions

We have described how to compute the eigenperiods of an SNREI Earth model. Modifications due to lateral inhomogeneity, anisotropy, and initial stress can be found by perturbation methods (Woodhouse and Dahlen, 1978). The eigenperiods of different modes of vibration are increased by the anelasticity of the Earth. In addition, anelasticity results in the decay of the amplitudes of the free oscillations with time. Necessary allowance for anelasticity can be made while constructing Earth models using normal-mode theory (Dziewonski and Anderson, 1981).

The study of the free oscillations of the Earth played a leading role in the development of geophysical inverse theory in the late 1960s (Backus and Gilbert, 1967, 1968, 1970) and 1970s. This theory enables us optimal extraction of information regarding the internal constitution of the Earth from a finite set of gross Earth data (Wiggins, 1972). Comprehensive exposition of the inverse theory has been given, amongst others, by Tarantola (1987) and Parker (1994). Dahlen and Tromp (1998) gave a comprehensive treatment of the problem of computing synthetic seismograms for a realistic three-dimensional model of the Earth, using normal-mode summation technique.

Major discoveries based on the observations of the Earth's free oscillations are (in chronological order):

1. Finite rigidity of the inner core (Dziewonski and Gilbert, 1971)
2. 1D density distribution inside the Earth (Gilbert and Dziewonski, 1975; Masters and Gubbins, 2003)
3. Construction of PREM (preliminary reference Earth model) by Dziewonski and Anderson (1981)
4. Construction of first global 3D upper mantle velocity model (Masters et al., 1982)
5. Construction of 1D attenuation model (e.g., Masters and Gilbert, 1983)
6. Tight upper bound on the maximum differential super rotation of the inner core (e.g., Laske and Masters, 1999)

Free oscillations of the Earth are typically observed after a large earthquake. Recently, continuous free oscillations of the Earth (primarily fundamental spheroidal modes between 2 and 7 mHz) have been discovered (Suda et al., 1998; Tanimoto, 2001). Observational evidence points to the

atmosphere (e.g., Kobayashi and Nishada, 1998) and oceans (e.g., Rhie and Romanowicz, 2004) as possible sources of excitation of these oscillations. Stochastic excitation of normal modes by surface pressure variations, caused by atmospheric and/or oceanic turbulence, may explain the overall amplitudes of observed modes of the continuous free oscillations of the Earth.

Acknowledgments

The authors are thankful to the reviewer, Dr. Rudolf Widmer- Schnidrig of the Black Forest Observatory, for his useful suggestions and comments and for providing the Fourier spectrum of the 2010 Maule (Chile) earthquake gravimeter record shown in Figure 3.

Bibliography

Aki, K., and Richards, P. G., 1980. *Quantitative Seismology*. San Francisco: W.H. Freeman.

Alterman, Z., Jarosch, H., and Pekeris, C. L., 1959. Oscillations of the Earth. *Proceedings of the Royal Society of London. Series A*, **252**, 80–95.

Alterman, Z., Eyal, Y., and Merzer, A. M., 1974. On free oscillations of the Earth. *Surveys in Geophysics*, **1**, 409–428.

Backus, G. E., and Gilbert, F., 1961. The rotational splitting of the free oscillations of the Earth. *Proceedings of the National Academy of Sciences of the United States of America*, **47**, 362–371.

Backus, G. E., and Gilbert, F., 1967. Numerical applications of a formalism for geophysical inverse problems. *Geophysical Journal of the Royal Astronomical Society*, **13**, 247–276.

Backus, G. E., and Gilbert, F., 1968. The resolving power of gross Earth data. *Geophysical Journal of the Royal Astronomical Society*, **16**, 169–205.

Backus, G. E., and Gilbert, F., 1970. Uniqueness in the inversion of gross Earth data. *Philosophical Transactions of the Royal Society of London. Series A*, **266**, 123–192.

Benioff, H., 1958. Long waves observed in the Kamchatka earthquake of November 4, 1952. *Journal of Geophysical Research*, **63**, 589–593.

Ben-Menahem, A., 1964. Mode-ray duality. *Bulletin of the Seismological Society of America*, **54**, 1315–1321.

Ben-Menahem, A., and Singh, S. J., 1981. *Seismic Waves and Sources*. New York: Springer-Verlag. Reprinted in 2000 by Dover Publications, New York.

Bromwich, T. J. I'. A., 1898. On the influence of gravity on elastic waves, and, in particular, on the vibrations of an elastic globe. *Proceedings of the London Mathematical Society*, **30**, 98–120.

Brune, J. N., 1964. Travel times, body waves and normal modes of the Earth. *Bulletin of the Seismological Society of America*, **54**, 2099–2128.

Buland, R., 1981. Free oscillations of the Earth. *Annual Review of Earth and Planetary Science*, **9**, 385–413.

Dahlen, F. A., 1968. The normal modes of a rotating, elliptical Earth. *Geophysical Journal of the Royal Astronomical Society*, **16**, 329–367.

Dahlen, F. A., and Tromp, J., 1998. *Theoretical Global Seismology*. Princeton: Princeton University Press.

Dziewonski, A. M., and Anderson, D. L., 1981. Preliminary reference Earth model. *Physics of the Earth and Planetary Interiors*, **25**, 297–356.

Dziewonski, A. M., and Gilbert, F., 1971. Solidity of the inner core of the Earth inferred from normal mode observations. *Nature*, **234**, 465–466.

Gilbert, F., and Dziewonski, A. M., 1975. An application of normal mode theory to the retrieval of structural parameters and source mechanisms from seismic spectra. *Philosophical Transactions of the Royal Society of London. Series A*, **278**, 187–269.

Jeans, J. H., 1923. The propagation of earthquake waves. *Proceedings of the Royal Society of London. Series A*, **102**, 554–574.

Jobert, N., 1956. Evaluation de la période d'oscillation d'une sphére elastique hétérogène, par application duprincipe de Rayleigh. *Comptes Rendus*, **243**, 1230–1232.

Jobert, N., 1957. Sur la période propre des oscillations sphéroidales de la terre. *Comptes Rendus*, **245**, 921–922.

Kanamori, H., and Anderson, D. L., 1977. Importance of physical dispersion in surface waves and free oscillation problems – review. *Reviews of Geophysics and Space Physics*, **15**, 105–112.

Kobayashi, N., and Nishada, K., 1998. Continuous excitation of planetary free oscillations by atmospheric disturbances. *Nature*, **395**, 357–360.

Lamb, H., 1882. On the vibration of an elastic sphere. *Proceedings of the London Mathematical Society*, **13**, 189–212.

Lapwood, E. R., and Usami, T., 1981. *Free Oscillations of the Earth*. Cambridge: Cambridge University Press.

Laske, G., and Masters, G., 1999. Limits on differential rotation of the inner core from an analysis of the Earth's free oscillations. *Nature*, **402**, 66–69.

Love, A. E. H., 1911. *Some Problems of Geodynamics*. Cambridge: Cambridge University Press. Reprinted in 1967 by Dover Publications, New York.

Madariaga, R., 1972. Toroidal free oscillations of the laterally heterogeneous Earth. *Geophysical Journal of the Royal Astronomical Society*, **27**, 81–100.

Masters, G., and Gilbert, F., 1983. Attenuation in the Earth at low frequencies. *Philosophical Transactions of the Royal Society of London. Series A*, **308**, 479–522.

Masters, G., and Gubbins, D., 2003. On the resolution of density within the Earth. *Physics of the Earth and Planetary Interiors*, **140**, 159–167.

Masters, G., Jordan, T. H., Silver, P. G., and Gilbert, F., 1982. Aspherical Earth structure from fundamental spheroidal mode data. *Nature*, **298**, 609–613.

Parker, R. L., 1994. *Geophysical Inverse Theory*. Princeton: Princeton University Press.

Pekeris, C. L., and Jarosch, H., 1958. The free oscillations of the Earth. In Benioff, H., Ewing, M., Howell, B. F., Jr., and Press, F. (eds.), *Contributions in Geophysics in Honor of Beno Gutenberg*. New York: Pergamon, pp. 171–192.

Pekeris, C. L., Alterman, Z., and Jarosch, H., 1961. Rotational multiplets in the spectrum of the Earth. *Physical Review*, **122**, 1692–1700.

Rhie, J., and Romanowicz, B., 2004. Excitation of Earth's continuous free oscillations by atmosphere-ocean-seafloor coupling. *Nature*, **431**, 552–556.

Shida, T., 1925. On the possibility of observing the free vibrations of the Earth. *Nagaoka Anniversary Volume*, 109–120.

Singh, S. J., 1992. Free oscillations of the Earth. *Current Science*, **62**, 227–232.

Stoneley, R., 1961. The oscillations of the Earth. In Ahrens, L. H., Press, F., Rankama, K., and Runcorn, S. K. (eds.), *Physics and Chemistry of the Earth*. New York: Pergamon, Vol. 4, pp. 239–250.

Suda, N., Nawa, K., and Fukao, Y., 1998. Earth' s background free oscillations. *Science*, **279**, 2089–2091.

Tanimoto, T., 2001. Continuous free oscillations: atmosphere-solid Earth coupling. *Annual Review of Earth and Planetary Science*, **29**, 563–584.

Tarantola, A., 1987. *Inverse Problem Theory*. Amsterdam: Elsevier.

Wiggins, R. A., 1972. The general linear inverse problem: implications of surface waves and free oscillations for Earth structure. *Reviews of Geophysics and Space Physics*, **10**, 251–285.
Woodhouse, J. H., 1980. The coupling and attenuation of nearly resonant multiplets in the Earth's free oscillation spectrum. *Geophysical Journal of the Royal Astronomical Society*, **61**, 261–283.
Woodhouse, J. H., and Dahlen, F. A., 1978. The effect of a general aspherical perturbation on the free oscillations of the Earth. *Geophysical Journal of the Royal Astronomical Society*, **53**, 335–354.

Cross-references

Body Waves
Inverse Theory, Monte Carlo Method
Propagation of Elastic Waves: Fundamentals
Seismic Wave Propagation in Real Media: Numerical Modeling Approaches
Seismic, Waveform Modeling and Tomography
Surface Waves

G

GEODESY, FIGURE OF THE EARTH

Kusumita Arora
National Geophysical Research Institute, Council of Scientific & Industrial Research (CSIR), Hyderabad, India

Definition

Geodesy is the discipline that deals with the measurement and representation of the Earth, its gravity field, and geodynamic phenomena (polar motion, Earth tides, and crustal motion) in three-dimensional time-varying space.

"Figure of the Earth" in geodesy most commonly deals with the representation of the external geometry of the Earth, without going into the causal mechanisms.

Background

The shape of the Earth has always been speculated upon – historical accounts of sailors record that they feared falling off the edge of the Earth if they sailed too far. Early ideas about the Figure of the Earth held the Earth to be flat and the heavens a physical dome spanning over it. Two early arguments for a spherical Earth were that lunar eclipses were seen as circular shadows that could only be caused by a spherical Earth, and that the pole star is seen lower in the sky as one travels south.

The flat disk advocated by Homer (in the ninth century B.C.) later gave way to the spherical body postulated by Pythagoras (in the sixth century B.C.) – an idea supported 100 years later by Aristotle (384–322 B.C.). According to old texts, the first geodetic experiment took place around 250 B.C. by Eratosthenes (about 284–192 B.C.) in Egypt. Eratosthenes noted that the Sun on the summer solstice day at Syene (close today to Assouan) was at the zenith while its zenithal distance at Alexandria (approximately on the same meridian) was recorded as $7°12'$. By knowing the geometrical distance between Syene and Alexandria thanks to a cadastre performed by Ptolemy Soter (about 360–283 B.C.), he deduced, the length of the terrestrial circumference as 40,500 km. Later Hipparchus (in the second century B.C.) recommended the use of latitudes and longitudes for cartography (in agreement with its Greek etymology).

The great Indian mathematician Aryabhata (476–550 A.D.) described the Earth as being spherical and rotating on its axis. He estimated the circumference of the Earth, accurate to 1% of today's measurement. The medieval Persian geodesist Abu Rayhan Biruni (973–1048 A.D.) is sometimes regarded as the "father of geodesy" for his significant contributions to the field. He solved complex geodesic equations based on spherical trigonometry in order to accurately compute the Earth's circumference and radius, which are very close to modern values.

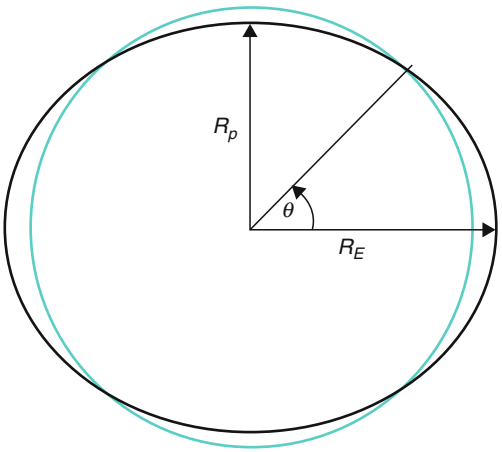
Mathematical Figures of the Earth

The actual shape of the Earth consisting of its topographic surface with its variety of landforms and water areas is not suitable for mathematical representations because the formulas that would be required to take these features into account would necessitate a prohibitive amount of computations. The Pythagorean concept of a spherical Earth offers a simple surface, which is mathematically easy to deal with. Many astronomical and navigational computations use it as a surface representing the Earth. While the sphere is a close approximation of the true Figure of the Earth and satisfactory for many purposes, to the geodesists interested in the measurement of long distances – spanning continents and oceans – a more exact figure is necessary. In the mid- to late twentieth century, research across the geosciences contributed to drastic improvements in the accuracy in the description of the Figure of the Earth.

The concept of a perfectly spherical Earth was eventually refined by Sir Isaac Newton based on his theory of

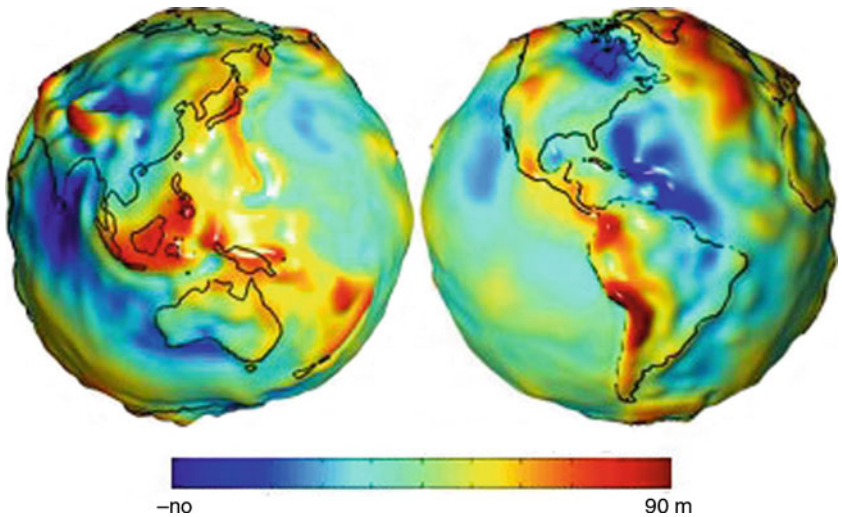
Harsh K. Gupta (ed.), *Encyclopedia of Solid Earth Geophysics*, DOI 10.1007/978-90-481-8702-7,

gravity, deriving that the Earth would not be perfectly spherical but slightly squashed at the poles and bulging at the Equator due to the forces acting upon it as it spins. Mathematically, this shape is described as an oblate spheroid or as an ellipsoid of revolution (Figure 1), an oval that revolves about its shortest dimension. This shape is most commonly used for exact measurements over long distances, across continents or oceans. Closer approximations utilize spherical harmonics for whole Earth shape or local approximations in terms of local reference ellipsoids. The idea of a planar or flat surface for Earth, however, is still acceptable for surveys of small areas where local topography is more important than the curvature. For such small areas, exact positions can be determined relative to each other without considering the size and shape of the whole Earth.



Geodesy, Figure of the Earth, Figure 1 An oblate spheroid (*black*) superimposed on sphere (*cyan*); R_P is the polar radius 6,357 km and R_E is the equatorial radius 6,378 km. θ is the latitudinal angle.

Another method estimates the Earth's surface using mean sea level of the ocean with all continents removed – this surface is equivalent to an equipotential surface and is called the Geoid. Due to variations in the Earth's mass distribution (oceans and land), the Geoid has an irregular shape that is described as "undulating." The ellipsoid of revolution mentioned above, is a mathematical approximation of the Geoid. The Geoid, unlike the ellipsoid, is so far deemed to be too complicated to serve as the computational surface on which to solve geometrical problems like point position. The separations between the two, referred to as Geoid undulations, Geoid heights, or Geoid separations, are irregular. The Geoid undulations over a spherical Earth are shown in Figure 2, where the red colored areas indicate positive deviations of the Geoid from the ellipsoid and the blues indicate negative deviations of the Geoid from the ellipsoid. Many of these deviations have obvious spatial correlations with surface expressions of substantial topographic features, which in turn indicate significant geodynamic activity. However, there are also undulations that do not bear such relations and the causal mechanisms remain to be established. An equipotential ellipsoid or level ellipsoid is an ellipsoid that is defined to be an equipotential surface. When an ellipsoid of revolution (semimajor axis a, semiminor axis b, flattening (a−b)/a) is defined, it can be made an equipotential surface of normal gravity potential, determined by means of the ellipsoidal surface, the enclosed mass, and the angular velocity (independently of the internal density distribution). The equipotential



Geodesy, Figure of the Earth, Figure 2 Geoid undulations derived from observed gravitational satellite orbit perturbations [from http://upload.wikimedia.org/~]; the greatest departure of the Geoid from ellipsoidal form is a 105-m low south of India. This is very small compared with undulations of the land surfaces or the 5-km mean difference in level between continents and ocean floors.

ellipsoid furnishes a simple, consistent, and uniform reference system for all purposes of geodesy and does not include an atmosphere.

The possibility that the Earth's equator is an ellipse rather than a circle and therefore the ellipsoid is triaxial has been a matter of scientific controversy for many years. Modern technological developments have furnished new and rapid methods for data collection and orbital data from satellites have been used to investigate the theory of ellipticity. A second theory proposes that observed long period orbital variations of the early Earth satellites indicate an additional depression at the South Pole accompanied by a bulge of the same degree at the North Pole. It is also contended that the northern middle latitudes were slightly flattened and the southern middle latitudes bulged by a similar amount. This concept suggested a slightly pear-shaped Earth (Figure 3). Modern geodesy tends to retain the ellipsoid of revolution and treat triaxiality and pear shape as a part of the Geoid figure: They are represented by the spherical harmonic coefficients C22, S22, and C30, respectively, corresponding to degree and order numbers 2,2 for the triaxiality and 3,0 for the pear shape.

Apart from rotation, there are other motional forces, which affect the shape of the equipotential surface. Precession is the effect of external torque exerted on the Geoid causing the spin axis to describe a circular cone. This motion is caused by the gravitational attraction of other celestial masses, particularly the sun, moon, and planets. Precession arises because of the ellipticity of the Geoid, with torque exerting on the equatorial bulge, in combination with the obliquity of the axis of rotation in relation to the plane of revolution (the ecliptic and fixed, heliocentric space). The period of precession is presently about 25,780 years.

Geodesy, Figure of the Earth, Figure 3 Concept of a pear-shaped Earth, from www.icsm.gov.au/mapping/~.

Polar Motion is the change in the position of the axis of rotation in relation to the surface of the Geoid, which was observed by five stations near latitude 39.1333° N, which began observations in 1899. The Conventional International Origin is defined as the mean position of the instantaneous pole during the period 1900–1905. Secular Polar Motion is the drift in the pole position thought to be due to tectonic movement. Seasonal Polar Motion is a fluctuation with an annual period with considerable fluctuation in amplitude ranging from 1.3 to 3.7 m.

Determining the exact Figure of the Earth is not only a geodetic operation or a task of geometry, but is also related to geophysics. Without any idea of the Earth's interior, we can state a "constant density" of 5.515 g/cm³ and, according to theoretical arguments such a body rotating like the Earth would have a flattening of 1:230. In fact the measured flattening is 1:298.25, which is more similar to a sphere and a strong argument in favor of very compact Earth's core. Therefore the density must be a function of the depth, reaching from about 2.7 g/cm³ at the surface to approximately 15 g/cm³ within the inner core. With increasing accuracy of the measured global gravity field from dedicated satellite missions, it is observed that the true vertical generally does not correspond to the theoretical vertical (deflection ranges from 2″ to 50″) because topography and all geological masses disturb the gravitational field. Therefore the gross structure of the Earth's crust and mantle can be determined by geodetic–geophysical models of the subsurface.

Reference coordinate system

Description of the shape and size of the Earth as well as measurement of distances on its surfaces and definitions of position locations of any particular point inevitably depends upon the use of a Reference Coordinate System. Reference ellipsoids are customarily expressed by the size of the semimajor axis (a) and flattening (f). The quantity f is a purely mathematical one. The mechanical ellipticity of the Earth (dynamical flattening) is determined by observation and differs from the geometrical because the Earth is not of uniform density. The 1967 Geodetic Reference System posited a 6,378,160 m semimajor axis and 1:298.247 flattening. The 1980 Geodetic Reference System posited a 6,378,137 m semimajor axis and 1:298.257 flattening. This system was adopted at the XVII General Assembly of the International Union of Geodesy and Geophysics (IUGG).

The World Geodetic System 84 (WGS 84) is the standard for use in cartography, geodesy, and navigation. It comprises a standard coordinate frame for the Earth, a standard spheroidal reference surface (the *datum* or *reference ellipsoid*) for raw altitude data, and a gravitational equipotential surface (the *Geoid*) that defines the *nominal sea level*. Presently WGS 84 uses the 1996 Earth Gravitational Model (EGM96) Geoid, revised in 2004. This Geoid defines the nominal sea-level surface by means of a spherical harmonics series of degree

360 (which provides about 100-km horizontal resolution). The deviations of the EGM96 Geoid from the WGS 84 reference ellipsoid range from about −105 m to about +85 m.

International Terrestrial Reference System (ITRS) was defined by the IUGG in 1991 as a geocentric nonrotating system identical to the International Astronomical Union (IAU) defined Geocentric Reference System with Geocentric Coordinate Time (TCG). Conventional Terrestrial System (CT) is geocentric and references north to the 1900–1905 mean pole and 0° longitude to Greenwich Observatory. Terrestrial coordinate systems are earth-fixed and spin and revolve with the Earth. Terrestrial Reference System (TRS) is a spatial reference system corotating with the Earth in space, with positions of points attached to the solid surface of the Earth having coordinates that undergo only small variations with time, due to geophysical, tectonic, or tidal deformations.

Nowadays the realizations of the fundamental reference system supported by the International Association of Geodesy, the ITRF, (International Terrestrial Reference Frame with the year of realization, at present the ITRF 2008) are very strongly recommended to be used in precise geodesy, for geosciences applications and for navigation satellite systems. Fortunately, the choice of fundamental parameters of WGS 84 was made in the beginning (in the 1980s) as close as possible to the ITRF (at this time defined by the BIH – the Bureau International de l'Heure). As a result, within 1 m there is no difference between what is defined as WGS 84 and the ITRF 2008, the latest international realization with a centimeter level of accuracy.

Datum

A datum is a framework that enables us to define coordinate systems. Definition of datum has improved considerably with advancements in measurement techniques through technology.

Traditional Geodetic Datum: To describe positions on the Earth accurately, traditionally two datum were required: a horizontal datum and a vertical datum.

Horizontal Datum: A horizontal datum is used to fix a position in the X and Y directions. A horizontal datum was traditionally defined using an ellipsoid and an origin. The ellipsoid was generally chosen for a best fit of the Geoid locally.

Vertical Datum: A Vertical Datum is used to fix a position in the vertical direction, up and down the Z-axis. A vertical datum is a line, value, or set of values from which heights are measured.

The International Terrestrial Reference Frame (ITRF) is defined by the International Terrestrial Reference System and the Datum consistent with it and is based on a number of very accurate observations all over the Earth, employing Very Long Baseline Interferometry (VLBI), Satellite Laser Ranging (SLR), Lunar Laser Ranging (LLR), and the Global Positioning System (GPS) techniques. The measurements relate to a 3D Earth-centered Cartesian coordinate system. In changing to this Geocentric Datum, the Global Positioning System [GPS] can be most effectively used.

Summary

Geodesy, Figure of the Earth outlines the basic parameters by which the shape of the Earth is defined. The Earth being a system in dynamic equilibrium, is under the influence of many forces, all of which contribute to its final structure and thereby its detailed shape. The subject of Geodesy has evolved sufficiently technologically and mathematically to take into account the details of the Earth's shape as well as the physics of the causal mechanisms.

Bibliography

Airy, G. B., 1845. Figure of the Earth. *Encyclopaedia Metropolitana*, **5**, 165–240.

Bomford, G., 1954. *Determination of the European geoid by means of vertical deflections*. Rpt of Comm. 14, IUGG 10th Gen. Ass., Rome.

Bomford, G., 1962. *Geodesy*. London: Oxford Publications.

Hooijberg, M., 1997. *Practical Geodesy*. Berlin: Springer. ISBN 3-540-61826-0.

Moritz, H., and Hofmann, B., 2005. *Physical Geodesy*. Wien: Springer.

Smith, J. R., 1986. *From Plane to Spheroid. Determining the Figure of the Earth from 3000 BC to the 18th century Lapland and Peruvian survey expeditions*. California: Landmark Enterprises. ISBN 0-910845-29-8.

Smith, J. R., 1996. *Introduction to Geodesy. The History and Concepts of Modern Geodesy*. New York: Wiley. ISBN 0-471-16660-X.

Todhunter, I., 1873. *The History of the Theories of Attraction and the Figure of the Earth*. London: Macmillan, Vol. 2.

Cross-references

Geodesy, Ground Positioning and Leveling
Geodesy, Networks and Reference Systems
Geodesy, Physical
Geoid Undulation, Interpretation
GPS, Tectonic Geodesy
Gravity Method, Satellite
Gravity, Global Models

GEODESY, GROUND POSITIONING AND LEVELING

Stelios P. Mertikas
Laboratory of Geodesy and Geomatics Engineering, Exploration and Positioning Division, Department of Mineral Resources Engineering, Technical University of Crete, Crete, Greece

Synonyms

Height determination; Terrestrial positioning

Definition

Geodesy. Science for the determination and representation of the figure, shape, size of the Earth and of other celestial bodies along with their external gravity field as a function of time.
Positioning. Determination of coordinates for points on earth, sea, or space with respect to a reference system.
Ground positioning. Determination of terrestrial coordinates.
Leveling. The determination of heights of points with respect to a reference surface.

Introduction

Geodesy, along with Astronomy, is considered the oldest earth science, and dates back to the Sumerians of 5,000 years ago. The word "geodesy" is composite and is derived from the Greek noun «Γη = earth» and the verb «δαίω = divide». It is, literally, the science used to "divide, distribute, and measure the Earth." The roots of geodesy can be found in ancient Egypt and Babylon where "the geometry of the Earth" was used for the recovery of the land property boundaries, after, for example, the flooding of crops by the Nile River. The word "Geodesy" appeared, for the first time, in the fourth century BC in Aristotle's book *Metaphysika*. Before Aristotle, "geodesy" was called the "geometry of the Earth." Ancient Greeks and Romans were the first to invent geodetic or *surveying* instruments for constructing roads, buildings, and tunnels, laying water pipes, culverts, land surveying, etc. Great scientists such as Pythagoras, Aristotle, Eratosthenes, Poseidonius, W. Snellius, J. Picard, Newton, Gauss, Cassini, and many others contributed to geodesy, surveying, high precision leveling and mapping.

Today, *geodesy* is the science of accurately measuring and understanding (1) the geometric size, shape, and orientation of the Earth, (2) its gravitational field, and (3) the variations of these quantities over time (Torge, 2001; Herring, 2010). The development of geodesy goes together with spectacular achievements in the fields of astronomy, mathematics, probability and statistics, physics, map projections, fast computing, etc. As a first approximation, "geodesy" can be considered as the science of measuring distances or heights between terrestrial objects, or measuring angles between lines and determining directions of lines for the establishment of coordinates of points on or near the earth's surface. This is also called *ground positioning* or *surveying*.

One of the main tasks of geodesy is *positioning*. This is the accurate determination of coordinates of points on the earth's surface, at sea, or in space with respect to a reference system (McCarthy and Petit, 2003). Positioning can be divided into *absolute positioning*, where coordinates are determined with respect to an inertial reference system, and into *relative positioning* when coordinates are derived with respect to other reference points on Earth. Positioning constitutes essential information for mapping, cadastral surveys, highway and dam construction, earthquake and landslide monitoring, geodetic control networks, crustal and structural deformation, etc.

Ground positioning has traditionally been carried out using measurements of distances, and angles (and or directions) between terrestrial lines with theodolites and measuring tapes (Anderson and Mikhail, 1997). Later, distances have been measured by other equipment using the electromagnetic radiation. Over the last two decades, the methods and instruments for measuring, acquiring, storing, calculating, reducing, and portraying geodetic ground positioning data have changed dramatically. New types of geodetic measurements, along with the traditional ones, are now being made with total electronic geodetic stations, terrestrial laser systems, inertial systems, ground-based radar interferometers, terrestrial wireless positioning systems, precise underwater positioning systems, and finally with the space-based systems such as the Global Navigation Satellite Systems (Kaplan and Hegarty, 2005).

Geodetic leveling is the precise determination of height difference between two or more points on the earth's surface. It ultimately aims at the determination of the height above a reference surface, usually the mean sea surface. Again in leveling, the ancient Egyptians and the Babylonians were the first to employ it to connect the Nile River with the Red Sea and to apply leveling for irrigation. In the following sections, the geodetic ground positioning and leveling will be briefly elaborated.

Ground positioning

The primary purpose of ground positioning is to provide coordinates of points on the earth's surface based either on (1) the astronomic determination of latitude, longitude, and azimuth of a terrestrial line or (2) the observation of horizontal angles, distances, zenith angles, and height differences.

A *horizontal angle* is defined as that angle measured in the horizontal plane of the observer between two vertical planes. The *horizontal plane* is the tangent plane to the equipotential surface of the gravity field at the point of observation. In early geodetic surveys, painstaking efforts were employed to accurately measure angles and distances. Angles were – and are still to this day – measured using theodolites and the distances were measured using tapes made from metal having a low coefficient of thermal expansion. Later, electronic instruments were used for measuring angles and distances.

Electromagnetic distance measuring devices have been used to determine the distance between geodetic stations on the ground. Electromagnetic pulses are transmitted to a retroreflector (or even to the terrain target) and the emitted signal finally returns to the instrument. Intervisibility requirements and attenuation of the signal, mainly due to the atmosphere, allow measurements to be made at ranges up to about 50 km; but the best results are obtained at shorter distances with an accuracy even higher than 0.1–0.3 mm over 100 m.

Point coordinates have thus been established by a process of triangulation, trilateration or traverse surveys. *Triangulation* is the process of determining the relative

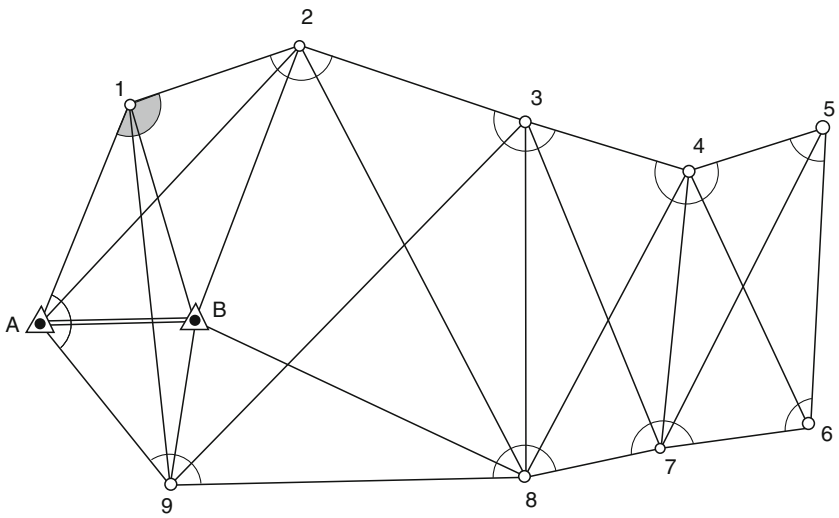
positions of a network of *control points* (Anderson and Mikhail, 1997) on the ground by theodolite measurements of directions or angles between intervisible stations (Figure 1). Angles have been more easily measured, as compared to distances, particularly when large distances over mountainous and rugged terrain were involved. As the name implies, triangulation employs horizontal angles and a limited number of lengths – called *geodetic baselines* – of connected triangles to determine trigonometrically the positions of stations (vertices of triangles). Triangles have been the simplest geometric figures that permit rigorous checks and adjustments of field observational errors. To compensate for the errors in the measurements, triangulation networks are adjusted mainly using the method of least squares. This method not only yields the final and adjusted coordinates of the vertices of the triangles but also gives their coordinate accuracies (Teunissen, 2000; Ghilani, 2010).

Trilateration is similar to triangulation process and again aims at the relative determination of positions using triangles, but now with their lengths measured by electromagnetic radiation with electronic instruments (Figure 2). Finally, *traverse surveys* consists of measurements of both horizontal (and vertical) angles at the traverse stations and distances along a series of connected lines on the ground. Unlike triangulation and trilateration, in which stations are commonly widely separated (e.g., 20–60 km), traverse surveys generally follow routes of highways, railways etc., with stations located close together (i.e., a few hundred of meters). The coordinates of these traverse stations are computed using basic trigonometric expressions as shown in Figure 3.

Astronomic azimuths are used to determine the orientation of the network with respect to the ellipsoid. The *astronomic azimuth* is the angle in the observer's local horizontal plane between the astronomic meridian and the vertical plane passing through the target point. Two coordinates, the geodetic latitude and longitude of each station of the network are then computed on this ellipsoid. The third coordinate of *height* above the reference ellipsoid is determined either from leveling and gravity data or from vertical angles.

All these geodetic measurements as well as other types (e.g., distance differences) are being made, today, by total geodetic stations, terrestrial laser systems, inertial systems, ground-based radar interferometry, wireless positioning, etc.

New kinds of theodolite measurement systems do not require a target and/or reflector to be monitored. They

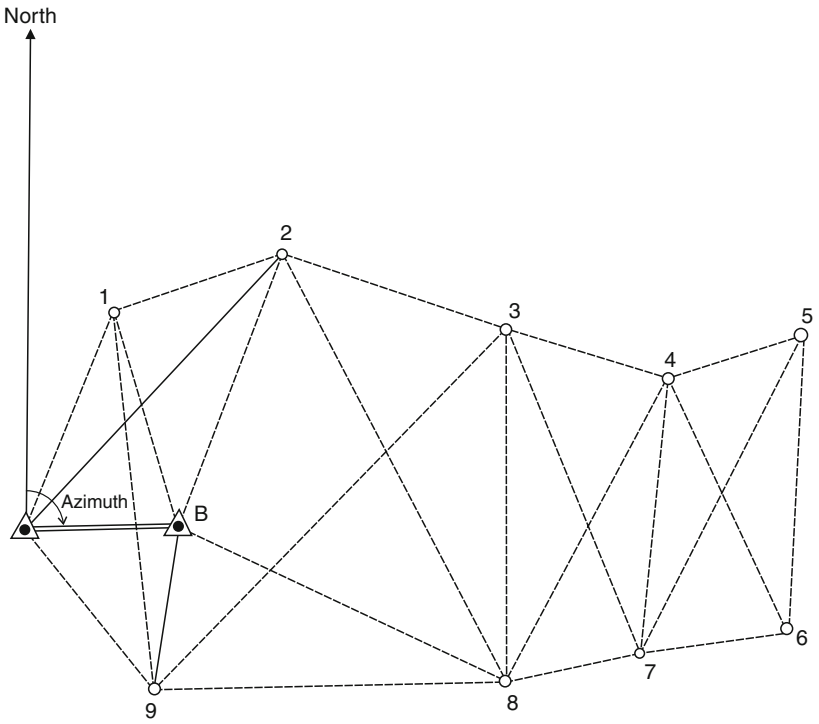


Known parameters:
1. Geodetic baseline AB (distance is known),
2. Latitude and Longitude of stations A and B,
3. Azimuth of length AB.

Measurements:
All horizontal angles to visible stations.

What is finally computed:
1. Distances to all control stations (triangle vertices),
2. Latitude and longitude of all stations,
3. Azimuth of all lines.

Geodesy, Ground Positioning and Leveling, Figure 1 An example of a triangulation network. The baseline AB is either known or measured, while all angles in the formed triangles are observed.



Known parameters:
1. Geodetic baseline AB,
2. Latitude and longitude of stations A and B,
3. Azimuth of line AB.

Measurements:
All distances to all inter-visible sites (stations).

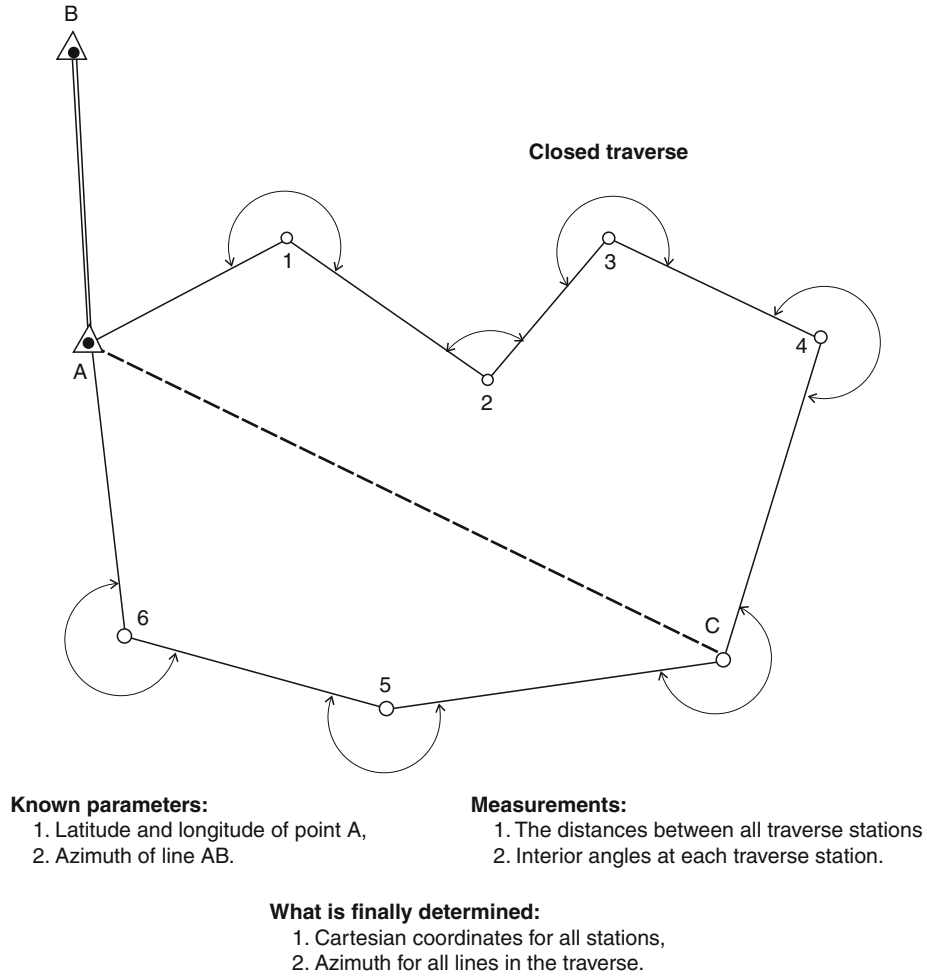
What is finally computed:
1. Angles and distances for all formed triangles,
2. Latitude and longitude for all stations,
3. Azimuth of all lines.

Geodesy, Ground Positioning and Leveling, Figure 2 A trilateration network where the azimuth of the baseline AB is determined to orient the network with respect to the north and all sides of the formed triangles are observed.

are able to measure point coordinates without targeting the object to be observed (Kahmen and Rieterer, 2004). Such systems use the texture on the terrestrial surface or of an object to locate "points of interest" to make angle and distance measurements. However, these new instruments require well-trained personnel to operate them.

Terrestrial laser scanners (Froehlich and Mettenleiter, 2004) can also provide detailed and highly accurate three-dimensional coordinates of terrestrial points rapidly and efficiently. Some scanners are suitable for indoor use and medium ranges (up to 100 m), while others are better for outdoor applications with longer ranges. Most systems use a laser beam, emitted by the instrument, which is directed into different directions, through a mechanical deflection mirror, to a number of reflective surfaces on the ground. The deflected laser light is detected and the instrument measures the time of flight for long ranges, and the phase measurement of the reflected signal, for medium ranges. For various scanning rates and/or a vertical scan angles up to 90°, the deflection mirror rotates continuously at adjustable speed and records target *range*, *angle*, signal *amplitude* and *time*. Some terrestrial laser scanners incorporate a camera to also capture a picture of an object being scanned. The accuracy of distance measurements is of the order of ± 1 cm, but this primarily depends on the intensity of the reflected laser light and on the reflectivity of the target surface. Terrestrial laser scanners have been used to characterize the kinematics of the landslide ground surface, by evaluating the involved displacement and deformation patterns. They have also been applied in surveying, mining, construction surveying, architecture, archaeology applications, as well as deformation monitoring, etc.

The standard wireless (i.e., Wi-Fi and the Ultra-Wide Band) networks employ a number of transmitters on the



Geodesy, Ground Positioning and Leveling, Figure 3 An example of a closed traverse. The geographic coordinates as well as the azimuth of the initial point A, with respect to point B, are known. Then the exterior (or interior) angles at each traverse station are observed along with the inter-station distances. The Cartesian coordinates for all stations of the traverse and the azimuths for all lines are determined using simple trigonometry.

ground and not only provide a way for finding the optimum communication relay routes in an ad hoc network of transmitters, but are also capable for ground positioning. The locations of the communication nodes (i.e., transmitters) of these networks are known. Positioning is achieved by measuring the *time of arrival* and the *time difference of arrival* for the communication signals. Thus, the positional accuracy of a network user depends on the ranging errors, the connectivity, and the available number of mobile nodes. The developed Ultra-Wide Band systems provide a dynamic three-dimensional accuracy of better than 5 mm (Kuhn et al., 2009). They are, at present, used for indoor wireless positioning, including robot tracking and monitoring but future uses to geophysical applications are soon to emerge.

Geodetic positioning techniques have also been extended to underwater applications using acoustic ranging to measure crustal deformation with centimeter level accuracies at several thousand of meters depths (Gagnon et al., 2005).

Another promising technique for geodetic positioning, height mapping, and deformation monitoring, such as tectonic seismic and volcanic deformation as well as ice and rock-glacier motion and ground subsidence, is provided by the *ground-based interferometric radar systems* (Werner et al., 2008). These instruments use real aperture radar antennas to illuminate a target area on the ground (Figure 4). The antennas are mounted in parallel on a rotational scanner. A radar image of the target area is formed, line by line, by rotating the antennas in azimuth and in the vertical axis. The range and the azimuth resolution of these radars are determined by the antenna beam width and the slant range. *Phase differences* (Cumming and Wong, 2005) between successive images, acquired from the same viewpoint, are used to determine line-of-sight displacements better than 1 mm. These instruments are

Geodesy, Ground Positioning and Leveling, Figure 4 A ground-based interferometric radar system to monitor deformation.

used for monitoring rapid moving landslides, and unstable mine slopes, as well as for terrain mapping. They can be used for short ranges from sensor to target (up to 1,000 m) to medium range applications (up to 4 km). The resulted accuracies are of the order of ± 1 mm and the entire observed area can reach up to thousands of square meters, while the radar system measures the displacements of all points in the area illuminated by the antenna.

Finally, space geodetic techniques are based on processing signals received from earth orbiting satellites to provide precise ground positioning and leveling (Herring, 2010). These systems, called Global Navigation Satellite Systems (GNSS), include the United States' Global Positioning System (GPS), the Russian Global Navigation System (Glonass), as well as the future ones such as the European Galileo, the Chinese Compass, the Indian Gagan, the Japanese Quasi-Zenith Satellite System, etc. Although originally developed for military and navigational purposes, new measuring and processing techniques (Hofmann-Wellenhof et al., 2007; Leick, 2003) were developed to improve the originally designed positioning accuracy of several meters. Today, these systems provide coordinates of ground locations at the millimeter level in the horizontal as well as in the vertical sense (Blewitt, 2010) over larger distances of several tens or hundreds of kilometers. In addition, they allow for three-dimensional positioning, do not require interstation visibility, their accuracy does not degrade with distance significantly, and provide continuous (e.g., every 1 or 20 Hz) positioning with respect to a global reference frame. In this vein, they have been applied for mapping, precise ground positioning to monitor plate tectonics, seismic and volcano ground deformation, etc.

The computational process, for all the above geodetic positioning techniques, requires their reduction on a reference surface, usually taken a flat horizontal surface or the *reference ellipsoid* (McCarthy and Petit, 2003) (Figure 5). In the flat horizontal surface the directions of the plumline (and thus gravity vector) are considered parallel. The ellipsoid is the curved surface approximating the size and shape of the earth and the Cartesian coordinate system is usually an earth-centered reference system for three-dimensional computations. Calculations involve solving equations derived from solid geometry, calculus, and statistics (Torge, 2001; Meyer, 2010).

The measured angles (directions) and distances are projected along normals onto the reference ellipsoid. The *geodetic datum* (McCarthy and Petit, 2003) is also defined as that which enables accurate location of features on earth through a mathematical model that describes the origin, orientation, and scale of coordinate systems.

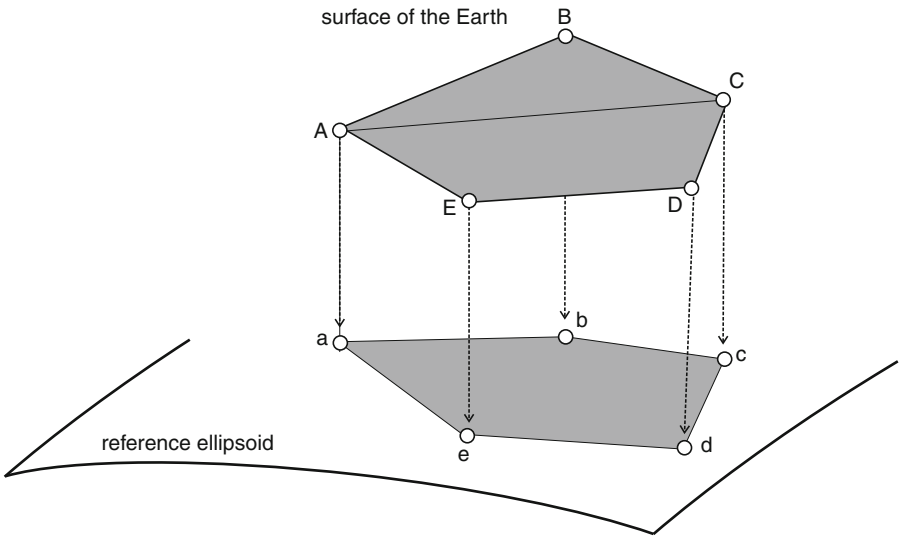
Leveling

Leveling is the general term applied to indicate the process by which elevations of points or differences in elevation between two or more points on the surface of the earth are determined. Differences in elevation are determined by taping, differential leveling, barometric leveling, and indirectly by trigonometric leveling. All height determinations require the establishment of a vertical datum (Ihde et al., 2008), usually taken as the mean sea level. This is realized by relative permanent objects on the ground, called *benchmarks*, whose elevation above an adopted geodetic datum is known.

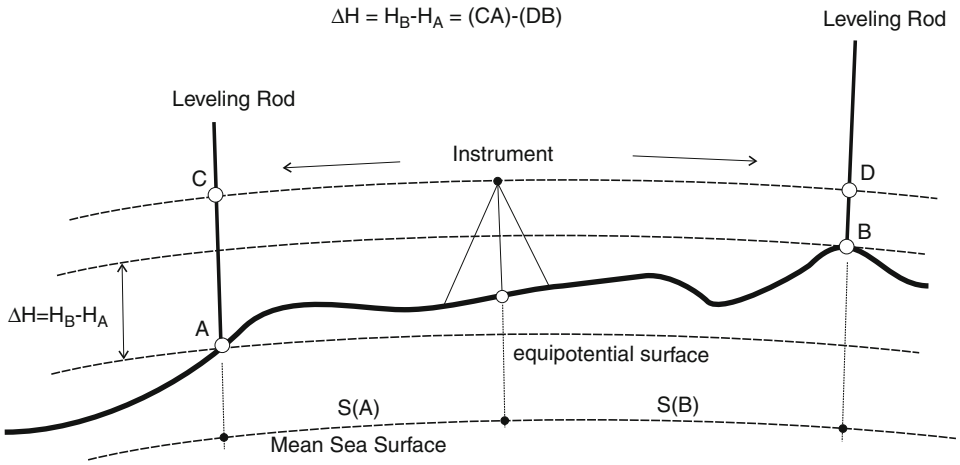
Differential leveling is accomplished by making measurements to two or more points through a telescope. The instrument's telescope, with suitable magnification, is regularly set horizontally with its line of sight being parallel to the equipotential surface of the local gravity field. A horizontal line of sight is established within the telescope

by a level vial or automatic compensator. The basic procedure is illustrated in Figure 6. A vertical measurement is made to the backsight of a graduated rod, held on fixed points, and is added to its height. Thus the height of the leveling instrument is determined. A second measurement is made to determine the height of the foresight rod. The instruments can then be moved forward and the leveling process is repeated until the elevation difference between the initial and the final point is established. Leveling procedures are limited by light of sight and to the distance between the level instrument and the rod. The readings on the rod can be taken either optically or electronically; the rod must be close enough to the level (about 40 m) to resolve the rod's readings in order to make accurate readings. Today, digital levels measure electronically to bar-coded rods, but can also measure optically to conventional rods. These are easy to use and because heights and distances are recorded electronically, reading and writing errors are avoided. They can be used for precise leveling of flat and sloping surfaces, and for establishing the vertical component of grade and ground profiles and for subsidence monitoring and control. The accuracy in height determination is of the order of less than ± 1 mm per km of distance in double leveling.

Differences in elevation are proportional to differences in the atmospheric pressure. Thus barometers, which are instruments to measure atmospheric pressure, can be used



Geodesy, Ground Positioning and Leveling, Figure 5 The observations of angles, distances, etc. measured on the earth's surface are reduced upon a reference ellipsoid and/or a flat horizontal surface.



Geodesy, Ground Positioning and Leveling, Figure 6 Differential leveling. Two leveling rods are placed vertically on points A and B, where the height difference is to be determined. The instrument is set up in the midway of A and B. A vertical measurement is made to the backsight of the graduated rod placed at A. A second measurement is also made to the foresight rod at B. The height difference is determined by subtracting the foresight from the backsight observation.

to determine relative elevations of points on the earth's surface. Elevations determined by barometric leveling probably are several meters in error even after they are corrected for the effects of temperature and humidity. Barometric measurements are made at a benchmark, while a second roving barometer is carried from point to point in the field.

The difference in elevation between two points, A and B, can also be established by measuring the vertical angle of the line from one point, A, to the other, B, as well as the slope distance, and then computing the difference in elevation using trigonometry. *Trigonometric leveling* is necessary when the elevations of inaccessible points must be determined. From two points of known position and elevation, the horizontal position of the unknown point is found by triangulation, and the vertical angles from the known points are measured. The technique is influenced by the atmosphere refraction (cloudy weather with good visibility with consistent temperature over measurement) and the earth curvature.

Summary

At present, the reference systems upon which all these geodetic positioning and leveling techniques are based, have been well defined with unprecedented accuracy and reliability (McCarthy and Petit, 2003; Ihde et al., 2008). The theory of errors for geodetic positioning and leveling has been well developed and applied in the measurements made. New data collection and processing technologies have emerged and matured. So, better results and reliable results are obtained, while the burden of adjustment computation and efficiency has further improved with fast computing and improved measurements and processing algorithms. All in all, Geodesy, ground positioning, and leveling have seen tremendous developments in recent years. New measurements (from optical to electronic) and new techniques (from angle measurement to phase and time and time-difference measurement of electromagnetic signals) and new instrumentation with increased accuracy (from meters to centimeters and millimeters at long distances) and capabilities (integration with Geographic Information Systems) have been emerged and matured for geophysical applications.

Bibliography

Anderson, M. J., and Mikhail, E. M., 1997. *Surveying: Theory and Practice*. Boston: McGraw-Hill.

Blewitt, G., 2010. *GPS and space-based geodetic methods*. In Herring, T. (ed.), *Geodesy: Treatise on Geophysics*. Amsterdam, The Netherlands: Elsevier. Chap. 11.

Cumming, G. I., and Wong, H. F., 2005. *Digital Processing of Synthetic Aperture Radar Data: Algorithms and Implementation*. New York: Artech House Publishers.

Froelich, C., and Mettenleiner, M., 2004. Terrestrial Laser Scanning – New Perspectives in 3D Surveying, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XXXVI-8/W2, 3–6 October, Freiburg, Germany.

Gagnon, K., Chadwell, C. D., and Norabuena, E., 2005. Measuring the onset of locking in the Peru-Chile trench with GPS and acoustic measurements. *Nature*, **434**, 205.

Ghilani, D. C., 2010. *Adjustment Computations: Spatial Data Analysis*. New York: Wiley.

Herring, T. (ed.), 2010. *Geodesy: Treatise on Geophysics*. Amsterdam, The Netherlands: Elsevier.

Hofmann-Wellenhof, B., Lichtenegger, H., and Wasle, E., 2007. *GNSS Global Navigation Satellite Systems: GPS, GLONASS, Galileo, and more*. Berlin: Springer.

Ihde, J., Mäkinen J., and Sacher, M., 2008. Conventions for the definition and realization of a European vertical reference system. In *Symposium of the IAG Sub-Commission for Europe (EUREF)*, June 18–21, Brussels.

Kahmen, H., and Rieterer A., 2004. High-precision object monitoring with image assisted theodolites: state of the art. In *International Workshop on Accelerator Alignment, CERN*, Oct 4–7, Geneva.

Kaplan, D. E., and Hegarty, Ch (eds.), 2005. *Understanding GPS: Principles and Applications*, 2nd edn. New York: Artech House.

Kuhn, M., Zhang, C., Mahfouz, M., and Fathy, A. E., 2009. Real-time UWB indoor positioning system with millimeter 3-D dynamic accuracy. In *Antennas and Propagation Society International Symposium, APSURSI '09*, June 1–5, IEEE, Charleston.

Leick, A., 2003. *GPS Satellite Surveying*, 2nd edn. New York: Wiley.

McCarthy, D. D., and Petit, G. (eds.), 2003. IERS Conventions (2003), International Earth Rotation and Reference Systems Service, IERS Technical Note No. 32, Verlag des Bundesamts fuer Kartographie und Geodaesie, Frankfurt am Main, 2004.

Meyer, H. T., 2010. *Introduction to Geometrical and Physical Geodesy: Foundations of Geomatics*. New York: ESRI Press.

Rueger, J. M., 2003. *Electronic Surveying Instruments: A Review of Principles, Problems and Procedures*, Monograph No. 18, School of Surveying and Spatial Information Systems, The University of New South Wales, Sydney, Australia.

Teunissen, P. J. G., 2000. *Adjustment Theory: An Introduction (Mathematical Geodesy and Positioning)*. Delft, The Netherlands: VSSD Publishers.

Torge, W., 2001. *Geodesy*. Berlin: Walter de Gruyter.

Werner, C., Strozzi, T., Wiesmann, A., and Wegmüller, U., July 6–11, 2008. *A Real-Aperture Radar for Ground-Based Differential Interferometry*. Boston: IGARSS.

Yu, K., Sharp, I., and Guo, Y. J., 2009. *Ground-Based Wireless Positioning*. West Sussex: Wiley.

Cross-references

Geodesy, Figure of the Earth
Geodesy, Networks and Reference Systems
Geodetic Pendulums, Horizontal Ultra Broad Band

GEODESY, NETWORKS AND REFERENCE SYSTEMS

Hayo Hase
Geodätisches Observatorium Wettzell, Bundesamt für Kartographie und Geodäsie, Bad Kötzting, Germany

Definition

Geodesy. Scientific discipline dealing with the measuring and representing of Earth in the domains of time, space, and the gravitational field.

Networks. Set of materialized geodetic reference points as carriers of a specific reference frame tied to each other by geodetic measurements.

Reference system. A reference system is based on one definition of the reference. A geodetic reference system is a definition of a specific coordinate system.
Reference frame. A reference frame is one (among many possible) realizations of a reference system. It is characterized by a set of reference points from geodetic networks and a reference epoch.

Introduction

The historical definition of geodesy was given by F. R. Helmert (1880): "Geodesy is the science of measuring and mapping of Earth's surface." According to this definition, the task of geodesy is expressed in Torge (2003): "The task of geodesy is to determine the figure and the external gravity field of the Earth and other objects in space as functions of time by surface- or space-based observations."

Therefore, geodesy deals with measurements and representation of Earth's geometry, its gravitational field, and geodynamic phenomena (Earth rotation, Earth tides, and crustal motion) in three-dimensional and time-varying space. Metrology defines the physical units, which are needed in geodesy. The measurements are related to various geodetic instruments and different methods. Earth-bound instruments are mostly operated in global networks, while a space-borne instrument senses the entire Earth while in orbit. Earth models contain a mathematical description approximating the physical reality in which observation data are modeled. The description of Earth's figure and its external gravity field requires the definition of reference systems, which are constituted by measurements (Schneider, 1992).

As it is very difficult to fix a static and very precise global coordinate system to the continuously changing Earth, a permanent monitoring of reference positions in geodetic networks is necessary. These permanent observing networks, in which fundamental stations for geodesy have an indispensable role, are the base for global geodetic reference frames. Such reference frames are the backbone for a better understanding of the system Earth, for any kind of geoinformation systems, and for space exploration (Hase, 1999).

Geodetic reference systems

Theory of reference systems

In geodesy, the terms *reference system* and *reference frame* are different: A reference system is based on physical theories and reflects an ideal idea. A reference frame is one among many possible realizations of a reference system and based on geodetic observations and hence available to users (Kovalevsky and Mueller, 1981).

Example: A global reference system can be defined in such a way that the origin of the coordinate system coincides with the center of mass. The center of mass is an abstract physical model and not accessible as a reference point. However, geodetic observations of satellite orbits are useful to realize a corresponding reference frame because the center of mass of Earth coincides with one of the two foci of the elliptical orbit of the satellite. The used data set of observations delivers one particular reference frame with respect to the definition of the reference system.

Reference systems are standardized by the introduction of conventional reference systems that are based on the decisions of international committees. These conventional reference systems become available to users by conventional reference frames (Schneider, 1999). Reference frames are always based on observation data, which are processed with introduced hypotheses of the used reduction models. Due to the arbitrary settings in the processing of observation data, the defined ideal of a reference system can only be approximated with different analysis strategies. This circumstance and the need for comparison call for a standardization of reduction models in order to compare different reference frames. The IERS-Conventions are an example for the standardization of physical constants and models with respect to global reference frames (IERS Conventions, 2003).

Two reference systems are of importance in global geodesy: (a) the barycentric space-bound celestial reference system (CRS), and (b) the geocentric Earth-bound terrestrial reference system (TRS). While the first system is related to distant objects in space, the latter one is related to terrestrial reference points or to reference points in the near Earth space. Both systems are tied by the Earth Rotation Parameters (EOP).

The figure of Earth is approximated by Earth models. The geometric models are spherical or ellipsoidal while the geophysical models are based on expansion of the spherical harmonic functions in order to model the geoid as an equipotential reference surface. In an ideal situation, models of both categories coincide with their origins and have parallel main axes.

Metrology

Geodetic measurements in the domains of space, time, and gravity field refer to physical units of the SI (Sistéme International). In general, the modeling of space-geodetic measurements in Newtonian space-time shows deficiencies. The underlying physical model for the definition of physical units must be based on the theory of general relativity (Damour et al., 1991). This requires that the physical units are defined in infinitesimal small spatial areas with only local and not global validity. The question is whether physical units defined in local laboratory space can be used on the global scale for space geodesy.

The units for timescale, spatial-scale, and gravitational potential cannot be defined independently from the object to be measured. The metrological units are only defined as precisely as the measurements of the geodynamic phenomena interacting with the laboratory and vice versa (Hase, 1999).

The introduction of the constant of the speed of light, c = 299792458 m/s, enables the introduction of the physical

units *meter* and *second* and a method to measure space-time intervals if the synchronization of two local clocks fulfills the Einstein synchronization condition. Generally, this condition is not fulfilled on Earth. Therefore, the spatial distance of the clocks must be defined in relation to the desired resolution of the measurement in order to neglect the spatial curvature effects.

The global timescale is ideally related to the center of the mass of Earth as the effect of external gravitational bodies disappears. For practical reasons, timekeeping is not possible at the center of Earth's mass. Therefore, the timescale definition of *Temps Atomique International* (TAI) refers to a geocentric reference system with a definition of the second related to the equipotential surface of the rotating geoid. The world time is the *Coordinated Universal Time* (UTC), which is TAI corrected by an integer number of leap seconds realized by local clocks. The time second is realized by cesium frequency standards. Its SI-definition is as follows: The second is the duration of 9192631770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium atom ^{133}Cs (BIPM, 1998).

The metric scale is a secondary unit derived from the definition of the time second and the constant of the speed of light. The meter is defined as the distance travelled by light in vacuum in 1/299792458 s. Depending on the definitions for speed of light and time second, the metric scale is only valid in the vicinity of the time second-defining clock. That is why large distances should be seen as space-time intervals that are characterized by two-way measurements or one-way measurements with clock transportation.

The mass of Earth could be seen as constant if there were no exchange of radiation with the ambient universe. With respect to the terrestrial reference frames, the solar system with its gravitating bodies is seen as sufficiently isolated. The universal gravitational constant, $G = 6.67259 \times 10^{-11}$ m³/kg/s² $\pm$ 8.54×10^{-15} m³/kg/s², is imprecisely known with its relative accuracy of 10^{-4} (Bursa, 1992). G is the only link between the SI unit for mass and the gravitational phenomena. Given that only in the magnitude 10^{-17}, the tidal potential at the height of the geoid has an effect on the definition of the time second, the value of G may be applied in geocentric as well as in barycentric reference systems.

The product of the universal gravitational constant with the mass of Earth is called the geocentric gravitational constant, $GM = 398600.4418$ m^3/s^2 $\pm$ 0.0008 m^3/s^2, and can be determined with higher accuracy (2×10^{-9}) than the individual factors (10^{-4}). It is used for geodetic Earth models.

The IERS Conventions and IERS Standards mirror the actual status of parameters and models to be used in order to obtain a consistent TRF approximating the underlying physical theories with the best available models and constants (IERS Conventions, 2003).

Global geodetic reference systems

With the beginning of the satellite age, the need for an accurate global geodetic reference system became evident. Also spatial and geographical information datasets extending over national boundaries are calling for a common global reference frame for collection, storage, visualization, and exchanging of geoinformation.

Global geodetic reference systems have always been based on definitions and measurements with respect to the available scientific knowledge and technology. In former times, solid Earth models had been sufficient for global reference frames, scientific geodetic studies, and for mapping applications. Modern geodetic space techniques such as Very Long Baseline Interferometry (VLBI), Satellite Laser Ranging (SLR), Global Navigation Satellite Systems (GNSS), and modern instrumentation of local sensors such as absolute gravity meter (AG) and superconducting gravity meter (SG), much higher resolutions in the measurements showed the variability of the observables as a function of time. Consequently, the solid Earth models had to be replaced by viscoelastic Earth models.

For the purpose of reference frames, the viscoelastic Earth models must also contain models of plate-tectonic motion because reference points at different continental plates expose different directions and velocities of proper motion over time. The exhaustion of the kinematic motion of individual reference points is achieved in the coordinate estimation under the condition that the squared sum of the individual point motion is made to zero. This condition is commonly expressed as *no-net-rotation* (NNR).

As it appears to be an impossible task to fix a static and very precise global coordinate system on the continuously changing Earth, a permanent monitoring of reference positions in geodetic networks became necessary. These permanent observing networks, in which fundamental stations for geodesy have an indispensable role, are the base for global geodetic reference frames.

Currently, two global geodetic reference systems have a major importance: ITRS and WGS84.

International terrestrial reference system (ITRS)

The most important conventional terrestrial reference system (CTRS) is the International Terrestrial Reference System (ITRS). The ITRS is defined as a cartesic coordinate system in which (a) the origin coincides with the center of Earth's mass (including the masses of the hydrosphere and the atmosphere), (b) the metric scale is defined by the fixed velocity for the propagation of light in a gravitating environment and the physical definition for the unit of time according to SI, (c) the orientation is obtained by observation of Earth rotation relative to space-fixed objects of the celestial reference frame (ICRF), and (d) the system has no global residual rotation with respect to horizontal motions at the Earth's surface (NNR) (IUGG, 1991). The definition of origin, scale, and orientation is the *geodetic datum* of a reference

system. If a geographic coordinate system (latitude, longitude, ellipsoidal height) is chosen, the corresponding parameters of the desired reference ellipsoid must also be defined and becomes in general part of that datum. The GRS80-ellipsoid (Geodetic Reference System, 1980) is often used for the ITRS as best fitting the global figure of Earth's surface.

The ITRF is currently based on observation data of more than 3,800 reference points. International Services for VLBI, SLR, GNSS, DORIS, and the collection of local surveys at collocation sites with multiple techniques being operated provide the ITRF database.

The ITRF is published with respect to certain epochs, i.e., ITRF2005 and ITRF2008. The ITRF-coordinates and -velocities of the reference points are estimated in the processing of observational data. Each reference frame refers to a specific epoch and shall allow linear extrapolation of the coordinates to any desired epoch. However, long-term series and certain geographical regions partly expose nonlinear evolution of point positions and require frequently updated regional reference frames or nonlinear motion models.

The updating of the ITRF is necessary not only due to the continuously acting geodynamic forces on the global network and the need to update reference points, but also the introduction of improved models in the data reduction. For the generation of each new frame, the NNR condition is applied. The IERS Conventions contain the adopted constants, parameters, and models that shall be used for data processing (IERS Conventions, 2003).

World geodetic system 1984 (WGS84)

WGS84's geometrical base is comprised of (a) a global coordinate frame for the Earth and (b) a standard spheroidal reference surface (the *datum* or *reference ellipsoid*) for raw altitude data. In addition, the geophysical base is a gravitational equipotential surface model (the *geoid*) that defines the nominal sea level. The geophysical base was periodically updated as the underlying Earth Gravitational Model (EGM96, EGM2008) improved by its ground resolution from 200 km to 10 km.

The geometrical definition of WGS84 identifies Earth's center of mass as the origin of the coordinate system. The *zero-longitude* is identified with the IERS Reference Meridian, which lies 5.31 arcsec east of the Greenwich Prime Meridian for historical reasons related to the NNSS Doppler System.

The WGS84 is defined and maintained by the US Department of Defense. The GPS control segment consists of only a few globally distributed reference stations, whose coordinates together with the broadcast ephemerides are expressed in this datum. The control segment network is also called *zero-order network* as it fixes the reference frame.

Consequently, any GPS user who wants to determine a position in WGS84 is dependent on how accurately known the satellite orbits with respect to the control segment reference stations are. Due to orbit perturbations, the broadcast ephemerides are only approximations and need frequent updates. In absolute positioning mode, results are said to be in the WGS84 datum. In relative positioning model, a baseline vector between two stations can be determined with much higher accuracy. However, the datum is almost completely defined by the base station, which has a much higher uncertainty than the baseline vector. It is therefore desirable to tie GPS baseline vectors to ITRF reference stations using the precise orbit product of the International GNSS Service (which is obtained from hundreds of network stations) and express sets of estimated coordinates related to the most recent ITRF.

The *absolute accuracy* is defined by the true relationship of a position to the coordinate datum. The ITRF is multi-technique based and with a higher number of network stations, for which reason the ITRF provides a higher absolute accuracy than the WGS84 in its current configuration.

As it can be recognized by the defining parameters (Tables 1 and 2), the ITRS and WGS84 coincide within a few centimeters.

Geodesy, Networks and Reference Systems, Table 1 Defining geometrical and geophysical parameters of the Geodetic Reference System, 1980 (GRS80). The dynamical flattening expressed by the dynamical form factor J_2 and determined by satellite orbit perturbations was given preference in the definition. The geometrical flattening f = 1/298.257222101 became a derived parameter (Moritz, 1980)

Defining parameters of GRS80	Symbol	Value
Semimajor axis or equatorial radius	a	6,378,137.0 m
Dynamical form factor	J_2	0.001,082,63
Geocentric gravitational constant (incl. atmospheric mass)	GM	$398{,}600.5 \times 10^{+9}$ m^3/s^2
Angular velocity of Earth rotation	ω	$0.729{,}211{,}5 \times 10^{-4}$ rad/s

Geodesy, Networks and Reference Systems, Table 2 Defining geometrical and geophysical parameters of the World Geodetic System 1984. As opposed to the GRS80, the geometrical flattening f is the defining quantity and the dynamical flattening given by the dynamical form factor J_2 = 0,001081874 is a derived quantity

Defining parameters of WGS84	Symbol	Value
Semimajor axis or equatorial radius	a	6,378,137.0 m
Flattening $(a - c)/a$	f	1/298.257,223,560
Geocentric gravitational constant (incl. atmospheric mass)	GM	$398{,}600.441{,}8 \times 10^{+9}$ m^3/s^2
Angular velocity of Earth rotation	ω	$0.729{,}211{,}5 \times 10^{-4}$ rad/s

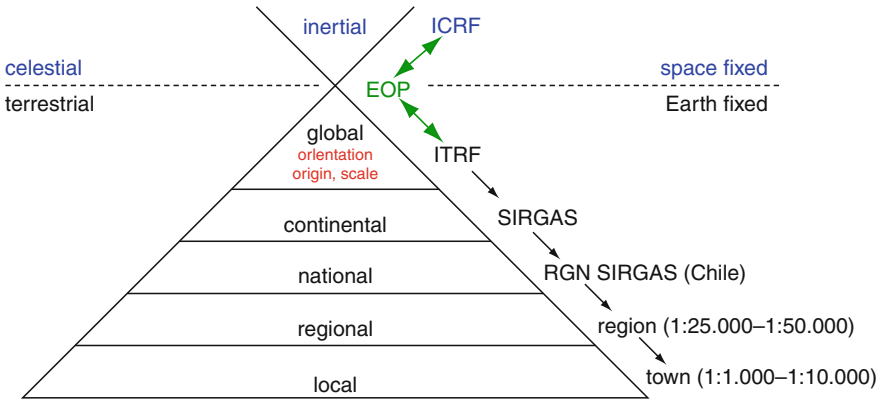
Continental reference systems

Like the surveying practice "working from the whole to the part," continental and national reference frames are established using subsets of reference points defining the global terrestrial reference frame. These frames can be ordered hierarchically beginning with the global terrestrial reference frame (TRF) in which the origin, orientation, and scale is defined. The continental reference network contains the corresponding subset of the global reference points providing the consistent frame and is densified by additional reference points. Consistency to the global reference frame is achieved by the network adjustment with the introduction of the global definition of origin, scale, and orientation represented by the subset of global reference points. National reference frames are established by composing the network of the corresponding global and continental reference points on national territory and additional densifying of national reference points. Thus, the densification of geodetic networks can be done in a consistent way to the most accurate global terrestrial reference frame, the ITRF. One example for consistency between hierarchical reference frames is the geocentric reference frame of the Americas, SIRGAS, and subsequent national reference frames in South-American countries (Figure 1) (SIRGAS). The GRS80 is mostly adopted as the reference ellipsoid for the continental reference frames for consistency reasons to the ITRF (Table 3).

Continental vertical reference systems have to overcome the national height system definitions and made consistent at the borders. Usually in continent-wide measuring campaigns at tide gauges and at national first-order reference points with leveling, GNSS and gravimetry are performed to resolve this task. An example is the European Vertical Reference System (EVRS) in its realization of the European Vertical Reference Frame 2007 (EVRF, 2007).

National geodetic reference systems

GNSS techniques turned the former particular national reference system into being part of a superior continental or global reference system. National geodetic networks are needed to provide a reliable common coordinate system for spatial datasets and geographical information data. At present, most national geodetic network infrastructures consist of continuously operating reference stations (CORS), which are linked to the ITRF. A GNSS CORS comprises a GNSS receiver with its antenna mounted on a stable monument at a safe and secure location equipped with a reliable power supply and a connection to the communication networks for data



Geodesy, Networks and Reference Systems, Figure 1 Hierarchical relationship between different reference frames. The International Terrestrial Reference Frame (ITRF) is tied to the space-fixed International Celestial Reference Frame (ICRF) via the Earth orientation parameters (EOP). The Earth-fixed ITRF defines the origin, orientation, and scale of this global reference frame. Subsequent frames on the continental, national, regional, or local level are consistent with each other if they are including reference points from the superior level (shown here is the example of Chile in South America). Therefore, the definition of the global reference frame is introduced. The shape of the triangle symbolizes the increasing number of reference points. On the top level, there are only a few global reference points (fundamental stations) necessary. On the bottom level, the maximum number of reference points is needed for small scale mapping and cadastre purposes.

Geodesy, Networks and Reference Systems, Table 3 Continental reference frames

Continental reference frames	Acronym	Meaning	Earth model
Africa	AFREF	African Geodetic Reference Frame	WGS84
Asia–Pacific	APRGP	Asia and Pacific Regional Geodetic Project	GRS80
Europe	EUREF	European Reference Frame	GRS80
North and South America	SIRGAS	Geodetic System for the Americas	GRS80

transmission. Different services make this infrastructure available to the users: (a) streaming real-time (raw) data over the Internet (NTRIP); (b) streaming real-time kinematic data (RTK) and differential GNSS data (DGNSS) for applications such as RTK-survey, GIS-data acquisition, and GNSS navigation; and (c) transfer of hourly or daily data files to an ftp-server.

The national coverage with CORS depends on the resources available and of the tectonic situation. A good example is the Japanese GPS Earth Observation Network System (GEONET) with about 1,200 CORS. This network with 30–50 km distance between each CORS is used not only for land survey but also for the monitoring of seismic and volcanic activities.

For historical reasons national geodetic systems may still refer to a best-fitting ellipsoid as a reference surface to the given national latitudes. With the introduction of the hierarchy of reference systems, the global Earth models are becoming more important especially for national reference systems due to consistency.

Vertical national reference systems are usually related to a tide gauge. The *zero-height* can be related to an average level of the seawater height (i.e., mean sea level). Some countries transferred the vertical reference from the tide gauge by spirit leveling to a distant benchmark at a secure location. Vertical reference systems are preferably related to the sea level as the accessible part of the geoid. Currently, vertical reference networks are measured not only by spirit leveling but also by GNSS combined with gravimetric measurements.

Historical reference systems

In the past, horizontal national reference systems were based originally on triangulation networks, when directions could be measured more precisely with theodolites than distances by rods. Triangulation networks had been oriented by astronomical azimuth measurements, setting the horizontal geodetic datum at the fundamental point. The scale was introduced by baseline enlargement networks.

The error in triangulation networks could reach 1–5 ppm with respect to the distance to the datum reference point. When electronics entered the geodetic instrumentation during the 1960s, electromagnetic signals (microwave, laser) allowed the measurement of distances more precisely than directions. Trilateration networks complemented or replaced the triangulation networks.

The introduction of space-geodetic methods in the 1980s, i.e., Navy Navigation Satellite System based on the Doppler effect and the follow-up GPS system, replaced most of the historical national reference systems based on classical triangulation or trilateration methods. The advantages are the global availability and the superior accuracy of the GNSS. The disadvantages are the dependency on the control segment of the GNSS and the availability of energy and communication in order to be able to use it.

The separation of horizontal and vertical reference networks is the consequence of the historical development of the geodetic methods. Modern geodetic space techniques such as VLBI, SLR, and GNSS are three-dimensional methods and can be used to overcome the classical separation. This becomes possible if the geometrical and geophysical properties of the geodetic observations are modeled correctly in the corresponding reference frame.

Global geodesy

International services

The task of global geodesy requires a global infrastructure of geodetic observatories. Currently, many institutions in various countries collaborate voluntarily to enable the production of terrestrial reference frames by sharing observational data, its processing, and geodetic products. Through the use of their human, technical, and financial resources, the contributing institutions are contributing to the products of the international services. This task is voluntarily organized by international services under the umbrella of the International Association of Geodesy (IAG). Even today, no country has been capable of reaching the same results on its own. This fact makes the nonprofit oriented work of international services indispensible to achieve the ITRF and the ICRF.

The international services can be classified according to their domains. In the space domain, the following are the services: International VLBI Service (IVS), International Laser Ranging Service (ILRS), International Global Navigation Satellite System Service (IGS), and International DORIS Service (IDS). In the time domain the Bureau International des Poids et Mesures (BIPM) in Paris hosts the office, which is generating the Universal Time (UT) scale as the time base. In the gravity field domain, the International Gravity Field Service (IGFS) forms the umbrella for the Bureau Gravimetrique International (BGI), International Geoid Service (IGeS), the International Center for Earth Tides (ICET), the International Center for Global Earth Models (ICGEM), and the International Digital Elevation Model Service (IDEMS).

The International Earth Rotation and Reference System Service (IERS) is primarily using the products of the IVS, ILRS, IGS, and IDS for the computation of the ICRF, ITRF, and EOP. All of the services mentioned above are part of the International Association of Geodesy (IAG).

With the development of the geodetic space techniques, it became obvious that an international coordination is beneficial for the users of geodetic products in navigation, land surveying, mapping, and scientific research. Therefore, most of the international services were created during the 1990s to continue previously existing collaborations related to projects.

The IAG introduced a project called Global Geodetic Observing System (GGOS) as its contribution to the Global Earth Observing System of Systems (GEOSS). GEOSS is an effort of the Group of Earth Observation (GEO), which unifies more than 80 countries and more

than 50 international organizations to improve the monitoring and hence the better understanding of global changes. The international services form the backbone for the realization of GGOS. The goal of GGOS is to provide a TRF with a global precision of 1 mm in position and 0.1 mm/year velocity determination (Plag and Pearlman, 2009). This ambitious goal can only be reached if more geodetic observatories with the characteristics of a fundamental station for geodesy will be implemented in a more homogeneous global distribution. The international services themselves try to improve their products so that they match the high precision demands of GGOS. This challenges the design of observing networks, improved observation schedules, new instruments and targets, improvements of the data acquisition and processing, and the homogenization and integration of the models of geodetic space techniques.

If the postulated precision can be achieved in the reference frames, trends in global changes will be resolved in shorter time periods. This compensates the lack of long-term monitoring data at many locations on Earth.

Fundamental station for geodesy

At fundamental reference points for global reference frames one encounters the necessary instrumentation for observing in the space domain, in the time domain and in the gravitational field domain of Earth. These domains correspond to the underlying physical model on which global geodetic reference systems need to be defined.

Each geodetic space method leads to a specific TRF, e.g., VLBI to a VLBI-TRF, SLR to a SLR-TRF or GNSS to one of the GNSS-TRF. Each one of the specific TRFs contains site positions and velocities of the specific instruments. Each one of the different techniques has some advantages, but also weaknesses regarding the realization of reference systems. Geodesists prefer to use all the techniques in a synergetic way in order to minimize the effect of otherwise undetectable systematic errors. Thus, the collocation of different geodetic space technique instruments at some sites is a must in order to achieve the best TRF. The space geodesy instruments also need to be complemented by local sensors in order to obtain a complete set of observables for the models used in data reduction.

A geodetic observatory in which all relevant space-geodetic observation methods are operated is called a fundamental station if it complies with the following characteristics: (a) *Permanency* of the operation by consideration of the timescales of geodynamic phenomena and therefore long-term assured production of geodetic data, (b) *Complementariness* of the operated geodetic space techniques for the best possible coverage of the System Earth affecting phenomena as to obtain the best possible realization of a precise global reference system, (c) *Redundancy* within the selected instruments for quality assurance of the observational data by independently obtained results of the same observables, (d) *Linking* of the technique-specific reference points by their corresponding space vectors, usually determined by a local survey within a local geodetic network, which is also used for the monitoring of the local site stability.

A fundamental station for geodesy enables the link and the transformation from one technique-specific reference frame (e.g., VLBI) to another (e.g., GNSS) by coordinate transformation. Fundamental stations are constitutive for the realization of the ITRF, which is based on observational data produced by VLBI, SLR, GNSS, and DORIS (Hase, 1999).

Instrumentation of a fundamental station

The Geodetic Observatory in Wettzell, Germany, and the Transportable Integrated Geodetic Observatory (TIGO) in Concepción, Chile (see Figure 2) are both examples of a fundamental station for geodesy. Both observatories operate radio telescopes for geodetic/astrometric VLBI, optical telescopes with a laser system for SLR, some GNSS receivers at the site plus an additional regional permanent network for monitoring the regional stability in the space domain. At both sites, *local geodetic networks* are installed that allow to tie the telescope reference points to the GNSS antenna reference points and hence the individual technique-specific reference frames.

As the geodetic space techniques basically measure time intervals, a fundamental station needs to be equipped with frequency standards defining the time unit: the second. The instrument for the time domain is a time and frequency laboratory, which operates at least three atomic clocks based on frequency standards. An issue in timekeeping is redundancy in instrumentation. Otherwise, drifts in time cannot be related to one specific clock. The geodetic observatories Wettzell and TIGO operate three cesium normals for long-term timescale generation and use one of the three available hydrogen masers for short-term time interval measurements.

Concerning the gravity field domain, a fundamental station must measure the gravity periodically with an absolute gravity meter and permanently with a superconducting gravity meter in order to determine the Earth tides. The daily gravitational deformations of Earth's crust reach several decimeter, which is therefore the largest time variable spatial parameter in intercontinental baseline measurements and needs to be considered in the analysis of geodetic space techniques.

Fundamental stations for geodesy produce observational data in the space, time, and gravitational field domain.

In order to perform the observations correctly or to interpret the measurements afterward, additional local sensors are necessary. Meteorological sensors deliver input data for the application of refraction corrections during observation with the telescopes and compute signal travel times through the atmosphere. Hydrological sensors are requested to model seasonal variations in the gravity data. Laser gyroscopes deliver local Earth rotation parameters (as opposed to global determined EOP by geodetic

Geodesy, Networks and Reference Systems, Figure 2 The Transportable Integrated Geodetic Observatory (TIGO) in Concepción, Chile. The foreground shows the Satellite Laser Ranging (SLR)-System, behind it a pillar with Global Navigation Satellite Systems (GNSS)-antennas, and on the right the Very Long Baseline Interferometry (VLBI) radio telescope. All geodetic space techniques are placed at one site to define reference points for the global reference frame ITRF. The space vectors between the different instruments can be measured by a local survey. Not visible are the gravity meter and atomic clocks that are part of TIGO.

space techniques) and seismometers deliver epochs of seismic events to model the nonlinearity of continental drift respective to the site motion.

Summary

The task of geodesy is to establish global geodetic reference systems. These reference systems are defined theoretically in context of the knowledge of physics. Global reference systems are realized by geodetic observatories and measuring reference stations that mark reference points in the space and time domain and in the gravitational field of Earth. Metrology provides the corresponding physical constants and units to be used in the measurements. The cooperative use of the same geodetic method at various measuring platforms creates a geodetic network. A set of observational data from a geodetic network defines a specific geodetic reference frame that relates usually to a specific epoch. Continuous changes of Earth's surface require a permanent monitoring. Different terrestrial reference frames can be tied together by collocating geodetic space techniques.

Fundamental stations for geodesy establish reference points in technique-specific TRF that are tied together by the knowledge of the local spatial vectors between the technique-specific reference points. The ITRF is based on observational data of various space-geodetic methods tied in a synergetic manner at collocated and fundamental stations. The WGS84 is based only on GPS-observations and coincides with the multi-technique-based ITRF within 1–2 cm. Global change studies drive the global precision of TRF to 1 mm in position and 0.1 mm in point velocity. The International Association of Geodesy undertakes an effort with the Global Geodetic Observing System to set up the necessary infrastructure.

The use of geodetic satellite methods in positioning requires a precise knowledge of orbital parameters. Two out of six parameters are dependent on the variable Earth's rotation. Monitoring Earth's rotation requires the creation of a Celestial Reference System. The ICRF is based on quasar positions measured by global VLBI networks. The transformation parameters from the space-fixed celestial reference frame to the Earth-bound terrestrial reference frame are given by the time variable Earth orientation parameters (EOP).

Bibliography

BIPM, 1998. *The International System of Units (SI)*, 7th edn. Paris: Bureau International des Poids et Mesures.

Bursa, M., 1992. Parameters of common relevance of astronomy, geodesy and geodynamics, *Bulletin Géodésique*, **66**(2), 193–197.

Damour, T., Soffel, M. H., and Xu, C., 1991. General relativistic celestial mechanics: I. Method and definition of reference systems. *Physical Review D: Particles and Fields*, **43**(10), 3273–3307.

Hase, H., 1999. Theorie und Praxis globaler Bezugssysteme, Mitteilungen des BKG. Frankfurt am Main: Bundesamtes für Kartographie und Geodäsie, Band 13.

Helmert, F. R., 1880. *Die mathematischen und physikalischen Theorien der höheren Geodäsie, Band I*. Leipzig: Verlag Teubner.

IERS Conventions, 2003. Dennis D. McCarthy and Gérard Petit. *IERS Technical Note 32*. Frankfurt am Main: Verlag des Bundesamts für Kartographie und Geodäsie, 2004. 127 pp., paperback, ISBN 3-89888-884-3 (print version) http://www.iers.org/IERS/EN/Publications/TechnicalNotes/tn32.html.

IUGG, 1991. Resolution on the conventional terrestrial reference system. *Journal of Geodesy*, **66**(2), 128–129. doi:10.1007/BF00989258.

Kovalevsky, J., and Mueller, I. I., 1981. Comments on conventional terrestrial and quasi-inertial reference systems. In Gaposchkin, E. M., and Kolaczek, B., (eds.), *Proceedings of the 56th Colloquium of the IAU*, September 8–12, 1980, Academic: Dordrecht.
Moritz, H., 1980. Geodetic reference system 1980. *Bulletin Géodésique*, **54**(3), 395–405.
Plag, H.-P., and Pearlman, M., 2009. *Global Geodetic Observing System*. Berlin: Springer.
Schneider, M., 1992. *Himmelsmechanik: Band I, Grundlagen, Determinierung*. Mannheim: BI Wissenschaftsverlag, p. 250.
Schneider, M., 1999. *Himmelsmechanik: Band IV, Theorie der Satellitenbewegung, Bahnbestimmung*. Heidelberg: Spektrum Akademischer. Kap. 67.2.
SIRGAS. http://www.sirgas.org/index.php?id=55.
Torge, W., 2003. *Geodäsie*. Berlin: Walter de Gruyter.

Cross-references

Earth Rotation
Earth Tides
Geodesy, Figure of the Earth
Geodesy, Ground Positioning and Leveling
Geodesy, Physical
Geodetic Pendulums, Horizontal Ultra Broad Band
Geoid Determination, Theory and Principles
Geoid, Computational Method
GPS, Data Acquisition and Analysis
GPS, Tectonic Geodesy
Gravity Field of the Earth
Gravity Method, Satellite
Gravity, Global Models
Satellite Laser Ranging
Very Long Baseline Interferometry

GEODESY, PHYSICAL

V. Chakravarthi
National Geophysical Research Institute, Hyderabad, Andra Pradesh, India

Geodesy is a branch of earth sciences that deals with the measurement and representation of the earth's shape and its gravitational field. Geodesy may be broadly divided into the areas of global (satellite) geodesy, geodetic surveys (national and supranational), and plane surveying (topographic surveying, cadastral surveying). Global geodesy includes the determination of the shape of the earth, its orientation in space and its external gravity field. A geodetic survey is for the determination of the earth's surface and gravity field over a region that typically spans a country or a group of countries. In plane surveying, the details of the earth's surface are determined on a local scale, and thus the curvature and gravity effects are generally ignored. A close relation exists between the above three areas of geodesy in the sense that geodetic surveys are linked to reference frames established by global geodesy, whereas plane surveys are generally referenced to control points established by geodetic surveys. Physical geodesy utilizes measurements and characteristics of the earth's gravity field as well as theories relating it to deduce the shape of the geoid (see *Geoid Determination, Theory and Principles*; *Geoid, Computational Method*) and thus the figure of the earth (Haffman and Moritz, 2005). With sufficient information regarding the earth's gravity field, it is possible to determine geoid undulations, gravimetric deflections, the earth's flattening and other broad characteristics of the earth, such as equatorial ellipticity – the C22, S22 terms of the spherical expansion of the gravity field, the C21, S21 terms related to the position of the rotation axis, etc., with a view to their application in geodesy.

Historical Development of Geodesy

Geodesy is one among the oldest sciences dealing with the planet earth. Superseding the use of sphere as a model for the earth, the oblate rotational ellipsoid became widely accepted as the model of choice in the early eighteenth century. The significance of the gravity field was recognized in the nineteenth century leading to the introduction of the geoid. Later in the second half of the twentieth century, satellite techniques paved the way for realization of the three-dimensional concept of geodesy. A drastic increase in the accuracy of geodetic observations required that time variations be taken into account, which in turn leads the development of four-dimensional geodesy. Extensive material on geodetic history can be found in Smith (1986).

Gravity field of the earth

A body rotating with the earth experiences the gravitational force of the masses of the earth, other celestial bodies, as well as the centrifugal force due to the earth's rotation (see *Earth Rotation*). The resultant is the force of gravity (see *Gravity Field of the Earth*). If the effects of centrifugal force and that of celestial bodies are removed from gravity, the left out component is gravitation. Newton's law of gravitation states that the force of attraction between two point masses m_1 and m_2 is directly proportional to the product of their masses and inversely to the square of the distance between the centers of respective masses and is given as

$$F = \frac{-Gm_1m_2}{r^2}\frac{k}{r}, \tag{1}$$

where, F is the force on m_2, r is the distance between m_1 and m_2, and G is the universal gravitational constant. The minus sign represents that the force is always attractive. By setting the mass at the attracted point to unity, the above equation transforms into the gravitational acceleration

$$g = \frac{-Gm}{r^2}\frac{k}{r}. \tag{2}$$

The vector k may be expressed by the position vectors r and r' in the Cartesian coordinate system as

$$k = r - r', r^T = (X, Y, Z), r'^T = (X', Y', Z'),$$

with the magnitude of

$$k = \sqrt{(X - X')^2 + (Y - Y')^2 + (Z - Z')^2}.$$

However, the earth is composed of an infinite number of differential mass elements, dm. The gravitation on the unit mass results from the integral over the individual contributions. The equation for gravitational acceleration then takes the form

$$g = -G \iiint \frac{r - r'}{|r - r'|^3} \mathrm{d}m. \tag{3}$$

The mass element dm can be expressed as

$$\mathrm{d}m = \rho \mathrm{d}u, \tag{4}$$

where, ρ is density of volume element du.

The representation of gravitational acceleration and related computations are simplified if the scalar quantity potential is used. Because the gravitational field is invariant to rotations

$$\mathrm{curl}\, g = 0 \tag{5}$$

The vector g may be represented as the gradient of a potential U (Sigl, 1985),

$$g = \mathrm{grad}\, U. \tag{6}$$

For a point mass m, we have

$$U = \frac{GM}{r}, \quad with \quad \lim_{r \to \infty} U = 0. \tag{7}$$

For the earth, we have

$$U = G \iiint \frac{\mathrm{d}m}{r} = G \iiint \frac{\rho}{r} \mathrm{d}u, \quad \lim_{r \to \infty} U = 0. \tag{8}$$

The potential indicates the work that must be done by gravitation in order to move the unit mass from infinity to the point under consideration.

From the above relations, it is clear that once the density function ρ was known for the earth, the gravitation can be calculated as a function of position. In reality, detailed density information is available for the upper layers of the earth only, while global models merely consider radial density changes (see *Earth, Density Distribution*). Consequently, gravity field observations have to be used in order to model the exterior gravity field (Torge, 2001).

Properties of the gravitational potential

Gravity potential U satisfies Laplace's equation (Murthy, 1998), i.e.,

$$\nabla^2 U = 0, \tag{9}$$

at all points unoccupied by matter. All the space derivatives of this potential also satisfy Laplace's equation. However, at points occupied by matter, gravity potential satisfy Poisson's equation

$$\nabla^2 U = -4\pi G d, \tag{10}$$

where d is the density at the point under consideration. Green's theorem connecting volume and surface integrations of potential fields has many important applications. If U and U' are two potential functions, whose first derivatives are finite, continuous, and single valued throughout a volume, v, bounded by a closed surface, s, then

$$\int_v (U \nabla^2 U' - U' \nabla^2 U) \mathrm{d}v = \int_s \left[U \frac{\partial U'}{\partial n} - U' \frac{\partial U}{\partial n} \right] \mathrm{d}s, \tag{11}$$

where n is the direction of the outward drawn normal to the surface. The theorem is independent of the size or shape of the surface or the nature of distribution of masses causing U and U'. If U' is assumed to be constant everywhere, and U is the Newtonian potential due to masses distributed both inside and outside the closed surface, s, then

$$\nabla^2 U' = 0, \quad \frac{\partial U'}{\partial n} = 0 \tag{12}$$

reducing equation to the form

$$\int_v \nabla^2 U \mathrm{d}v = \int_s \frac{\partial U}{\partial n} \mathrm{d}s. \tag{13}$$

If U_{in} and U_{out} are, respectively, the potentials due to masses inside and outside s, then U_{out} obeys Laplace's equation, while U_{in} satisfies Poisson's equation, i.e.,

$$\int_s \frac{\partial U}{\partial n} \mathrm{d}s = -4\pi GM. \tag{14}$$

This is called Gauss' law, which states that the total normal gravitational flux over a closed surface is $-4\pi G$ times the enclosed mass. This law holds good irrespective of the type of mass distribution in s. Basic relationships can therefore be established between the observations in the gravity field and the parameters describing the surface.

Disturbing potential

The difference between the actual (measured) gravity potential, W, and the normal gravity potential (constructed matching the known reference ellipsoid with an equipotential surface), U, is called the anomalous or disturbing potential, T (Hofmann and Moritz, 2005). The disturbing potential, T, is numerically smaller than U or W, and captures the detailed, complex variations of the true gravity field of the earth from point-to-point, as distinguished from the overall global trend captured by the smooth mathematical ellipsoid of the normal potential. The relation between the geoidal undulation, N, disturbing potential, T, and the force of gravity (F) computed from the normal gravity potential can be expressed by Bruns' formula as

$$N = \frac{T}{F}. \tag{15}$$

The Stoke's formula which enables one in determining the geoid from gravity data in physical geodesy can be expressed as

$$N = \frac{R}{4\pi F_0} \iint_{\sigma} \Delta g \; S(\Psi) \mathrm{d}\sigma. \tag{16}$$

Here, Δg stands for gravity anomalies, R is mean radius of the earth, F_0 is a mean gravity value over the earth, and $S(\Psi)$ is the Stokes function, a kernel function derived by Stokes in a closed analytical form. The symbol $\iint_{\sigma}$ signifies an integral extended over the whole unit sphere.

Geoid and reference ellipsoid

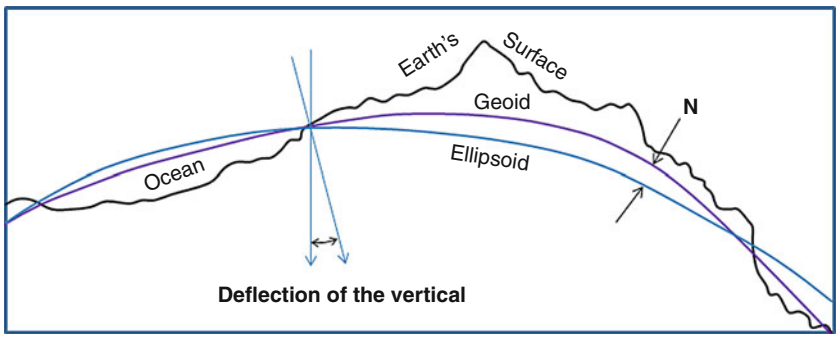
The equipotential surface of the earth's gravity field that coincides with the Mean Sea Level (MSL) in the absence of currents, air pressure variation, etc., is called the geoid. Most of the efforts in physical geodesy are concentrated on the geoid determination with steadily increasing accuracy (Heiskanen and Moritz, 1967). The geoid is important for understanding more about ocean currents and is also used as a reference for traditional height systems and monitoring sea-level change. However, as in the case of any other equipotential surface of the earth's gravity field, the geoidal surface is also an irregular surface and too complicated to serve as the computational surface. The mathematical surface that best fits the geoid is the reference ellipsoid. The relation between the geoid and the ellipsoid is shown in Figure 1. The ellipsoid lies below the geoidal surface in elevated regions and above the MSL over the oceans. The geometrical separation between the geoid and the reference ellipsoid is called the geoidal undulation, N (see *Geoid Undulation, Interpretation*), which generally varies globally between ±110 m.

Gravity measurements are used in physical geodesy to determine the figure of the earth (see *Geodesy, Figure of the Earth*). The earth shape can be equated to a sphere or an ellipsoid depending on the physical problem being dealt. The justification in equating the earth's shape to a model other than that of sphere can be found from two observations, namely, (1) for a rotating sphere having radius, R, and angular velocity, ω, the gravitational attraction at the equator and pole should differ by $R\omega^2$. For the values of R and ω relevant to the geoid, this figure works out to be 3.4 gal as against the observed one of 5.17 gal in gravity values at the equator and the pole. This discrepancy suggests an appreciable departure of the mean shape of the geoid from that of a sphere, also (2) arc measurements at the pole and equator showed that the meridian degree at the pole is longer than that at the equator. The centrifugal acceleration causes the Earth's surface to flatten at the poles and to bulge at the equator. These effects would adequately explain the observed discrepancy in difference of gravity at the equator and the pole. It is thus reasonable to equate the earth's shape to that of an oblate ellipsoid. The earth's gravity field on the ellipsoidal surface varies only with the geographic latitude, φ, and is defined by the International Gravity Formula (Wilcox, 1989, also see *International Gravity Formula*)

$$g(\varphi) = 978.0327(1 + 0.0053024 \sin^2\varphi - 0.0000058 \sin^2 2\varphi)$$

Methods of in situ measurements

Two distinctly different types of gravity measurements are made: absolute gravity measurements and relative gravity measurements. If the value of acceleration of gravity can be determined at the point of measurement directly from the data observed at that point, the gravity measurement is absolute (see *Gravity Measurements, Absolute*). If only the differences in the value of the acceleration of gravity are measured between two or more points, the measurements are relative (see *Gravity Method, Surface*). The most important gravity field quantities measured on the earth's surface are gravity potential differences determined by leveling in combination with gravity measurement; the modulus of the gravity vector determined by absolute and relative gravity measurements; the direction of the gravity vector determined by geodetic-astronomical observations; and the components of the second-order gravity gradient tensor determined by torsion balance measurements.



Geodesy, Physical, Figure 1 Relation between earth's topography, geoid and reference ellipsoid.

Satellite geodesy

Satellite geodesy utilizes artificial satellites and the moon as extraterrestrial targets or sensors. For a point mass earth model, the orbital motion of a satellite is described by Kepler's laws. The actual gravitational field and nongravitational forces create orbital perturbations. Satellites used for geodetic applications differ in design, equipment, and orbit according to the respective observation technique and the mission purpose. Classical measurement methods based on orbital perturbations, introduced and employed from 1960s to 1980s demonstrated the efficacy of satellite observations for establishing large region geodetic control networks and gravitational field determination (see *Gravity Method, Satellite*). Now the Global Positioning System (GPS) has become the most important method for a fast and precise determination of geodetic positions (see *Geodesy, Ground Positioning and Leveling*). Today, the GPS governs three-dimensional positioning at all scales, while laser distance measurements primarily contribute to global reference networks (see *Geodesy, Networks and Reference Systems*). By monitoring the ocean surface, satellite altimetry contributes to gravity field modeling, and high-resolution global gravity recovery is now possible from satellite-to-satellite tracking and gravity gradiometry missions (e.g., GRACE and GOCE missions). The theory of satellite orbits and measurement methods was described in detail by Seeber (1993) and Kaula (1966).

Determining the geoid

An intimate relationship exists between the height of the geoid, the earth's gravitational attraction, and the distribution of mass within the earth. The shape of the geoid does place important constraints on the mass distribution. Moreover, since the local vertical is defined as the direction of the gravitational attraction, which is always perpendicular to the geoid, geoidal anomalies cause deflections of the vertical, which are of vital practical importance for surveyors. Determining the geoid is therefore an important objective in geodesy and geophysics.

In the pre-satellite era, the geoid was determined from the gravity field, which entailed measuring gravitational attraction or deflections of the vertical (or both) at a large number of locations around the Earth. Accuracy and resolution were limited by the number and distribution of observations that were feasible. Since the advent of artificial satellites, geoid determinations have become easier, more accurate, and of higher resolution (see *Geoid, Computational Method*; *Geoid Determination, Theory and Principles*; *Gravity Method, Satellite*). Since the early 1980s, the mean sea level (or oceanic geoid) has been measured directly by a radar-altimeter mounted on a satellite. It is important to remove the effects of waves, storms, ocean currents, and other similar effects, but ultimately this method gives a precision of a few centimeters in geoid height. Although they cannot be used over land, such measurements have yielded unprecedented high-resolution coverage of the geoid over the sea (Rapp, 1982a, b; Sandwell and Schubert, 1982; Cazenave and Monnereau, 1987; Wunsch and Stammer, 1998; Fu and Cazenave, 2001). In recent past, under the NASA Earth System Science Pathfinder (ESSP) Program, the GRACE mission was launched to accurately map the variations in the Earth's gravity field. The mission has provided the scientists an efficient and cost-effective way to map the Earth's gravity fields with high accuracy. The results from this mission yielded crucial information about the distribution and flow of mass within the Earth and its surroundings. The gravity variations that GRACE study include changes due to surface and deep currents in the ocean, runoff and ground-water storage on land masses, and exchanges between ice sheets or glaciers and the oceans. More details on the mission can be found at http://www.csr.utexas.edu/grace. Further, satellite gradiometer mission (see *Gravity, Gradiometry*) named as Gravity field and steady-state Ocean Circulation Explorer (GOCE) was designed and launched by the European Space Agency (ESA) to determine gravity field anomalies with an accuracy of 1 mGal and geoid with an accuracy of 1–2 cm. The GOCE will be gathering data around 20 months to map the Earth's gravity field with unparalleled accuracy and spatial resolution. The final gravity map and model of the geoid will provide users worldwide with a well-defined data product that will lead to a better understanding of the physics of the Earth's interior to gain new insights into the geodynamics associated with the lithosphere, mantle composition and rheology, uplift and subduction processes, a better understanding of the ocean currents and heat transport, a global height-reference system, to serve as a reference surface to study topographic processes and sea-level change, better estimates of the thickness of polar ice sheets and their movement. Further details of the mission can be found at http://www.esa.int/esaLP/ ESAYEK 1VMOC_LP goce_0.html.

Gravity reduction

Gravity measurements provide values for the acceleration of gravity at points located on the physical surface of the earth. Before these measurements can be used for most geodetic purposes, they must be converted into gravity anomalies. Gravity measured on the physical surface of the earth must be distinguished from normal gravity referring to the surface of the reference ellipsoid. The measured gravity values show point-to-point variation on the earth's surface because the measured field depends on latitude and elevation of the station and also due to the effects of topographic masses above sea level (Murthy, 1998). The important corrections to the measured gravity data include latitude or normal correction, free-air correction, Bouguer correction, topographic correction, and isostatic correction (see *Gravity, Data to Anomalies*). The latitude correction calculates the normal gravity on the ellipsoidal surface, the free-air correction extrapolates it to the level of observation, the topographic and Bouguer

corrections account for the gravity effect of the material between the geoid and the ground surface. Isostatic correction removes long-wavelength variations in the gravity field inversely related to topography. The magnitudes of these corrections more often equal to the measured field itself and therefore should be evaluated carefully and precisely.

Gravity models

Two important models are in vogue: (1) effect models and (2) source models (Sünkel, 1989). Effect models deals with the calculation of four parameters to define the best fitting ellipsoid to the shape of the earth. These parameters are the product of the gravitational constant and the total mass of the earth, a dimensionless dynamical form parameter in terms of a second degree spectral coefficient of the earth's gravity field which is closely related to the flattening of the ellipsoid of revolution, the equatorial radius, and the angular velocity. However, the surface defined by the above parameters describes the global feature of the shape of the earth and its gravity field (see *Geoid, Computational Method*; *Gravity Method, Principles*). For more elaborate models, spherical harmonic coefficients have to be derived from surface and satellite data (Rapp and Cruz, 1986). Ditmar et al. (2003) discussed the regularization issues of the computation of spherical harmonic coefficients of the gravity gradiometric data acquired by the GOCE satellite. The effects of high-frequency temporal gravity field sources due to ocean tides, atmosphere, and hydrological mass variations on the GOCE satellite-to-satellite tracking and gradiometer measurements and on the corresponding geopotential model recovery are reported by Han et al. (2006) (see *Gravity Field, Temporal Variations from Space Techniques*). On the other hand, source models presume a density model having radially changing but laterally constant mass density. Such a density model can be described by a few parameters and can therefore be compared with an ellipsoidal model.

Conclusions

The gravity field of the earth, which forms the backbone for the studies relating to geometrical geodesy and physical geodesy, can be described in terms of potential, its derivatives both vertical and horizontal in several scales involving global, regional, and/or local information. The gradiometer missions will probably replace, even though not fully but to a large extent, the surface observations, thereby paving the way for more detailed information on geoid with unprecedented accuracy.

Bibliography

Cazenave, A., and Monnereau, M., 1987. Seasat gravity undulations in the central Indian Ocean. *Physics of the Earth and Planetary Interiors*, **48**, 130–141.

Ditmar, P., Kusche, J., and Klees, R., 2003. Computation of spherical harmonic coefficients from gravity gradiometry data to be acquired by the GOCE satellite: regularization issues. *Journal of Geodesy*, **77**, 465–477.

Fu, L. L., and Cazenave, A., 2001. *Satellite altimetry and earth sciences, a handbook of techniques and applications*. London: Academic.

Haffman, B. W., and Moritz, H., 2005. *Physical Geodesy*. New York: Springer Wien.

Han, S. C., Shum, C. K., Ditmar, P., Visser, P., Van Beelen, C., and Schrama, E. J. O., 2006. Aliasing effect of high-frequency mass variations on GOCE recovery of the earth's gravity field. *Journal of Geodynamics*, **41**, 69–76.

Heiskanen, W. A., and Moritz, H., 1967. *Physical Geodesy*. New York: W. H. Freeman.

Kaula, W. M., 1966. *Theory of Satellite Geodesy*. Waltham, Toronto, London: Blaisdell.

Moritz, H., 1980. *Advanced Physical Geodesy*. Karlsruhe: Wichmann.

Murthy, I. V. R., 1998. *Gravity and Magnetic Interpretation in Exploration Geophysics*. Bangalore: Geological Society of India.

Rapp, R. H., 1982a. *A global atlas of sea surface heights based on the adjusted Seasat altimeter data*. Ohio State University Department of Geodetic Science and Surveying Report 333.

Rapp, R. H., 1982b. *A summary of the results from the OSU analysis of seasat altimeter data*. Ohio State University Department of Geodetic Science and Surveying Report 335.

Rapp, R. H., and Cruz, Y. J., 1986. *Spherical harmonic expansions of the Earth's gravitational potential to degree 360 using 30' mean anomalies*. Ohio State University Department of Geodetic Science and Surveying Report 376.

Sandwell, D. T., and Schubert, G., 1982. Geoid height-age relation from SEASAT altimeter profiles across the Mendocino fracture zone. *Journal of Geophysical Research*, **87**, 3949–3958.

Seeber, G., 1993. *Satellite Geodesy*. Berlin, New York: De Gruyter.

Sigl, R., 1985. *Introduction to Potential Theory*. Cambridge: Abacus Press.

Smith, J. R., 1986. *From Plane to Spheroid*. Rancho Cordova: Landmark Enterprises.

Sünkel, H., 1989. *Geodesy Physical*. Berlin: Springer. Encyclopedia of Earth Sciences Series.

Torge, W., 2001. *Geodesy*. Germany: Walter de Gruyter GmbH & Co.

Wilcox, L., 1989. *Gravity Anomalies: Interpretation*. Berlin: Springer. Encyclopedia of Earth Sciences Series.

Wunsch, C., and Stammer, D., 1998. Satellite altimetry, the marine geoid, and the oceanic general circulation. *Annual Review of Earth and Planetary Sciences*, **26**, 219–53.

Cross-references

Earth Rotation
Earth, Density Distribution
Geodesy, Figure of the Earth
Geodesy, Ground Positioning and Leveling
Geodesy, Networks and Reference Systems
Geodetic Pendulums, Horizontal Ultra Broad Band
Geoid Determination, Theory and Principles
Geoid Undulation, Interpretation
Geoid, Computational Method
Gravity Field of the Earth
Gravity Field, Temporal Variations from Space Techniques
Gravity Measurements, Absolute
Gravity Method, Satellite
Gravity Method, Surface
Gravity, Data to Anomalies
Gravity, Global Models
Gravity, Gradiometry
International Gravity Formula

GEODETIC PENDULUMS, HORIZONTAL ULTRA BROAD BAND

Carla Braitenberg
Department of Geosciences, University of Trieste, Trieste, Italy

Synonyms

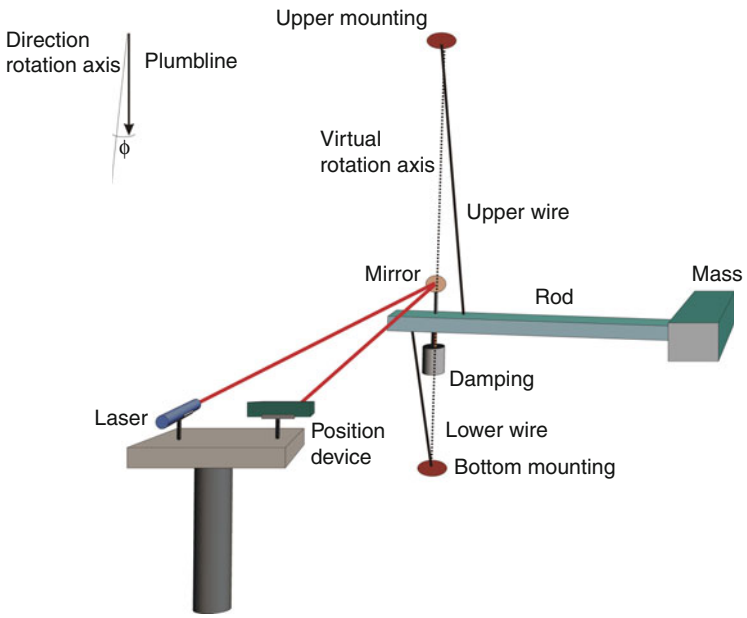
Ultra broad band long base tiltmeter

Definition

The ultra broad band horizontal geodetic pendulum is a device designed to measure the earth crustal movement in the range of periods typical of secular movements to seismic deformations. It measures the variation of the plumb line (vertical) with respect to the direction of two earth-fixed reference points (Zadro and Braitenberg, 1999).

Instrument description

The vertical pendulum is made of a mass suspended by a rod or wire that can oscillate in the vertical plane about a horizontal axis. For small amplitudes the oscillation period T depends on the moment of inertia K, the distance s of the center of mass from the rotation axis, the mass m, and the gravity acceleration g following the equation $T^2 = \frac{4\pi^2 K}{mgs}$. If the pendulum is allowed to oscillate about an inclined axis, the angle of the rotation axis with the vertical being φ, the period of oscillation is defined by $T^2 = \frac{4\pi^2 K}{\sin\varphi\, mgs}$. For small values of the angle φ, the rotation axis is near to vertical and the pendulum is called horizontal pendulum. Compared to the vertical pendulum of equal dimensions, the horizontal pendulum has an increased oscillation period. The first realizations of a horizontal pendulum were done in 1830 (Lorenz Hengler), and developed further by Zöllner (Zöllner, 1871, 1872). The construction of Zöllner is made of a horizontal rod on which at one end a mass is attached. At the other end and at a small distance from it the rod is connected by two wires to a housing structure (see Figure 1). The two wires are near to vertical above each other. Supposing the upper fixed point is shifted with respect to the lower fixed point along the Meridian, so that the line connecting the two fixed points makes an angle φ with the vertical, then the rod has its equilibrium position in the Meridian plane. If the direction of the vertical changes due to the influence of the gravitational attraction of the moon or sun toward east by a small angle α, the rod will rotate toward east by the angle $\frac{\alpha}{\sin\varphi}$. A similar rotation of the rod will be produced by the inclination of the housing structure or by a horizontal eastward movement of the housing structure as could be provoked by the passage of a seismic wave. In order to have an instrument sensitive in both NS and EW directions, usually a couple of orthogonally mounted pendulums is installed at one station. The movement of the rod is recorded by a space-sensitive device, as a magnetic induction coil mounted at the extreme of the rod or by an optical device that records the light ray reflected by a mirror mounted on the rod (e.g., Braitenberg et al., 2006).



Geodetic Pendulums, Horizontal Ultra Broad Band, Figure 1 Cartoon of a horizontal geodetic pendulum. The rod with the mass rotates in the horizontal plane about the virtual rotation axis. The angle φ of the virtual rotation axis with the plumb line is essential for the amplification factor of the pendulum.

Grotta Gigante horizontal pendulums

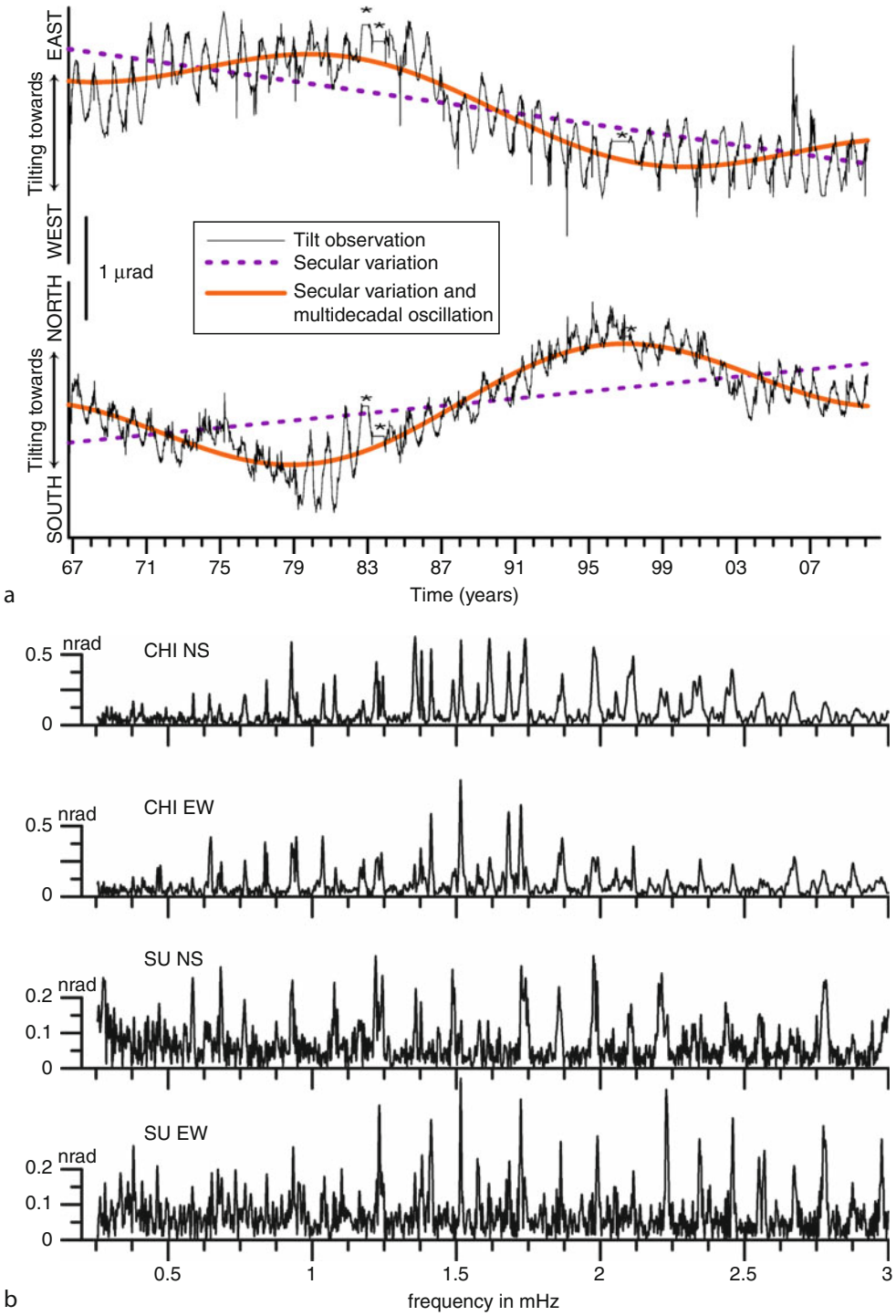
The largest prototype of a horizontal pendulum was built in 1959 by Antonio Marussi (Marussi, 1959) in a natural cave (Grotta Gigante) situated in the Italian Carst at 275 m above sea level near the town of Trieste, Italy. Trieste is a port on the Adriatic Sea, 150 km East of Venice. The upper and lower wires of the horizontal pendulum-rod are fixed directly into the rock, at a vertical distance of 95 m. The oscillation period of the pendulum is near to 6 min, and damping is critical. The original recording of the pendulum was on photographic paper, which was changed to a digital system in the year 2002. The instrument is very stable, and records multi-decennial crustal movements. The digital system has a sampling rate of 25 Hz, which makes it possible to observe fast movements of the pendulum caused by the passage of seismic waves. The smallest resolved tilt is 0.009 nrad (1 nrad = 10^{-9} rad).

The continuous record of tilt extends from 1966 to the present and demonstrates that the cave deforms continuously due to very different causes that act independently from each other. Each generates a characteristic signal contributing to the complex deformation of the cave. The graph in Figure 2a displaying the entire tilt sequence shows a secular drift that adds to an oscillation with a period of 33 years. The long-term movement is a tilt of the cave toward NW, parallel to the coastline, and is probably tied to a deformation of the Adria plate that is bent below the load of the sediments in front of the South Alpine mountain chain. The cause of the multi-annual oscillation is presently unknown, and could be tied to large-scale deformations imposed by the active deformation due to the collision of the Adria and the Eurasian plate, causing the seismicity at the contact between the sedimentary plain and the South Alpine mountains. The tilting of the cave with an annual period (mean period = 365.6 days and mean amplitude = 300 nrad for years 1966–2008; oscillation amplitude defined as half the peak-to-peak signal) has a characteristic orientation with extreme values in mid September and mid March. The annual tilting is due to the combined effect of the deformations due to the annual surface temperature variations (sinusoidal with an amplitude of 9.2°C) and to the variable loading of the Adriatic Sea.

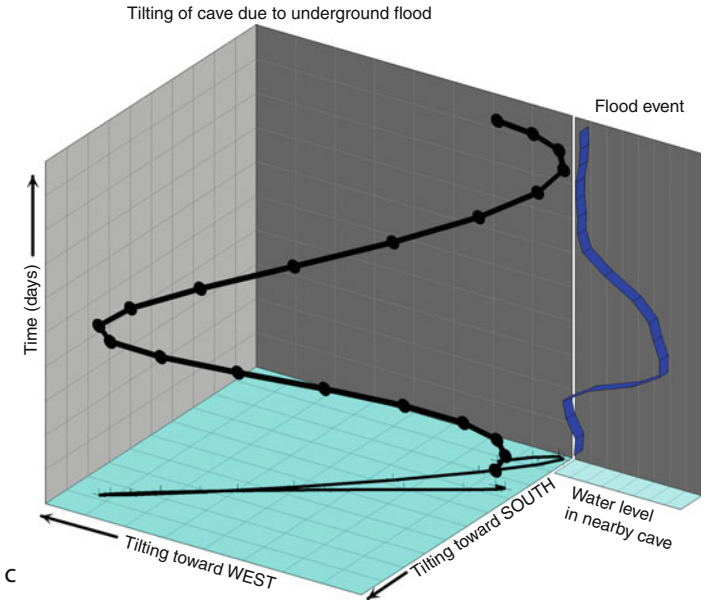
The earth tides generate a tilting that is due to the regular deformation of the earth in response to the gravitational attraction of moon and sun, with predominantly diurnal and semidiurnal periods. The amplitude of the earth tide tilting is 39 nrad, the gravitational attraction and the load of the ocean tide contribute to an additional tilt signal at the same tidal frequencies, but phase shifted with respect to the astronomical earth tide, leading to a total tidal tilting signal with a maximum amplitude of 90 nrad. The fault rupture of an earthquake leads to the generation of the well-known seismic waves that propagate along the earth surface and through earth's interior. Another movement caused by earthquakes is the oscillation of the earth at its resonance-frequencies, called the free oscillations. The crustal rupture causes the earth to pulsate in the spherical and torsional oscillation modes. The gravest spheroidal mode 0S2 where the earth deforms with an ellipsoidal pattern has a 54-min period, the more complicated modes having shorter periods. The gravest torsional mode 0T2 has a period of 43.9 min, one earth hemisphere rotating with respect to the other around a central axis (Bolt and Marussi, 1962). The geodetic pendulums of the Grotta Gigante station are the only instruments in the world that recorded the free oscillations generated by the greatest (M = 9.5, Chile, 1960; NEIC, 2010), third greatest (M = 9.1, Sumatra Andaman Islands, 2004; NEIC, 2010), and fifth greatest earthquake (M = 8.8; Chile, 2010; NEIC, 2010) ever recorded in instrumental seismology. The observations showed that the amplitude of the free oscillations of the Chile 1960 earthquake were up to four and a half times greater than those generated by the Sumatra-Andaman 2004 event (Braitenberg and Zadro, 2007), the greatest enhancement being found for the torsional modes (see spectrum in Figure 2b). The cave deforms not only due to periodic signals, similar to the ones we described before, but also due to transients, as pressure variations and underground water flow. The underground water flow is particularly interesting in carstic areas: The water from rainfall penetrates into the rock through pores and cracks until it finds an impermeable layer along which it flows downstream toward the sea. The Italian Carst is particular in this respect, as an entire surface river with an average flow of 9 m^3/s reaching peaks of 100 m^3/s during strong rainfalls, with extreme values up to 390 m^3/s, disappears after its passage through a natural cave (Skocjanske Jame, 2010). The underground river remains concealed for 33 km, until it emerges as a spring from the foot of the Carst and flows into the Adriatic Sea. The path of the flow is unknown, but its presence is identified by the geodetic pendulums. The cave responds in a characteristic direction to floods, with a signal that lasts in total between 4 and 8 days. The length of the deformation is indicative of the time the underground river takes to accommodate for the increased influx of water through the rainfall. The hydrologic-induced tilt signal is up to 300 nrad according to the existing records. Table 1 summarizes the different causes that generate a tilting of the cave together with the expected amplitude.

Besides the horizontal geodetic pendulums, other types of instruments exist with which tilt is observed. In the horizontal water tube, the differential water level at each end of the tube is measured by laser interferometry (e.g., Ferreira et al., 2006). The length of the tube extends over several tens of meters. Presently active instruments are found, e.g., at the Finnish Geodetic Institute, Metsähovi mine (Ruotsalainen, 2008) and in California (Piñon Flat Observatory, University of California, San Diego). Vertical pendulums are mounted in boreholes for measuring tilt and the displacement of the pendulum is detected by a capacitive transducer with resolution of 1 nrad; examples

are the tiltmeters of the Geodynamic Observatory Moxa, Thuringia, Germany. Application of a tilt array of this type to monitor pore pressure variations and minor earthquakes are given by Jahr et al. (2008). Portable short-base tiltmeters are used for volcanic monitoring, where the expected tilts are larger than for crustal deformation studies and therefore require less resolution and stability in time. The electronic tiltmeters record the movement of an air bubble in a conducting fluid. These instruments are used for volcanic monitoring of the USGS.



Geodetic Pendulums, Horizontal Ultra Broad Band, Figure 2 (Continued)



Geodetic Pendulums, Horizontal Ultra Broad Band, Figure 2 The graph shows the continuous time series of tilt recorded by the horizontal geodetic pendulum of Grotta Gigante. (**a**) *Black continuous line* are the NS and EW records with daily sampling. The regular oscillation is the annual tilting. The multi-decadal variation is the sum of an oscillation of 33 years (*thick line*) and a linear variation (*dashed line*). The spikes are due to tilting caused by floods of the underground carstic water flow. Asterisks mark data interruptions. (**b**) Spectral amplitudes in the frequency range of the free oscillations following the earthquakes of Chile 1960 (CHI-NS, CHI-EW) and Sumatra-Andaman 2004 (SU-NS, SU-EW). (**c**) The tilting of the cave due to the underground flooding: The *thick line* shows tilting of the cave during the flood and the z-axis is the time. The *thinner line* shows the tilting during the flooding in the EW and NS direction. Furthermore, the water level in a nearby cave during a flood is shown.

Geodetic Pendulums, Horizontal Ultra Broad Band, Table 1 Identified causes that deform the Grotta Gigante cave with characteristic periods and amplitudes

Cause of deformation	Characteristic period	Tilt maximum observed amplitude	Notes
Plate tectonics secular movement	Secular movement, near to static	31.4 nrad/year	
Plate tectonics multi-annual oscillation*	33 years	316 nrad	
Annual variation*	365.6 days	300 nrad	Orientation toward N60°E in mid March and toward N240°E in mid September
Earth tides*	Daily, half daily	39 nrad	
Ocean loading*	Daily, half daily	60 nrad	
Free oscillations* of the earth	54 min to a few minutes	Up to 0.7 nrad	
Underground water flow	A few days	Up to 300 nrad	

*Oscillation amplitude defined as half of the peak-to-peak excursion

Summary

The ultra broad band horizontal geodetic pendulum measures the inclination of a reference axis with respect to the vertical. The prototype installed in the Giant Cave in North-Eastern Italy has shown that the deformations that affect the upper crust are due to tectonic effects (seismic waves, free oscillations of the earth, coseismic deformation, tectonic plate deformation) and due to environmental effects (thermoelastic deformation, hydrologic water flow, ocean loading). The existing record of the greatest seismic event ever measured, the Chile 1960 earthquake, allows to make an absolute magnitude comparison with recent events due to the fact that amplitude factor, the instrumentational setup, and the location are perfectly known and controllable today.

Bibliography

Bolt, B. A., and Marussi, A., 1962. Eigenvibrations of the Earth observed at Trieste. *Geophysical Journal of the Royal Astronomical Society*, **6**, 299–311.

Braitenberg, C., Romeo, G., Taccetti, Q., and Nagy, I., 2006. The very-broad-band long-base tiltmeters of Grotta Gigante (Trieste, Italy): secular term tilting and the great Sumatra-Andaman islands earthquake of December 26, 2004. *Journal of Geodynamics*, **41**, 164–174.

Braitenberg, C., and Zadro, M., 2007. Comparative analysis of the free oscillations generated by the Sumatra-Andaman islands 2004 and the Chile 1960 earthquakes. *Bulletin Seismological Society of America*, **97**, S6–S17, doi:10.1785/0120050624.
Ferreira, A. M. G., d'Oreye, N. F., Woodhouse, J. H., and Zürn, W., 2006. Comparison of fluid tiltmeter data with long-period seismograms: surface waves and Earth's free oscillations. *Journal of Geophysical Research*, **111**(B11), B11307/17.
Jahr, T., Jentzsch, G., Gebauer, A., and Lau, T., 2008. Deformation, seismicity, and fluids: Results of the 2004/2005 water injection experiment at the KTB/Germany. *Journal of Geophysical Research*, **113**, B11410, doi:10.1029/2008JB005610.
Marussi, A., 1959. The University of Trieste station for the study of the tides of the vertical in the Grotta Gigante. In *Proceedings of the III International Symposium on Earth Tides,* Trieste, 1960. pp. 45–52.
NEIC, 2010. National Earthquake Information Center, http://earthquake.usgs.gov/
Ruotsalainen, H. E., 2008. Recording deformations of the Earth by using an interferometric water level tilt meter. In Korja, T., Arhe, K., Kaikkonen, P., Korja, A., Lahtinen, R., and Lunkka, J. P. (eds.), Fifth symposium on the structure, composition, and evolution of the Lithosphere in Finland, Institute of Seismology, University of Helsinki, Report S-53, 103–106.
Skocjanske Jame, 2010. http://www.park-skocjanske-jame.si/
Zadro, M., and Braitenberg, C., 1999. Measurements and interpretations of tilt-strain gauges in seismically active areas. *Earth Science Reviews*, **47**, 151–187.
Zöllner, K. F., 1871. Über einen neuen Apparat zur Messung anziehender und abstoßender Kräfte, Kgl. sächs. Gesellsch. der Wissensch. zu Leipzig, 27. Nov. 1869 und 1. Juli 1871.
Zöllner, K. F., 1872. Zur Geschichte des Horizontalpendels, Kgl. sächs. Gesellsch. der Wissensch. zu Leipzig, math.-phys. Klasse, November 1872.

Links to geodynamic observatories with active tiltmeters

Finnish Geodetic Institute: http://www.fgi.fi/tutkimus/tiedot_aiheesta_eng.php?projekti=19
Geodynamic Observatory Moxa: http://www.geo.uni-jena.de/Homepage-Moxa-englisch/start.html
Plate Boundary Project, Earthscope, http://www.earthscope.org/observatories/pbo
University California San Diego: http://pfostrain.ucsd.edu
USGS volcano monitoring with tiltmeters: http://volcanoes.usgs.gov/activity/methods/deformation/tilt/index.php

Cross-references

Earth Tides
Free Oscillations of the Earth
Great Earthquakes
International Geophysical Year
Seismic Instrumentation

GEODYNAMICS

Alessandro M. Forte
GEOTOP Département des Sciences de la Terre et de l'Atmosphère, Université du Québec à Montréal, Montréal, QC, Canada

Definition

Geodynamics (noun, functioning in the singular). A broad, multidisciplinary branch of geophysics that investigates the causes and consequences of the forces acting on and inside the solid Earth and the rock movements and deformations they produce.
Geodynamic, geodynamical (adjective). Relating or pertaining to the field of geodynamics.

The term geodynamics originates from two root words of Greek origin: *geo-*, relating to the Earth, and *dynamics*, the branch of physics that involves the study of forces acting on bodies and the changes in motion they produce. As the name therefore suggests, the focus of geodynamics is on the forces that act on the solid Earth and the motions and deformation they produce.

Historically, in the late nineteenth and early twentieth century, when geodynamics was emerging as a distinct field of study, it also included the earliest systematic investigations of earthquakes and propagating seismic waves. These studies later evolved into a separate branch of geophysics called seismology. Modern geodynamics investigates forces and movements over timescales that exceed the characteristic periods of propagating earthquake waves and corresponding Earth oscillations (i.e., approximately ranging from a second to an hour). On these longer timescales, for geodynamic processes ranging from Earth tides (hours to years) to postglacial rebound (thousands of years) and ultimately plate tectonics (millions of years), the deformation is sufficiently slow that the acceleration of the material is entirely negligible (i.e., the deformation is quasistatic).

The forces studied in geodynamics mainly originate from lateral changes in density $\Delta\rho$ on or within the Earth, thereby producing buoyancy or Archimedes forces $\Delta\rho\, g$ per unit volume, where g is Earth's gravitational field. For example, lateral density variations due to thermal plumes in the convecting mantle give rise to buoyancy forces $\Delta\rho\, g = \alpha\rho\Delta T\, g$, where α is the coefficient of thermal expansion and ΔT are the lateral temperature anomalies between the plumes and ambient mantle. These thermal buoyancy forces drive the convective circulation in the mantle and the corresponding surface motions of the tectonic plates. Similarly, the topographic depressions Δh at the Earth's surface left behind by the rapid melting of the massive ice sheets that once covered Canada and Scandinavia about 20,000 years ago, give rise to buoyancy forces $\Delta h\rho\, g$ per unit area that are focused at the surface (here ρ is the average density of the crust). Postglacial rebound is the slow viscous creep of the mantle and the resulting uplift of the crust produced by these surface forces.

There are also external forces $\rho\,\nabla\phi$ where, for example, $\nabla\phi$ is the gradient of the gravitational potential generated by the attraction of the Sun and Moon, or it may represent the centrifugal force due to Earth's diurnal rotation. For the former case, these forces produce the tidal deformations in the Earth or, for the latter case, they produce the oblate ellipsoidal shape of the Earth. These tidal and rotational forces change over an enormous range of timescales, from a few hours to hundreds of millions of years, and they produce globally distributed deformations throughout the Earth's interior and at the surface.

A major area of geodynamics research is the measurement and the quantitative modeling of forces and deformations in the crust, underlying lithosphere and deep mantle. An understanding of the relationship between force and deformation requires detailed knowledge of the rheology of the solid Earth that depends on the ambient temperature, pressure, and the timescales over which the forces are applied. For example, on timescales corresponding to the tides, the solid Earth behaves like an imperfectly elastic body. However, on the much longer timescales of mantle convection and plate tectonics, the solid mantle effectively behaves like a fluid medium with an extremely high viscosity.

The effective viscosity η of the mantle is one of the most important parameters in geodynamics because it controls the speed at which the mantle may flow and deform in response to buoyancy forces. The characteristic flow speed is $v = \Delta\rho g L D/\eta$ where L represents the lateral dimension (distance) of the density anomalies and D is the depth (typically that of the mantle) over which they exist. Flow velocities in the mantle are of the order of several centimeters per year, similar to those of the moving surface plates. This flow velocity determines the efficiency of convective heat transport across the mantle and hence governs the thermal (and chemical) evolution of the Earth. For this reason, a great deal of work in the second half of the twentieth century, and up to the present day, is dedicated to constraining the value and spatial variation of mantle viscosity from the analysis of various geodynamic data sets (e.g., from postglacial uplift rates, geoid anomalies, long-term sea level variations, plate motions).

The range of geophysical phenomena studied in geodynamics has increased greatly over the course of the twentieth century. For example, an early treatise on geodynamics by Love (1911) provided a detailed mathematical description of a range of problems that included the isostatic support of continents, the height of mountains chains and the forces required to maintain them, the Earth tides, the propagation of earthquake waves and the oscillations of the entire Earth produced by earthquakes. With the advent of plate tectonics and the acceptance that some form of thermal convection in the mantle is necessary to drive the plate motions, the scope of geodynamical investigations has been greatly enlarged.

Modern geodynamics research deals with a multitude of problems relating to plate tectonics and heat transfer in the Earth, as well as numerous studies of the rheology of the mantle and the crust-lithosphere system that bear directly on the efficiency of heat transport across these different structures. The very wide range of problems now included in contemporary geodynamics is well illustrated by Turcotte and Schubert (2002) and Schubert et al. (2004). Depending on the specific problem, it is difficult, if not impossible, to devise analog laboratory experiments that can properly incorporate the complex spatial variations of the physical properties of the solid Earth. This difficulty is further compounded when it is necessary to model geodynamic processes over extremely large spatial and temporal scales, as in the case of mantle convection and plate tectonics. To address these major challenges, the rapidly growing field of *computational geodynamics* has emerged, in which very complex problems in Earth dynamics are now routinely solved by numerical simulations on high-performance computers (e.g., Ismail-Zadeh and Tackley, 2010).

Bibliography

Ismail-Zadeh, A., and Tackley, P., 2010. *Computational Methods for Geodynamics*. New York: Cambridge University Press.
Love, A. E. H., 1911. *Some Problems of Geodynamics*. New York: Cambridge University Press.
Schubert, G., Turcotte, D. L., and Olson, P., 2004. *Mantle Convection in the Earth and Planets*. New York: Cambridge University Press.
Turcotte, D. L., and Schubert, G., 2002. *Geodynamics*, 2nd edn. Cambridge: Cambridge University Press.

Cross-references

Continental Drift
Earth Tides
Earth's Structure, Global
Earthquakes and Crustal Deformation
Geoid
Gravity Field, Temporal Variations from Space Techniques
Isostasy
Lithosphere, Continental
Lithosphere, Continental: Thermal Structure
Lithosphere, Mechanical Properties
Lithosphere, Oceanic
Lithosphere, Oceanic: Thermal Structure
Mantle Convection
Mantle Viscosity
Plate Driving Forces

GEOELECTROMAGNETISM

Antal Ádám, László Szarka
Geodetic and Geophysical Research, Institute of the Hungarian Academy of Sciences, Sopron, Hungary

Definitions and scope

Geoelectromagnetism, in the broad sense, includes the application of classical electrodynamics to various interrelated regions of space (Sun, interplanetary field, magnetosphere, ionosphere, atmosphere), on and below the Earth surface (near-surface regions, crust, upper mantle [or lithosphere and asthenosphere using alternative nomenclature], lower mantle, and Earth's core). Over the Earth surface the dominant phenomena of geoelectromagnetism are electromagnetic waves, in a wide spectrum. The electromagnetic waves do not penetrate deep into the Earth. Instead, the propagation of the electromagnetic field inside the Earth is a diffusion process.

In a narrower sense, geoelectromagnetism means the application of time variations of natural electromagnetic fields to the study of the electric resistivity distribution

in the Earth's interior, which gives indispensable information about the structure, composition, and processes of the subsurface.

The term electromagnetic geophysics is also used, but electromagnetic geophysics in general includes all exploration techniques, which may employ not only natural sources, but various man-made sources as well. In order to investigate very deep structures, extremely large-size sources of natural electromagnetic field variations should be used. In the following, we restrict ourselves to those geoelectromagnetic methods that apply natural electromagnetic field variations as the source to study deeper (lithosphere, asthenosphere, and upper mantle) geological structures.

Introduction

Of three electromagnetic parameters (magnetic permeability, electrical resistivity, and permittivity) associated with the subsurface geology, only the magnetic permeability and the electrical resistivity are relevant because of the diffusive nature of the electromagnetic field propagation. The magnetic permeability is usually assumed to be constant, which – apart from some exceptional cases – has been seen to be a good assumption. The electrical conductivity σ of various earth materials varies over an extremely wide range; therefore, its spatial distribution reflects most structural/physical/petrological changes in the Earth. In practice, instead of the conductivity, the electrical resistivity ρ, where $\rho = 1/\sigma$, is often used. The ohm-meter (or Ωm) is the unit of measurement.

These methods are classified on the basis of the measured geoelectromagnetic field components. When only the electric field variations are measured, we speak about the *telluric* or *Earth current* technique, or *tellurics* (TT) for short. If only the geomagnetic field components are measured (usually as a function of the period of the time variations of the electromagnetic field), the method is known as "*Geomagnetic Deep Sounding*" (*GDS for short*), or by its other name: "*MagnetoVariation Sounding*" (MVS). Electrical conductivity anomalies (i.e., the zones of concentration of electric current) can be located by using the so-called "*Magneto-Variation Profiling*" (MVP), where the so-called induction vectors point to the current concentrations. A more sophisticated technique is the "*Magnetotelluric Sounding*" (MTS) or shortly "*magnetotellurics*" (MT), where both the electric and magnetic field variations are measured in a broad period range (see Skin Depth).

Although the Geomagnetic Deep Sounding (the MagnetoVariation Sounding) is more or less coeval with observatory records (see, e.g., Lamb, 1883), the history of almost all these variants dates back to about six decades. The MVP and its induction vectors were introduced in the second half of 1950s in studying the anomalous behavior of geomagnetic variations in North Germany (e.g., Schmucker, 1959; Wiese, 1965). The theoretical basis of the magnetotelluric (MT) method was laid by Tikhonov (1950) and Cagniard (1953). The first MT field measurements were carried out in the early 1960s. Application of the tellurics (i.e., the simplest technique) to investigate large sedimentary basins started in the middle of the twentieth century in Hungary (Kántás, 1954) and elsewhere.

In the following, the principles of these geoelectromagnetic methods will be presented. We summarize the governing equations of MVS, MVP, and MT techniques and illustrate their field application. The telluric measurements being the special case of the magnetotelluric method used in the so-called S-interval (see Figure 4) where the phase approximates zero, is not discussed here in detail.

Magnetovariational sounding (MVS) and magnetovariational profiling (MVP)

By simultaneously measuring the time variations of the horizontal and vertical geomagnetic field components at a site, it is possible to determine the subsurface electrical resistivity. Carrying out the measurements in a broad period interval, the depth-dependence of the electrical resistivity can be estimated (see Skin Depth).

For a spherical Earth, this technique is able to provide information about the Earth's conductivity down to the depths of 1,200–1,800 km in the mantle (Bailey, 1973; Rokityansky, 1982). For shallower studies (i.e., down to depths of less than approximately 100 km) the Earth's surface may be considered as plane.

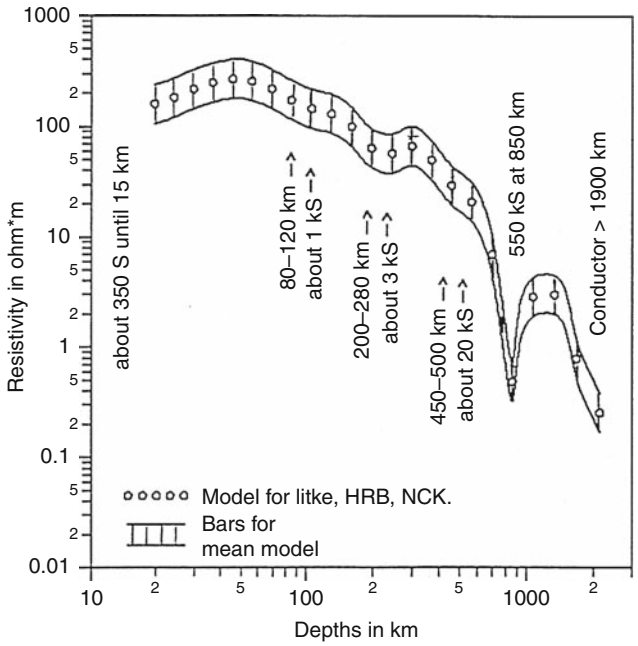
The magnetovariational profiling (MVP) aims at locating large lateral inhomogeneities of the Earth's crust and upper mantle. In the recent book by Berdichevsky and Dmitriev (2008), this theme is discussed in detail under the title "*Magnetovariation Studies*."

Deep mantle MVS studies

We follow Rokityansky (1982) nomenclature concerning the deep mantle studies. The field source of so-called Dst geomagnetic variations (used recently by Yu Semenov et al., 1997; Yu Semenov et al., 2008) is the ring current at the equator. When the vertical magnetic field of the ring current source is approximated by the first term in its spherical harmonic expansion, that is, $H_z \sim P_1^0(\cos\theta)$ where P_1^0 is the associated Legendre function in the first term of the expansion, the ratio of vertical (H_z) to horizontal $\left(\sqrt{H_x^2 + H_y^2}\right)$ field component can be converted into the complex apparent resistivity (Rokityansky, 1982). (In standard practice, x and y are the horizontal components, with the x-axis pointing northward, and the positive z-axis directed vertically downwards into the ground.) If the magnetic permeability (μ), the radius of the Earth (R), the geomagnetic colatitude of an observatory (Θ), and the angular frequency (ω) in the assumed time-dependence $e^{i\omega t}$ are known, then the apparent resistivity ρ is given as

$$\rho(\omega) = \frac{i\omega\mu}{4} R^2 \tan^2\Theta \left| \frac{H_z(\omega)}{\sqrt{H_x^2 + H_y^2}} \right|^2 .$$

Yu Semenov et al. (1997) used the daily mean values of geomagnetic observatories. Their first step in the analysis was a rotation of the observatory reference systems to determine the maximum Dst field (Schultz and Larsen, 1983). For the agreement between the horizontal Dst field and the first spherical harmonic model of the source, a coherence test was carried out. The period ranges, which were in contradiction with the source model, were excluded from further analysis. The transfer functions between the vertical (H_z) and horizontal $\sqrt{H_x^2 + H_y^2}$ components were computed only if the coherence values were greater than 0.7 between them. The transfer functions were converted into apparent resistivity. Yu Semenov et al. (1997) combined the magnetovariational data with the results of magnetotelluric sounding in observatories and estimated, by using a robust technique, the resistivity model below the observatories. The mean geoelectric structure of the mantle below the Pannonian Basin is shown in Figure 1 on the basis of data from Nagycenk, Litke (Hungary), and Hurbanovo (Slovakia). A more detailed combined MV + MT study in East-Central Europe pointed out a significant difference in the Precambrian and Phanerozoic mantle structures, separated by the Tornquist-Teisseyre Zone (TESZ line; Yu Semenov et al., 2008).



Geoelectromagnetism, Figure 1 The mean geoelectrical structure of the mantle in the Pannonian Basin using the magnetic records of Nagycenk, Litke (Hungary), and Hurbanovo (Slovakia) for magnetovariational soundings (Yu Semenov et al., 1997).

Schmucker (2003) proposal for the generalized horizontal spatial gradient method (gSHG) is promising for a reliable combination with the MT method as stated recently by Vozar and Semenov (2010) modeling. Kelbert et al. (2009) inverted long period geomagnetic response functions to derive a global-scale 3D model of electrical conductivity variations in the Earth's mantle. Their results – one order magnitude conductivity increase in the subduction zones – support the presence of at least some water in these transition zones.

Magnetovariational profiling

It is evident from the Biot–Savart law that on the surface, just over and near the center of a horizontal lineament of high electrical conductivity in the subsurface, in which the telluric currents are gathered, the measured vertical magnetic component H_z is zero. More precisely, H_z is changing sign along a profile, crossing the lineament. At the same time, the horizontal component $\sqrt{H_x^2 + H_y^2}$ has a maximum over the lineament, with a moderate decrease on both sides. Therefore, the $H_z \Big/ \sqrt{H_x^2 + H_y^2}$ ratio can be drawn as a vector, that is, a quantity having magnitude and direction. It is called "*induction vector*" or "*induction arrow*." Along a measuring profile crossing the lineament, the induction arrow will change sign. Induction arrow profiles at various periods provide useful depth information. The origin of the method goes back to the resolution of a paradox, namely when the vertical component of substorms in two German geomagnetic observatories (Wingst and Niemegk) proved to be of opposite sign. This paradox was attributed later to the North German anomaly (Schmucker, 1959).

The source of the method is the same as that of all other geoelectromagnetic methods and a similar time series analysis is needed to separate the variations of different periods (frequencies) for single station and array measurements too. Array and single site transfer functions are applied.

Array transfer function

The normal field is that which would exist in the presence of one-dimensional layered structure in which the inhomogeneity is embedded. The anomalous field is then most simply defined as the total measured field minus the normal field.

Due to the linearity of the Maxwell field equations, there exists a linear relationship between the anomalous and normal magnetic field components (Schmucker, 1970).

$$H_{xa} = T_{xx}H_{xn} + T_{xy}H_{yn} + T_{xz}H_{zn} + \varepsilon_x$$
$$H_{ya} = T_{yx}H_{xn} + T_{yy}H_{yn} + T_{yz}H_{zn} + \varepsilon_y$$
$$H_{za} = T_{zx}H_{xn} + T_{zy}H_{yn} + T_{zz}H_{zn} + \varepsilon_z$$

where H_x, H_y, and H_z are the field components and ε is the error of the relation. All these T quantities are period dependent and complex. Subscript "a" and "n" refer to anomalous and normal parts, respectively. The assumption that the source field is uniform, allows the normal part of the H_z component to be taken as zero. Accordingly, the third term on the right-hand side of equations can be neglected.

Single site transfer function
With the following condition

$$H_{zn} \ll H_{za},\ H_{xa} \ll H_{xn},\ H_{ya} \ll H_{yn}$$

the relation between the vertical and horizontal components is

$$H_{za} = T_{zx}H_x + T_{zy}H_y,$$

where T_{zx} and T_{zy} are complex values.

Induction vectors (arrows) are computed separately from the real and imaginary parts of the T values. In practice two variants are used (Wiese, 1965; Parkinson, 1959). The vector lengths in both cases are the same:

$$L_r = \sqrt{(\mathrm{Re}T_{zx})^2 + (\mathrm{Re}T_{zy})^2},$$

$$L_i = \sqrt{(\mathrm{Im}T_{zx})^2 + (\mathrm{Im}T_{zy})^2}.$$

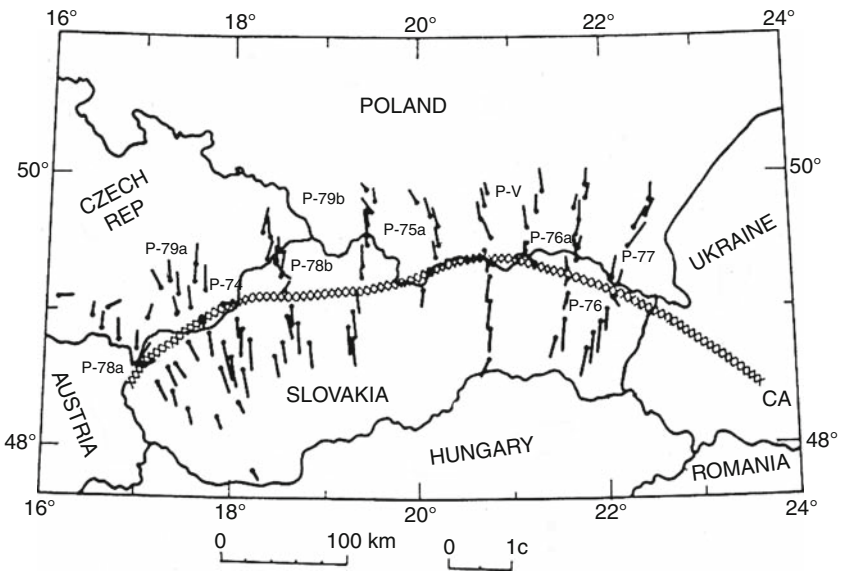
The azimuth of the Wiese vector (arrow)

$$\phi_w = \tan^{-1}\left(\frac{T_{zy}}{T_{zx}}\right)$$

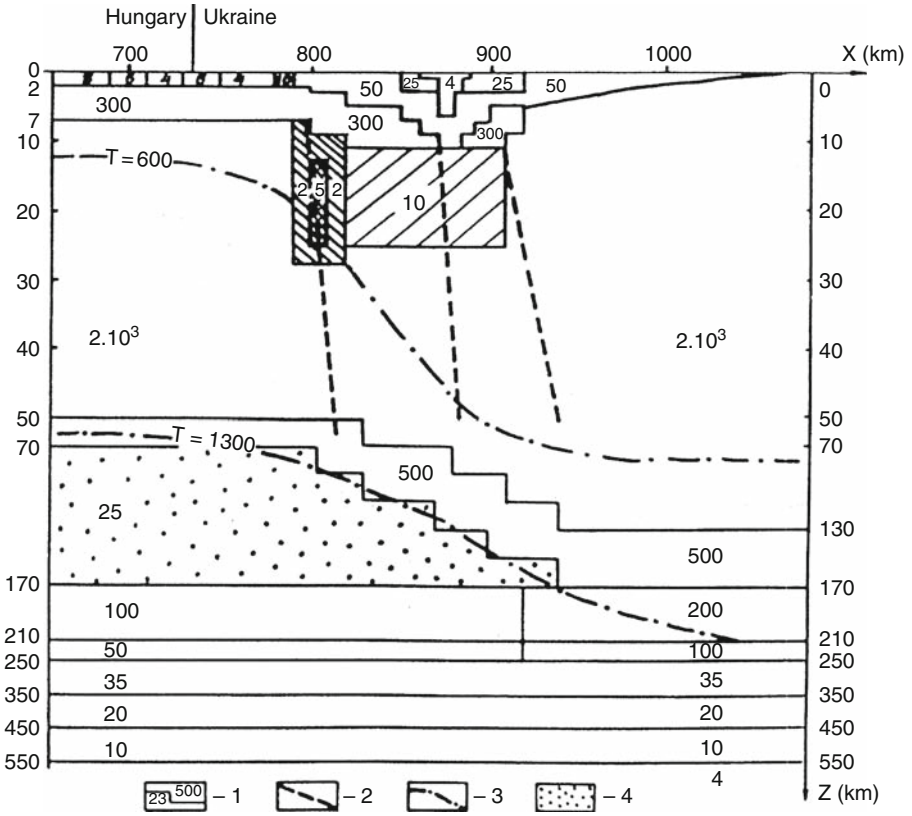
gives the direction of the maximum correlation between the upward *vertical* and horizontal field component. This definition of the azimuth implies that the x-axis is directed northwards. The magnitude (length) represents the ratio of vertical field amplitude to the amplitude of the horizontal component with which it correlates (Arora, 1997). The Wiese vectors point away from areas of higher conductivity. Conversely, the azimuth of the Parkinson arrow points towards the high electrical conductivity region. The induction arrows when displayed on the map clearly identify the location and trend of subsurface conductivity anomalies (Arora, 1997).

The MVP (induction vector) has an extended literature. Rokityansky (1982) provided a series of model studies. Arora (1997) published several alternative approximations, which might better determine the structure of the conductive bodies, for example, "perturbation arrows," "hypothetical event analysis," *Z/H* "pseudosection," and 2D modeling.

Among the field studies conducted in East-Central Europe in which induction vectors were plotted, it is the Carpathian conductivity anomaly (indicated at first by Wiese, 1965) that has received the greatest attention, attracting scientists from all neighboring countries. The distribution of the induction vectors is shown in Figure 2 (after Jankowski et al., 2005) and one of the best models is shown in Figure 3 (Zhdanov et al., 1986). For the assumed conductive region between 10 and 20 km there are two alternative hypotheses: (a) porous rocks saturated by mineralized water (ionic conduction), (b) conducting rock complex connected to graphite.



Geoelectromagnetism, Figure 2 Wiese induction vectors measured along the Carpathian conductivity anomaly (Modified after Jankowski et al., 2005).



Geoelectromagnetism, Figure 3 Geoelectric and geothermal model of the Carpathian conductivity anomaly and the deepening asthenosphere toward the East European Platform along a NE–SW profile (so-called KAPG geotraverse II) crossing the Transcarpathian region in Ukraina (After Zhdanov et al., 1986). Legend: 1, geoelectric boundaries and resistivity values of blocks in ohm.m; 2, deep faults; 3, isotherms; 4, conducting asthenosphere.

At present the induction vectors form an integral part of routine MT interpretation software (e.g., WinGlink) and the vertical magnetic components are also incorporated in MT inversion.

Magnetotelluric soundings

Basic formulas

Magnetotellurics (MT) means a joint measurement of the natural electric and magnetic field variations on the surface. Its energy source is partly the same as that of MVS and MVP. Namely, at frequencies $f < 5$ Hz the electromagnetic (EM) source field is generated by ionospheric as well as magnetospheric phenomena (micropulsations, substorms, etc.) and at frequencies $f > 5$ Hz the EM field is due to worldwide lightning. The source field penetrates into the Earth, where it is reflected from all resistivity interfaces. The measured field is the sum of the primary (source) and secondary (subsurface-dependent) fields. The signal propagation in the subsurface is diffusive. *Skin depth* δ is the depth at which a uniform field is attenuated to $1/e$ of its surface amplitude in a conductor of uniform conductivity:

$$\delta = \sqrt{2/\mu\sigma\omega},$$

where σ is the conductivity (Sm^{-1}), μ is the magnetic permeability (Hm^{-1}), and $\omega \equiv 2\pi f$ is the angular frequency with f denoting the frequency (Hz). The free-space value of magnetic permeability ($\mu = \mu_0 = 4\pi \times 10^{-7}\mathrm{Hm}^{-1}$) is generally assumed. As shown in the above equation, the lower the frequency is, the larger is the magnetotelluric depth of investigation.

The time variations of the two horizontal (E_x and E_y) components of the electric (or telluric) field are measured by using non-polarizable electrodes in a broad period range. The H_x and H_y horizontal geomagnetic field components are measured either with induction coils or fluxgates. From the time series of the horizontal electric and magnetic field components their frequency spectra are determined. In the simplest case, from $E_x(\omega)$ and $H_y(\omega)$ a complex impedance $Z(\omega)$ is computed as follows:

$$Z_{xy}(\omega) = \frac{E_x(\omega)}{H_y(\omega)}.$$

From Z the apparent resistivity ρ (i.e., the electrical resistivity of a fictitious homogeneous half-space) is defined as

$$\rho = \frac{1}{\omega\mu}\left|Z_{xy}\right|^2.$$

The phase difference between the electric and geomagnetic components is given as

$$\varphi = \arg Z_{xy} = \tan^{-1}\frac{\mathrm{Im}Z_{xy}}{\mathrm{Re}Z_{xy}}.$$

In the general case, the impedance is a period-dependent tensor:

$$\begin{vmatrix} E_x \\ E_y \end{vmatrix} = \begin{vmatrix} Z_{xx} & Z_{xy} \\ Z_{yx} & Z_{yy} \end{vmatrix} \begin{vmatrix} H_x \\ H_y \end{vmatrix}$$

where Z_{xy} and Z_{yx} are the antidiagonal or main impedance elements and Z_{xx} and Z_{yy} are the diagonal or additional elements.

For a 1D model (where the geoelectric structure varies exclusively with depth), the impedance elements are, theoretically, in the following relationships:

$$Z_{xy} = -Z_{yx} \text{ and } Z_{xx} = Z_{yy} = 0.$$

For a 2D model (where the resistivity section does not change in one of the horizontal direction named by strike direction), the apparent resistivity calculated from E_x and H_y components will differ from that derived from E_y and H_x, that is,

$$Z_{xy} \neq Z_{yx}$$

Furthermore, $Z_{xx} = Z_{yy} = 0$ is valid exclusively when the horizontal axes are aligned along and perpendicular to the strike direction.

One of the basic modes is called as *TE mode (or E-polarization),* where the electric field exists only in *strike direction. The other basic mode is called* as *TM mode (or H- or B-polarization),* where it is the magnetic field, which has one single component parallel to the strike direction. In the TE mode, there is also a measurable vertical magnetic component (H_z), which is usually also measured; in the TM mode, there is a measurable vertical electric component (which is not measured). The TE and TM mode apparent resistivities represent the geoelectric structure in two different ways. The TE mode is more sensitive to deep conductive anomalies, while the TM mode – due to the charge accumulation at the resistivity boundaries – is more sensitive to the inhomogeneity boundaries.

When the model is 3D, the complete impedance tensor has four nonzero complex elements.

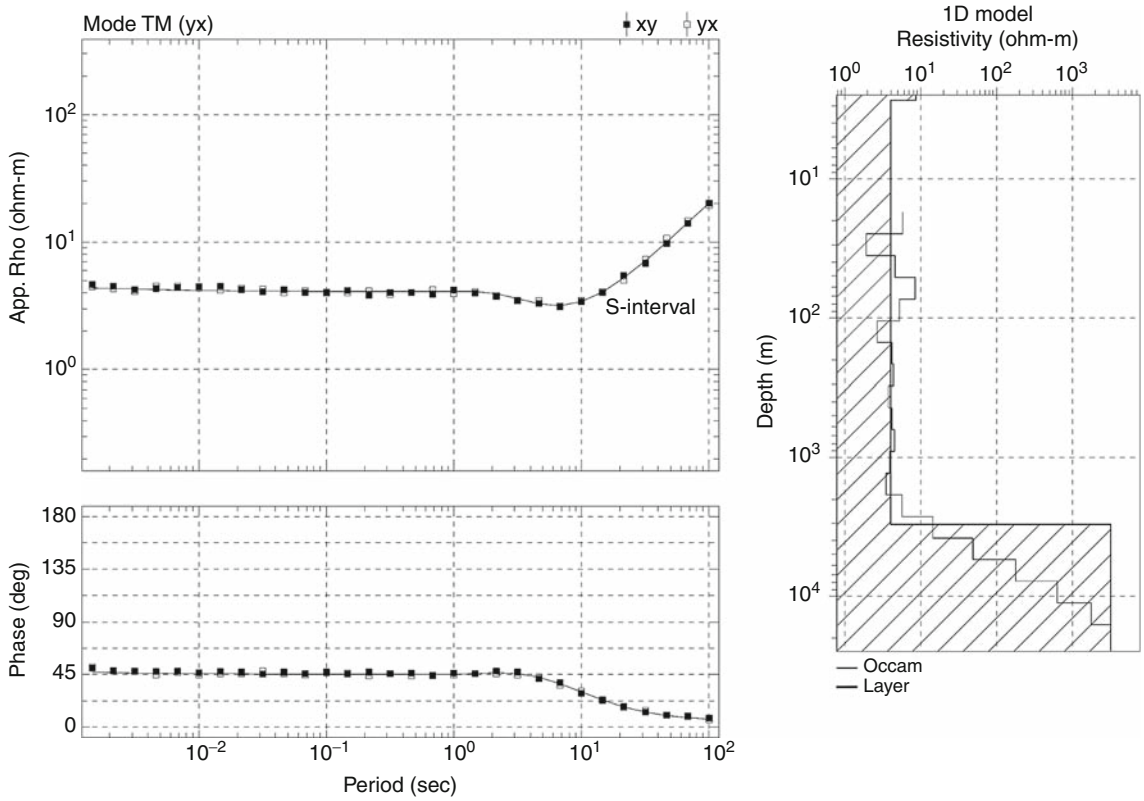
MT sounding curves are obtained by plotting apparent resistivity and impedance phase as functions of period (or square root of the period) over a wide period range. In Figure 4, a 2 layer 1D resistivity model is shown. Note that the resistivity curve is fully logarithmic while the phase curve is semi-logarithmic. $Z_{xy}(\omega)$ and $Z_{yx}(\omega)$ serve as the basis of the TE and TM mode $\rho(\omega)$ and $\varphi(\omega)$ sounding curves. Computing the impedance elements in rotated coordinate systems, one obtains so-called polar diagrams, which characterize the anisotropy of the resistivity distribution.

Interpretation

Complex impedance elements recorded over a range of periods and at many sites along a profile or over an area, can be used to estimate the true resistivity distribution of the subsurface. This procedure is called inversion. Before carrying out inversion, the impedance elements should be corrected from distortions of both technical and structural origin. The "remote reference" method – using the data of simultaneous measuring sites in the processing – reduces, for example, the effect of man-made electromagnetic noises (Gamble et al., 1979). Different *decomposition* techniques (e.g., (Larsen, 1977; Groom and Bailey, 1989; Bahr, 1991); corrected by Prácser and Szarka, 1999; Caldwell et al., 2004) enable suppression of the effect of a near-surface small 3D inhomogeneity causing the "*static shift*" of galvanic origin. At the same time they inform how well the impedance data correspond to a 2D resistivity model, and give an estimate for the strike direction, too. The papers by Bibby et al. (2005) and by Weaver et al. (2006) also explore the role of the phase tensor in determining dimensionality and strike direction. Due to the fact that the inversion is inherently nonunique, additional geophysical and geological information is needed in order to get a realistic model of the subsurface. The most frequently used 2D inversion algorithms in magnetotellurics are those by Rodi and Mackie (2001) and Siripunvaraporn and Egbert (2000) (see Figure 7). The fitness of the modeled and measured data is characterized by the root-mean-square (rms) values. Various EM imaging procedures, exploiting the whole information content of the impedance tensor, may also be useful, as proposed by Szarka et al. (2005) and Weaver et al. (2000). For this purpose, Szarka and Menvielle (1997) recommended the system of independent rotational invariants of the impedance tensor. The usefulness of this approach was demonstrated in detail by Novák (2009) using field data. In recent years, many papers have been published on 3D modeling and inversion and some have been presented at the biennial International Association of Geomagnetism and Aeronomy (IAGA) Workshops on Electromagnetic Induction in the Earth, including reviews of these topics, e.g., Avdeev (2005).

Application to geosciences and future directions

MVS, MVP, MT, and to a lesser extent tellurics have received a very wide application in studying the structure and physical processes of the Earth at various local, regional, and global scales. The development of field applications has been disseminated at the biennial IAGA electromagnetic induction workshops since 1972 (see MTNet website, http://www.dias.ie/mtnet/working group/wg_wsreviews.html) and at assemblies of IAGA. For example, at the last (11th) IAGA Assembly (Sopron,



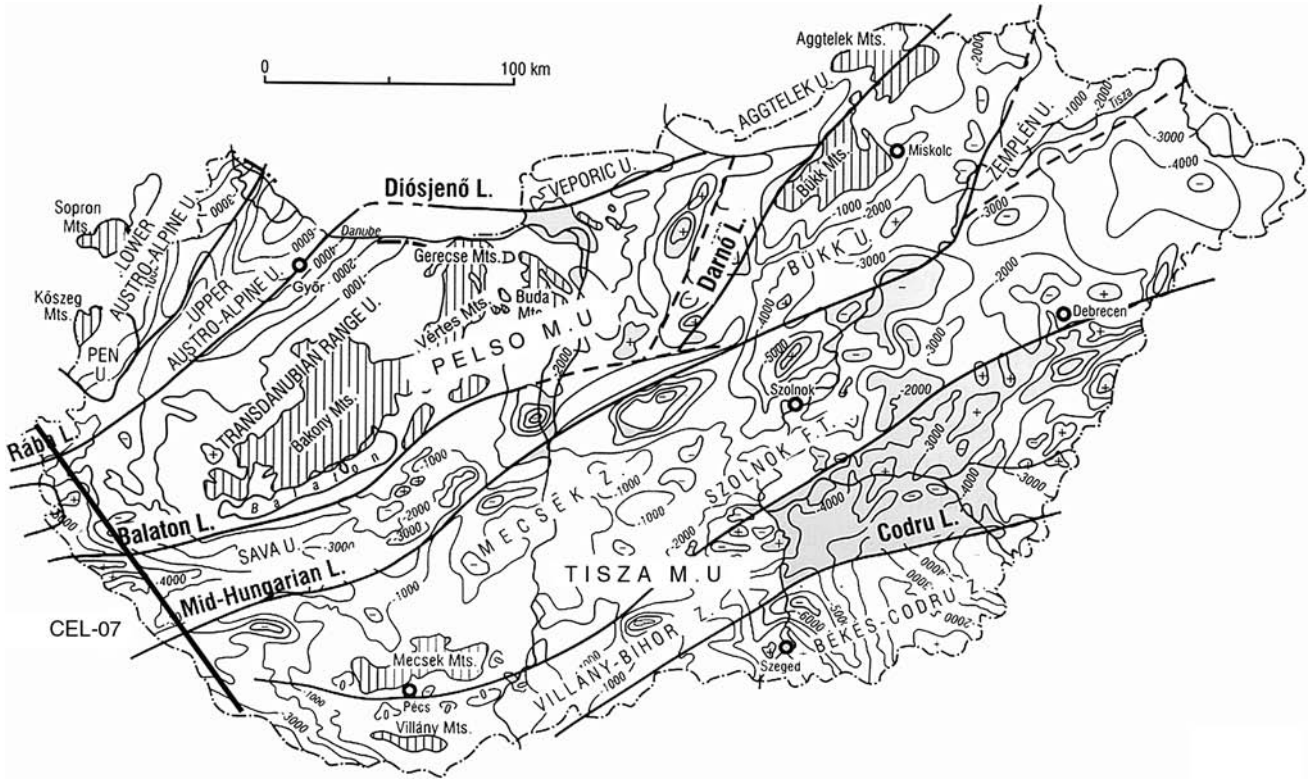
Geoelectromagnetism, Figure 4 Magnetotelluric sounding curves for a layer (1D) model with Rho1 = 4 Ω m, h_1 (thickness) = 3 km, Rho2 = 2,000 Ω m. According to the Occam and few-layer inversions the "S-interval" starts at a period of 10 s. At longer periods the telluric (Earth currents) method can be used to determine the thickness of the low resistivity sedimentary layer on a more resistive basement.

23–30 August 2009, www.iaga2009sopron.hu), a wide spectrum of geoelectromagnetic studies, from the surface to the middle of mantle, was presented: theoretical and practical aspects of near-surface and environmental problems, oceanic geophysical exploration, formation and deformation of the continental lithosphere, crustal tectonic processes, deep mantle structure, and data interpretation methods.

In order to illustrate the wealth of geoelectromagnetic applications, the authors present a selection from their own results from the Pannonian Basin, which is a very complex geological-geophysical unit of the Carpatho-Pannonian region. Such a summary cannot be done without mentioning other geoelectromagnetic investigations of the most well-known global geodynamic phenomena, such as the active plate tectonics in the Himalayas, Andes, San Andreas Fault, Afar region, etc. Earth being a "dynamic planet," great efforts have also been devoted to mitigate the geohazards: earthquakes, volcanoes (as accompanying phenomena of the plate tectonics), and landslides (Bedrosian, 2007; Korja, 2007).

Before discussing the electrical conductivity anomalies/structures in the Pannonian Basin, we give a short description about its geology from those aspects that constrain the spatial distribution of electrical resistivity in the subsurface. The Pannonian Basin has a fairly thick and relatively uniform low electrical resistivity sedimentary cover of Late Tertiary origin, which was investigated in detail by telluric/magnetotelluric soundings. According to Haas (2001), the Pre-Neogene basement "shows a mosaic pattern of heterogeneous structural elements, a collage of allochthonous terrains derived from different parts of the Tethyan realm" (Figure 5). In the basement there are two large Mega-units of composite terrains, which are separated by the ENE-WSW trending Mid-Hungarian Lineament. The South-Pannonian Mega-unit is the so-called TISIA (TISZA) Terrain and the North-Pannonian one is the so-called ALCAPA Mega-unit, which was formed from different smaller terrains. One of them is the PELSO in its southern part.

The evolution of the Pannonian Basin is significantly determined by the Miocene synrift, accompanied with a strong thermal heating, and leading to appraisal of crustal and mantle horizons and an extension of the basin by 10 to 30%, pre-dominantly in the NE–SW direction (Horváth and Berckhemer, 1982). These features are extremely dominant in the deep extensional sedimentary sub-basins (e.g., in the Békés basin).

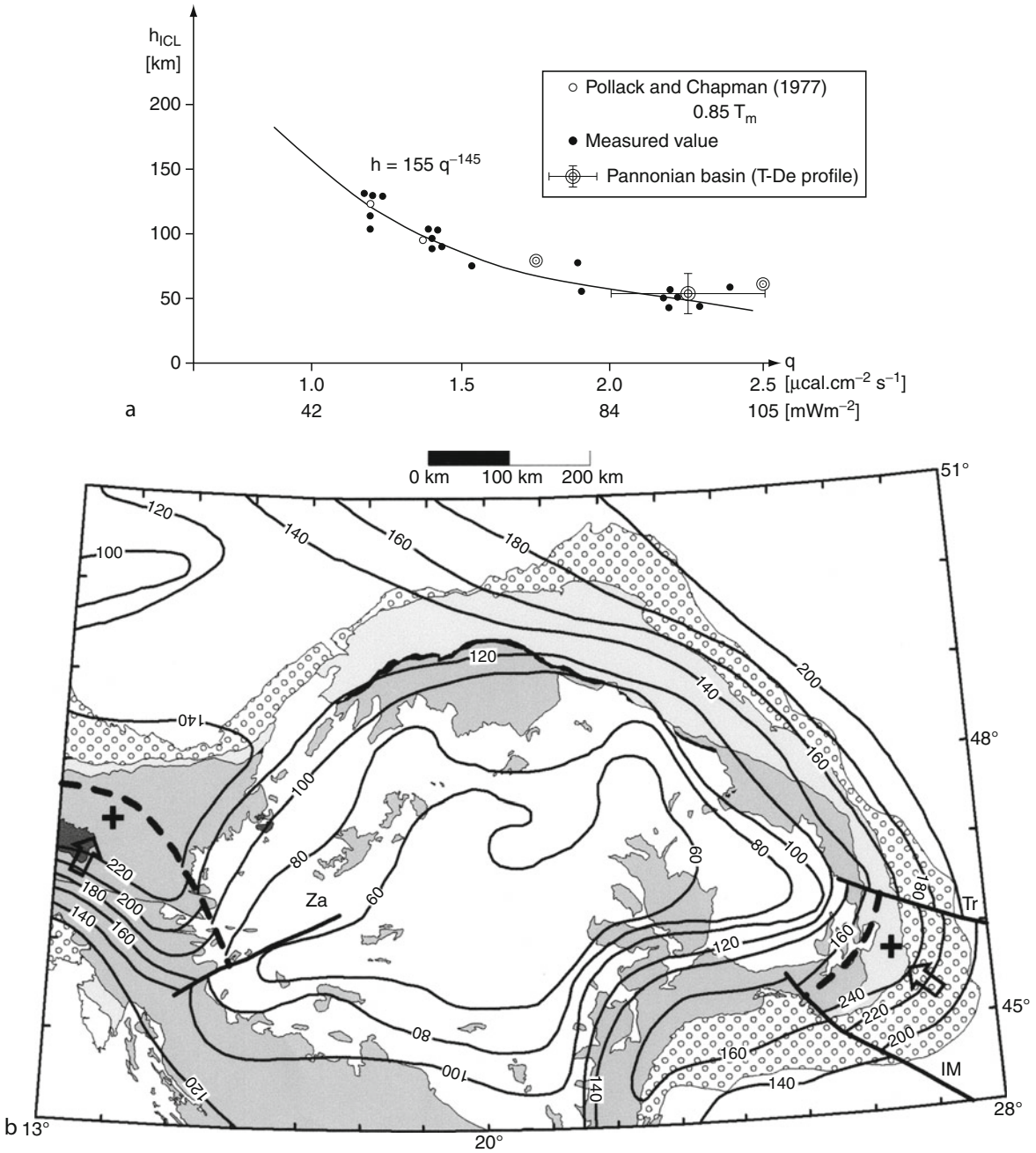


Geoelectromagnetism, Figure 5 Pre-Tertiary structural units of Hungary and depth of the pre-tertiary basement (Haas, 2001).

The spatial distribution of electrical resistivity/conductivity is influenced in the Pannonian Basin by the position, boundaries, and mosaic structure of the Mega-units and tectonic and thermal effects in the lithosphere, caused by synrift extension that has only been slowly compensated by the presently ongoing compression of the basin. The conductivity structures conserve the history of basin. Due to upheating of the Pannonian Basin – indicated by the high heat flow values at the surface of about 100 mW/m^2 – two geoelectric conductors are in an elevated position, in comparison to the surrounding mainly outer Carpathian: Paleozoic and Proterozoic regions.

Application of geoelectromagnetic methods is illustrated in five snapshots, representing various depth ranges: asthenosphere, mega-blocks, middle crust, upper crust, and near-surface.

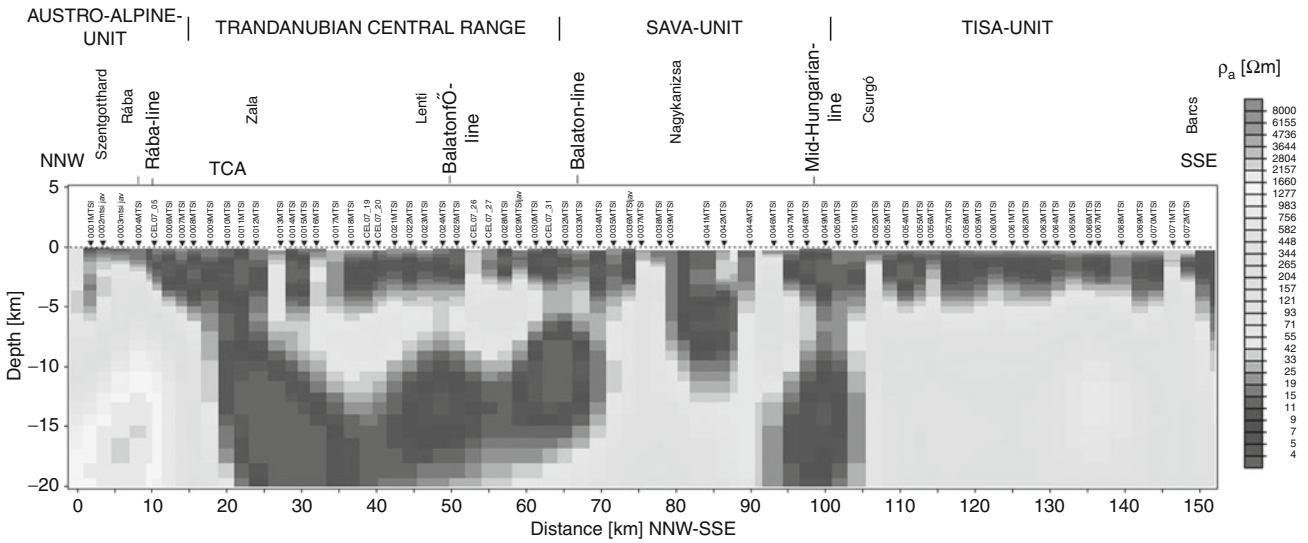
1. *Asthenosphere.* The depth of the high-conductivity asthenosphere – in which the temperature approximates the solidus of the mantle peridotite, so that a partial melting can start – is in the Pannonian Basin at about 60 km depth, indicated by Ádám (1965). In platforms and shield areas the asthenosphere is at much deeper position – if it exists at all. Its estimated value is about 250 km as concluded by extended projects such as BEAR in Finland (Korja and BEAR Working Group, 2003) and CEMES (Yu Semenov et al., 2008) in East-Central Europe (See also Figure 3). The MT indications of the conductive asthenosphere at an elevated position within Pannonian Basin were confirmed with reflection seismics (Posgay, 1975) and also with a thermobarographic diagram of well-equilibrated peridotite xenolith from the PELSO (Embey-Isztin et al., 2001). The elevated position is in close correlation with enhanced heat flow values (Figure 6a, Ádám, 1978). Its smallest depth of about 50 km was observed in the Békés extensional subbasin (Ádám et al., 1996). The map of asthenospheric depth in and around the Pannonian Basin is shown by Figure 6b as the result of seismic, seismological, and magnetotelluric measurements (See in Horváth et al., 2006, based partly on the MT data of Ádám and Wesztergom, 2001).
2. *Middle crust.* In the middle crust the transition between the brittle and ductile zone is also temperature dependent, which might be due to dehydration as proposed by Hyndman and Shearer (1989). Ádám (1978, 1987) found a relation between the mid-crustal conductivity increase and the regional heat flow values. In the Pannonian Basin there is only a weak indication of this anomaly at a depth of about 15–17 km, but this is supported by the lower limit of seismic focal depths and by strength decrease. The electrical conductivity of the lower crust was reviewed by Jones (1992).
3. *Boundaries of mega-units.* The boundaries of the Mega-units and their mosaics in the Pannonian Basin



Geoelectromagnetism, Figure 6 (**a**) Relation between the depth of the conductive asthenosphere and the regional heat flow (Ádám, 1978), (**b**) The map of the asthenospheric depth in and around the Pannonian Basin as results of seismic/seismological/magnetotelluric results (Horváth et al., 2006; Ádám and Wesztergom, 2001).

were studied among others along the deep seismic CEL-07 profile in SW Hungary with 72 deep MTS. The distance between neighboring magnetotelluric stations was in average only 2 km. The CEL-07 profile is shown together with the complicated basement tectonic structure in Figure 5.

The resistivity profile of the TE mode inversion along CEL-07 is shown in Figure 7 (the TM mode inversion profile is strongly distorted by charge accumulation at the boundary of electrical inhomogeneities falsifying the values of the resistivity distribution). The TE mode inversion clearly expresses the conductivity anomalies at the boundaries of the Mega-units: the Balaton + Balatonfő double lineaments at the southern border of the PELSO M.u. (mosaic of the ALCAPA M.u.) and the Mid-Hungarian Line at the northern

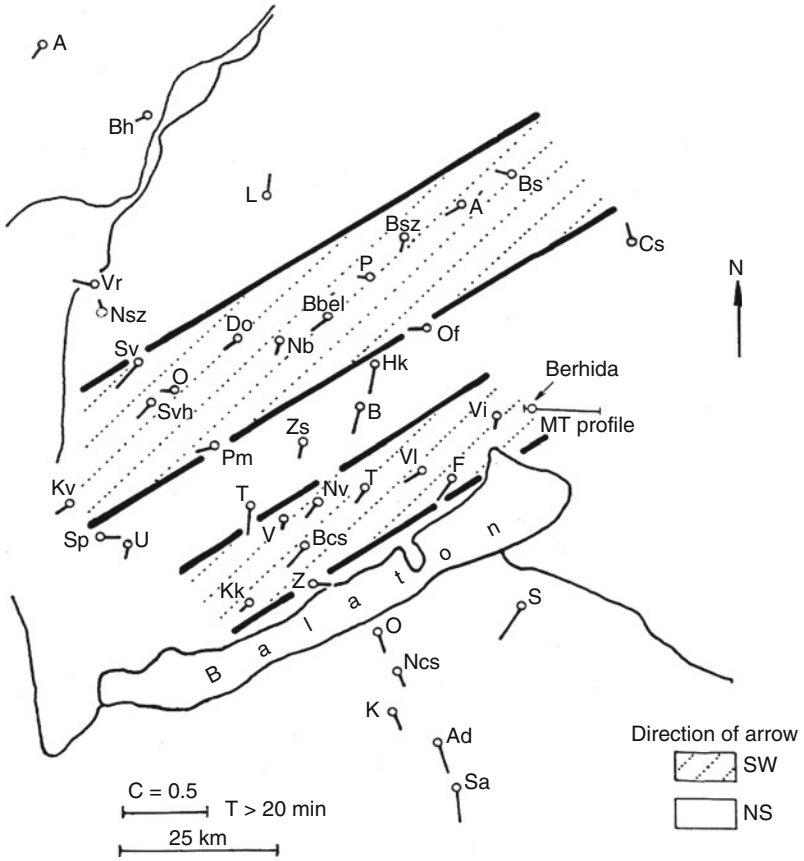


Geoelectromagnetism, Figure 7 The resistivity profile of the TE + Hz nonlinear conjugate gradient inversion of the magnetotelluric measurements along the CEL-07 profile (RMS = 1.344132). See the conductors corresponding to the deep boundary fractures (Ádám et al., 2005).

border of TISIA M.u. The horizontal electric conductance of the lineaments is about 2,500–3,000 S, starting at a depth of about 7 km. The source of the conductivity anomaly is certainly hot saline fluid containing also SiO_2 solution. Deep borehole data refer to accumulation of CO_2, increase of He concentration, etc. in the deep fractures. A high value of convective heat transfer, especially in volcanic eruption centers, also increases the electrical conductivity in these tectonic zones, which (e.g., the Balaton Line) is also indicated by the zero value of the induction vectors (Ádám et al., 2005).

4. *Crustal conductivity anomalies.* On the northern part of the CEL-07 resistivity profile there appears another conductivity anomaly that is assumed to be the continuation or ending of one of the extreme high fracture zones related to the Transdanubian Conductivity Anomaly (TCA) studied by 300 deep MT soundings (Ádám, 2001). The TCA consists of a conductive layer at about 5 km depth and of three significant fracture zones in the NE–SW direction. Their horizontal electric conductance value can reach 10,000 S. Such high conductance values make it probable that the sources of the anomaly are graphite/graphitic shists as well as fluid, which are partly within, and partly under the thick Mesozoic limestone sequence of the Bakony Mts. In this area there is significant seismic attenuation (Zsíros, 1985), and this must be due to an electrical conductor (Glover and Ádám, 2008). The large fractures as shown by the strongly reduced induction vectors (Figure 8) are connected to the hypocenters of the earthquakes (e.g., Mór in, 1809, Berhida in 1985, rarely exceeding a magnitude of 6 on Richter's scale). Glover and Ádám (2008) found a correlation between the depth to the zone of this high electrical conductivity anomaly and earthquake focal depths. According to Ogawa (2007) the earthquakes occur at the boundaries of the electric conductors, where only aseismic slip or elastic deformation can happen, rather than within the conductors themselves. In this way the energy can be transported towards the brittle zone, causing devastating earthquakes. In the presence of the hypothetical enhancement of magnetic permeability at the Curie depth (Kiss et al., 2005), MT interpretation of upper and middle crustal inhomogeneities should be carried out with a special care.
5. *Near-surface.* In order to study the near-surface depth range, either higher frequencies or local sources should be applied. Various high-frequency variants have been developed: instead of the classical MVP the so-called VLF; instead of the classical magnetotellurics: audiomagnetotellurics (AMT), radiomagnetotellurics (RMT), or VLF-EM. If the source is finite, we quickly arrive at various near-surface electromagnetic induction and direct current methods.

As a result of the various EM induction methods becoming increasingly important, the number of geoelectromagnetic papers published in international journals has shown an upward trend. Recently, Bedrosian (2007) reviewed MT measurements, which were carried out in the Himalayas (Lemonnier et al., 1999), in Southern Tibet (Li et al., 2003), in the Andean margin (Brasse et al., 2002), etc. In each of these dynamic areas the electrical conductivity anomalies appear at their most active segments, which has been also indicated by thermal activation (in forms of high heat flow), earthquakes, volcanoes, recent crustal movements, regional metamorphism



Geoelectromagnetism, Figure 8 Wiese induction vectors along the deep fractures of the TCA (Wallner, 1977).

(e.g., graphitization), partial melting, etc. Korja (2007) gives in his review a comprehensive description about the conductive anomalies in Europe. As Bedrosian (2007) stated, for a correct interpretation of MT (EM) data, due to the nonuniqueness of the geophysical inverse problems, independent geological and geophysical information is necessary. This conclusion is highly confirmed by our experiences with geoelectromagnetic results from the Pannonian Basin.

For further reading on various topics in geoelectromagnetism covered in this entry, both practical and theoretical, the reader is referred to the specialized monographs by Wait (1962), Berdichevsky and Zhdanov (1984), Weaver (1994), Simpson and Bahr (2005), and Zhdanov (2009) among others in sequence of the publication year.

Bibliography

Ádám, A., 1965. Einige Hypothesen über den Aufbau des oberen Erdmantels in Ungarn. *Gerlands Beitrage zur Geophysik*, **74**(1), 20–40.

Ádám, A., 1978. Geothermal effects in the formation of electrically conducting zones and temperature distribution in the Earth. *Physics of the Earth and Planetary Interiors*, **17**(2), P21–P28.

Ádám, A., 1987. Are there two types of conductivity anomaly (CA) caused by fluid in the crust? *Physics of the Earth and Planetary Interiors*, **45**, 209–215.

Ádám, A., 2001. Relation of the graphite and fluid bearing conducting dikes to the tectonics and seismicity (Review on the Transdanubian crustal conductivity anomaly). *Earth Planets and Space*, **53**, 903–918.

Ádám, A., and Wesztergom, V., 2001. An attempt to map the depth of the electrical asthensophere by deep magnetotelluric measurements in the Pannonian Basin (Hungary). *Acta Geologica Hungarica*, **44**, 167–192.

Ádám, A., Szarka, L., Prácser, E., and Varga, G., 1996. Mantle plumes or EM distortions in the Pannonian Basin? (Inversion of the deep magnetotelluric (MT) soundings along the Pannonian Geotraverse). *Geophysical Transactions*, **40**, 45–78.

Ádám, A., Novák, A., and Szarka, L., 2005. Tectonic weak zones determined by magnetotellurics along the CEL-7 deep seismic profile. *Acta Geodaetica et Geophysica Hungarica*, **40**(3–4), 413–430.

Arora, B. R., 1997. Chap 4. Geomagnetic deep sounding. In Arora, B. R., and Sri, Niwas (eds.), *Natural Source Electromagnetic Induction in the Earth*. New Delhi: New Age International, pp. 80–128.

Avdeev, P. B., 2005. Three-dimensional electromagnetic modelling and inversion from theory to application. *Surveys in Geophysics*, **26**, 767–799.

Bahr, K., 1991. Geological noise in magnetotelluric data: a classification of distortion types. *Physics of the Earth and Planetary Interiors*, **66**, 24–38.

Bailey, R. C., 1973. Global geomagnetic sounding – Methods and results. *Physics of the Earth and Planetary Interiors*, **7**, 234–244.

Bedrosian, P. A., 2007. MT+, Integrating magnetotellurics to determine earth structure, physical state and processes. *Surveys in Geophysics*, **28**, 121–167.

Berdichevsky, M. N., and Dmitriev, M. N., 2008. *Models and Methods of Magnetotellurics*. Berlin: Springer, 563 pp.

Berdichevsky, M. N., and Zhdanov, M. S. 1984. *Advanced Theory of the Deep Geomagnetic Sounding*. Amsterdam: Elsevier, 408 pp.

Bibby, H. M., Caldwell, T. G., and Brown, C., 2005. Determinable and non-determinable parameters of galvanic distortions in magnetotelluric. *Geophysical Journal International*, **163**, 915–930.

Brasse, H., Laezaeta, P., Rath, V., Schwalenberg, K., Soyer, V., and Haak, V., 2002. The Bolivian Altiplano conductivity anomaly. *Journal of Geophysical Research*, **107**, doi:10.1029/2001/J.B. 000391.

Cagniard, L., 1953. Basic theory of the magnetotelluric method of geophysical prospecting. *Geophysics*, **18**, 605–635.

Caldwell, T. G., Bibby, H. M., and Brown, C., 2004. The magnetotelluric phase tensor. *Geophysical Journal International*, **158**, 457–469.

Embey-Isztin, A., Downes, H., and Dobosi, G., 2001. Geochemical characterization of the Pannonian Basin mantle lithosphere and asthenosphere: an overview. *Acta Geologica Hungarica*, **44**, 259–280.

Gamble, T. B., Goubau, W. M., and Clarke, J., 1979. Magnetotellurics with a remote reference. *Geophysics*, **44**, 53–68.

Glover, P. W. J., and Ádám, A., 2008. Correlation between crustal high conductivity zones and seismic activity and the role of carbon during shear deformation. *Journal of Geophysical Research*, **113**, B 12210: 1–8. doi:10.1029/2008JB005804.

Groom, R. W., and Bailey, R. C., 1989. Decomposition of magnetotelluric impedance tensors in the presence of local three-dimensional galvanic distortion. *Journal of Geophysical Research*, **94**, 1913–1925.

Haas, J. (ed.), 2001. *Geology of Hungary*. Budapest: Eötvös University Press, 317 pp.

Horváth, F., and Berckhemer, H., 1982. Mediterranean backarc basin. In Berckhemer, H., and Ksu, H., (eds.), *Alpine-Mediterranean Geodynamics. Geodynamics Series* Vol. 7. Washington: American Geophysical Union, Boulder: Geological Society of America, pp. 141–173.

Horváth, F., Bada, G., Szafián, P., Tari, G., Ádám, A., and Cloetingh, S., 2006. Formation and deformation of the Pannonian Basin constraints from observational data. In Gee, D. G., and Stephenson, R. A. (eds.), *European Lithosphere Dynamics*. London: Geological Society, Memoires no. 32, pp. 191–206.

Hyndman, R. D., and Shearer, P. M., 1989. Water in the lower continental crust modelling magnetotelluric and seismic reflection results. *Geophysical Journal International*, **98**, 343–365.

Jankowski, J., Praus, O., and Józwiak, V., 2005. Carpathian anomaly of electrical conductivity: history of its discovery and interpretation. *Publications of the Institute of Geophysics Polish Academy of Sciences*, **C-95**(386), 17–27.

Jones, A. G., 1992. Electrical conductivity of the continental lower crust. In Fountain, D. M., Arculus, R. J., and Kay, R. W. (eds.), *Continental Lower Crust*. Amsterdam: Elsevier, pp. 81–143.

Kántás, K., 1954. The results and perspectives of the Hungarian telluric research (in Hungarian). *Bányászati Lapok*, **87**, 17.

Kelbert, A., Schultz, A., and Egbert, G., 2009. Global electromagnetic constraints on transition-zone water content variation. *Nature*, **460**, 1003–1006.

Kiss, J., Szarka, L., and Prácser, E., 2005. Second order magnetic phase transition in the Earth. *Geophysical Research Letters*, **32**, L24310, doi:10.1029/2005GL024199.

Korja, T., 2007. How is the European lithosphere imaged by magnetotellurics. *Surveys in Geophysics*, **28**, 239–272.

Korja, T., and BEAR Working Group 2003. Is there an electrical asthenosphere beneath the Fennoscandinavian Shield. In *EGU-AGU Joint Meeting*, Nice, 2003–2004, Abstract EAE03-A-14804.

Lamb, H., 1883. On electrical motions in a spherical conductor. *Philosophical Transactions of the Royal Society of London*, **174**, 519–549.

Larsen, J., 1977. Removal of local surface conductivity effects from low frequency mantle response curves. *Acta Geodaetica Geophysica et Montanistica Acad. Sci. Hung.*, **12**, 183–186.

Lemonnier, C., Marquis, G., Perrier, F., Avouac, J. P., Chitraker, G., Kafle, B., Sapkota, S., Gautam, U., Tiware, D., and Bano, M., 1999. Electrical structure of the Himalaya of Central Nepal: high conductivity around the mid-crustal ramp along MHT. *Geophysical Research Letters*, **26**, 3261–3264.

Li, S., Unsworth, M. J., Booker, J. R., Wei, W., Tan, H., and Jones, A. G., 2003. Partial melt or aqueous fluid in the midcrust of Southern Tibet? Constraints from INDEPTH magnetotelluric data. *Geophysical Journal International*, **153**, 289–304.

Novák, A., 2009. Electromagnetic imaging in geophysics with tensorial invariants: from the near surface to Transdanubian deep structures. PhD Dissertation (in Hungarian), Sopron (Hungary) 187 pp.

Ogawa, Y., 2007. Correlation of resistivity structure, seismicity and deformation from some case studies in Japan. *Lecture at the JAS003 Conference of the IUGG General Assembly* in Perugia.

Parkinson, W. D., 1959. Direction of rapid geomagnetic fluctuation. *Geophysical Journal of the Royal Astronomical Society*, **2**, 1–14.

Posgay, K., 1975. Mit Reflexionsmessungen bestimmte Horizonte und Geschwindigkeits Verteilung in der Erdkruste und Erdmantel. *Geophysical Transactions*, **23**, 13–18.

Prácser, E., and Szarka, L., 1999. A correction to Bahr's "phase deviation" method for tensor decomposition. *Earth Planets and Space*, **51**, 1019–1022.

Rodi, W., and Mackie, R. I., 2001. Nonlinear conjugate gradients algorithm for 2-D magnetotelluric inversion. *Geophysics*, **66**, 174.

Rokityansky, I. I., 1982. *Geoelectromagnetic Investigation of the Earth's Crust and Mantle*. Heidelberg: Springer, 381 pp.

Schmucker, U., 1959. Erdmagnetische Tiefensondierung in Deutschland 1957/1959. Magnetogramme und erste Auswertung. *Abhandlungen der Akademie der Wissenschaften in Göttingen. Mathematisch-Physikalische Klasse. I*, **H5**, 51.

Schmucker, U., 1970. Anomalies of geomagnetic variations in the south-western United States. *Bulletin of Scripps Institute Oceanography*, **13**, 165.

Schmucker, U., 2003. Horizontal spatial gradient sounding and geomagnetic depth sounding in the period range of daily variation. In Hördt, A., and Stoll, J. (eds.), *Protokoll über das Kolloquium elektromagnetische Tiefenforschung*. Potsdam, Germany: Bibliothek des Wiss. Albert Einstein, pp. 228–237.

Schultz, A., and Larsen, J. C., 1983. Analysis of zonal field morphology and data quality for a global set of magnetic observatory daily main values. *Journal of Geomagnetism and Geoelectricity*, **35**, 835–846.

Simpson, F., and Bahr, K., 2005. *Practical Magnetotellurics*. Cambridge: Cambridge University Press, 270 pp.

Siripunvaraporn, W., and Egbert, G. D., 2000. An efficient data subspace inversion for two-dimensional magnetotelluric data. *Geophysics*, **65**, 791–803.

Szarka, L., and Menvielle, M., 1997. Analysis of rotational invariants of the magnetotelluric impedance tensor. *Geophysical Journal International*, **129**, 133–142.

Szarka, L., Ádám, A., and Menvielle, M., 2005. Field test of a quick-look imaging method based on rotational invariants of the 3D magnetotelluric tensor. *Geophysical Prospecting*, **53**, 325–334.
Tikhonov, A. N., 1950. On investigation of electrical characteristics of deep strata of Earth's crust (in Russian). *Doklad Akademia Nauk SSSR*, **73**, 295–297.
Vozar, J., and Semenov, V. Y., 2010. Compatibility of induction methods for mantle soundings. *Journal of Geophysical Research*, **115**, B 03101, 9 pp, doi:10.1029/2009JB006390.
Wait, J. R., 1962. *Geoelectromagnetism*. New York: Academic Press, 268 pp.
Wallner, Á., 1977. The main features of the induction arrows in the area of the Transdanubian conductivity anomaly. *Acta Geodaetica Geophysica et Montanistica*, **12**, 145–150.
Weaver, J. T., 1994. *Mathematical Methods for Geoelectromagnetic Induction*. Taunton: Wiley, p. 316.
Weaver, J. T., Agarwal, A. K., and Lilly, F. E., 2000. Characterization of the magnetotellurid tensor in terms of its invariants. *Geophysical Journal International*, **141**, 321–336.
Weaver, J. T., Agarwal, A. K., and Lilly, F. E., 2006. The relationship between the magnetotelluric tensor invariants and the phase tensor of Caldwell, Bibby and Brown. *Exploration Geophysics*, **37**, 261–267.
Wiese, H., 1965. *Geomagnetische Tiefentellurik*. Berlin: Akad. Verlag, 146 pp.
Yu Semenov, V., Ádám, A., Hvozdara, M., and Wesztergom, V., 1997. Geoelectrical structure of the Earth's mantle in the Pannonian Basin. *Acta Geodaetica et Geophysica Hungarica*, **32**(1–2), 151–168.
Yu Semenov, V., Pek, J., Ádám, A., Józviak, W., Ladanyvskyy, B., Logvinov, I. M., Pushkarev, P., and Vozar, I., 2008. Electrical structure of the upper mantle beneath Central Europe: Results of the CEMES Project. *Acta Geophysica*, **56**, 957–981.
Zhdanov, M. S., 2009. *Geophysical Electromagnetic Theory and Methods*. Amsterdam: Elsevier, 848 pp.
Zhdanov, M. S., Golubev, N. G., Iv Varentsov, M., Abramova, L. M., Schneer, M. N., Berdischevsky, M. N., Zhdanova, O. N., Gordienko, V. V., Kulik, S. N., and Bilinsky, A. I., 1986. 2D model fitting of a geomagnetic anomaly in the Soviet Carpathians. *Annales Geophysicae*, **B4**(3), 335–342.
Zsíros, T., 1985. An estimation of seismic hazard in Hungary. *Gerlands Beiträge zur Geophysik Leipzig*, **94**, 111–122.

Cross-references

Electrical Properties of Rocks
Instrumentation, EM
Magnetotelluric Data Processing
Magnetotelluric Interpretation

GEOID

Paramesh Banerjee
Earth Observatory of Singapore, Nanyang Technological University, Singapore

Synonyms

Equipotential surface at mean sea level

Definition

Geoid. The gravitational equipotential surface of the earth that matches with the best approximation of the "mean sea level."

Tectonic forces. Forces originating from within the earth that alters surface configuration of the earth through uplift, movement, and deformation of the parts of the earth's crust.

Introduction

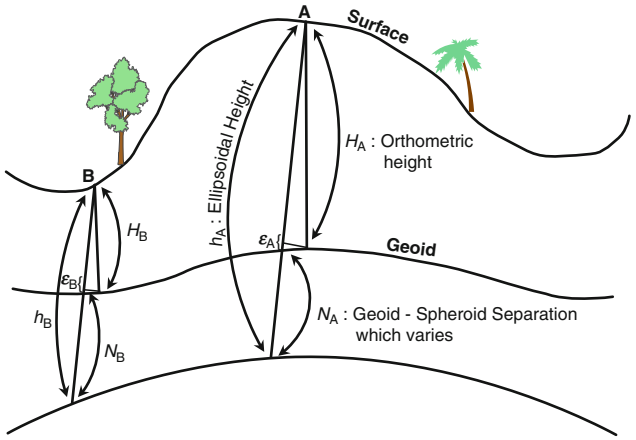
The earth does not have a perfect geometrical shape. Its shape matches very closely to that of an oblate spheroid. Topographical variations of several kilometers in the form of mountain ranges and oceanic trenches exist. Surface of the oceans, which conforms to a uniform gravitational potential, acts as an effective and useful way to define the shape of the earth. This oceanic surface serves as a reference surface for the earth.

Continuously changing configurations of the earth's surface are mostly engineered by the motion of the tectonic plates. The tectonic motion of a lithospheric plate and its deformation are responses to the force acting on it. Lateral density variation in the lithosphere produces both horizontal as well as vertical stress gradients. Gravitational Potential Energy gradient is thus directly proportional to the geoid height variation. Recent advances in space geodetic techniques make it now possible to map the geoid with unprecedented accuracy in both spatial and temporal domains, thus providing a new tool to study tectonics.

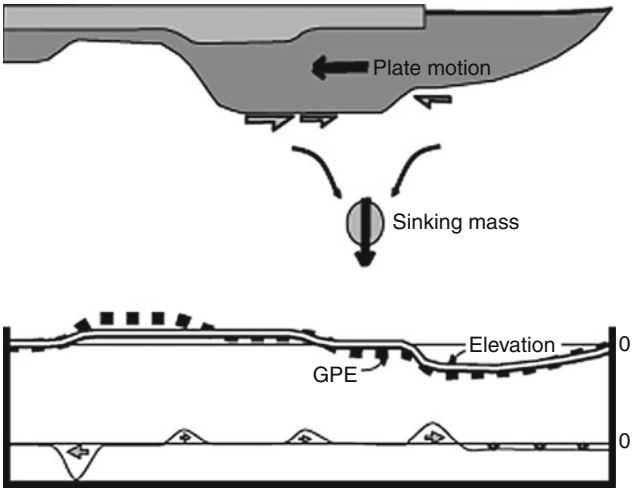
Geoid

Like any liquid in hydrostatic equilibrium, globally interlinked ocean water surface (without the effects of tides and waves), defines a gravity equipotential surface. Earth's gravity potential field contains infinite number of equipotential surfaces, almost parallel to each other (in case of an idealized earth). That particular equipotential surface which matches closest to the actual realization of the global mean sea level is the geoid. The Geoid surface is relatively smoother compared to the topographical surface of the earth. Whereas Earth's surface excursions range between +8,000 m at Mt Everest and −11,000 m at Maryana Trench, maximum geoid surface variation is about 200 m.

Over the oceans, realization of the geoid is easy as it is the "mean sea level". Geoid over the continents lies below the surface of the earth, and can be visualized as water level in the imaginary canals cutting across the continents and interconnecting the surrounding oceans. Because surveyor's spirit level or plumb bob follows the gravity equipotential, leveling traverse starting from the nearest tidal station running inside the continent, measures height above mean sea level or geoid, and is called orthometric height. All height measurements using conventional surveying techniques use geoid or "mean sea level" as the reference level. Though geoid is much smoother surface compared to the topographic surface of the earth, the geoid itself, because of inhomogeneous distribution of mass within the earth, undulates at much higher spatial wavelength domain. The undulation of the geoid is measured in reference to a "Reference spheroid" (Figure 1).



Geoid, Figure 1 A simplified cartoon showing geoid and spheroid levels both over the continents and oceans. Geoid anomaly is mostly positive over the continents and negative over the oceans.



Geoid, Figure 2 Different tectonic forces in play in a subduction zone where the subsiding slab sinks in weaker asthenosphere (slab pull), which excites mantle flow, thus creating a traction at the base of the plate (shown by small *gray arrows* at the base of the plate). Continental crust and oceanic crust are shown in *light gray* and *black* respectively, whereas mantle lithosphere is in *dark gray*. The lower figure shows surface topography, gravitational potential energy (GPE), and GPE gradient (lower line) across the plate. GPE gradient produced by the lateral variation of the lithospheric density/topography structure creates forces within lithosphere as shown by the *arrows*. Figure 1; Humphreys and Coblentz, 2007. Reproduced by permission of American Geophysical Union. Copyright [2007], American Geophysical Union.

Geoid anomaly at any point can be computed from gravity field measurements, but requires worldwide gravity data. Perturbations of the orbits of the artificial satellites have been used to model global gravitational field, covering both continents and oceans. Satellite and radar altimetry today provides the most refined technique of measuring the geoid over the oceans with an accuracy better than 10 cm. Starting with the first satellite by the then USSR in 1957, phenomenal improvement in measuring Earth's gravity field has been achieved. Important altimetric satellite missions contributing significantly in modeling ocean geoid include GEOS3(1975), SEASAT (1978), GEOSAT(1985), ERS-1(1991), TOPEX/POSEISDON(1992), ERS-2(1996), etc. Launching of twin-satellite system GRACE in 2002 was another significant event toward measuring Earth's gravity and geoid, both in temporal and spatial domains, covering land as well as oceans, with an unprecedented accuracy.

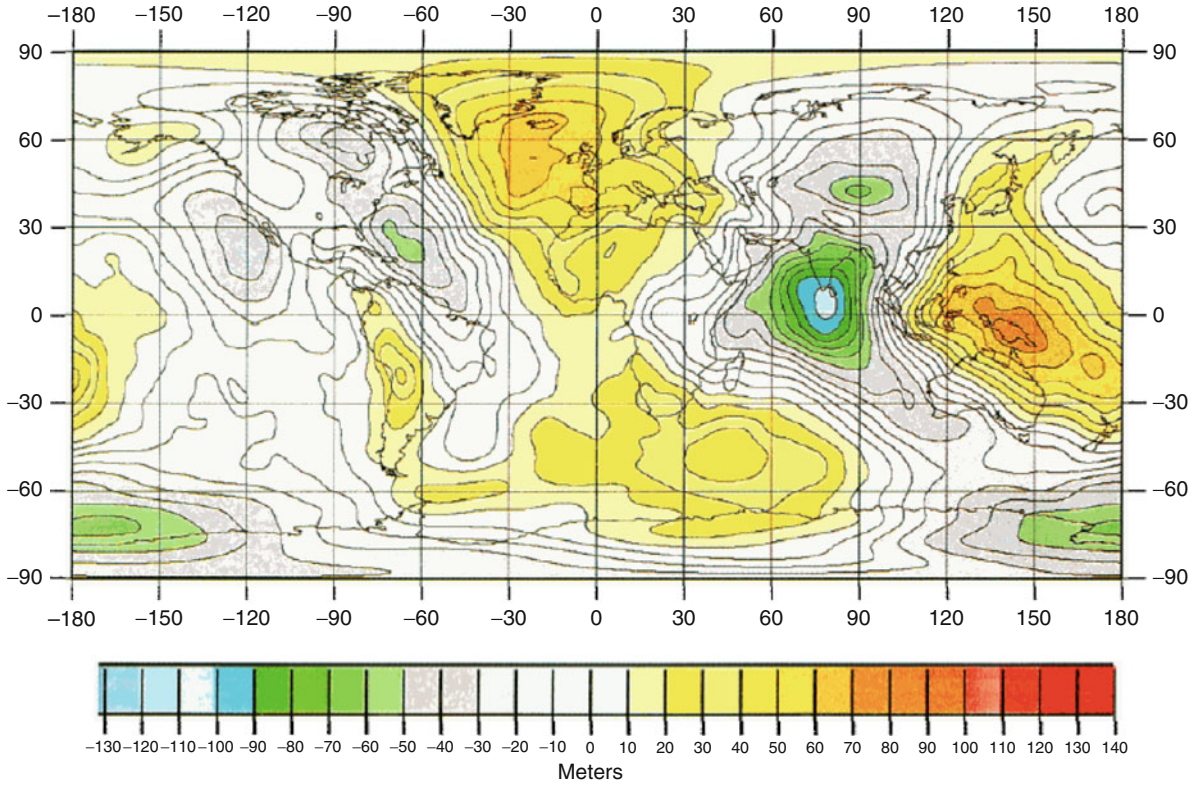
Tectonic forces

Tectonic forces acting on a lithospheric plate result in plate motion as well as accumulation of stresses, both in vertical and horizontal directions. Surface deformation and earthquakes are direct consequences of these stresses. Origin of the tectonic forces mostly lies in the flow in the asthenosphere which may either drive or resist the motion of the overlying plate. In the subduction zone, gravitational pull of the down going slab (slab pull) plays a more important role in driving the plate motion than the "ridge pushes" where slabs are driven away from the mid-oceanic ridges. Plate motion also creates traction on the base of the overlying continents (Humphrey and Coblentz, 2007). Another important source of force is generated by the lateral variation in density and thickness in the lithosphere which generates a gradient in gravitational potential energy (GPE). Figure 2 illustrates how a subducting slab generates GPE gradient and basal traction force. Tibetan plateau is an example of high GPE where continental collision produced high topography, thickened crust, extensional stresses on the plateau, and compression at the edges.

Geoid height variations are linearly proportional to the GPE gradients in an isostatically balanced lithosphere (Turcotte and Schubert, 2002). Geoid gradients can thus be used to study lithospheric stress gradients. Lithosphere-scale density inhomogeneities are reflected in shorter wavelength geoid anomalies. This, in combination with integrated density moment computed using a crustal density model at longer wavelength, can be used to produce GPE (Humphreys and Coblentz, 2007).

Geoid anomaly and tectonics

Deviation of the actual geoid surface from the reference spheroid is known as geoid anomaly. Figure 3 shows the global geoid height anomaly (Bowin, 2000). The largest geoid anomaly is a 100 m deep "hole" south of India, in



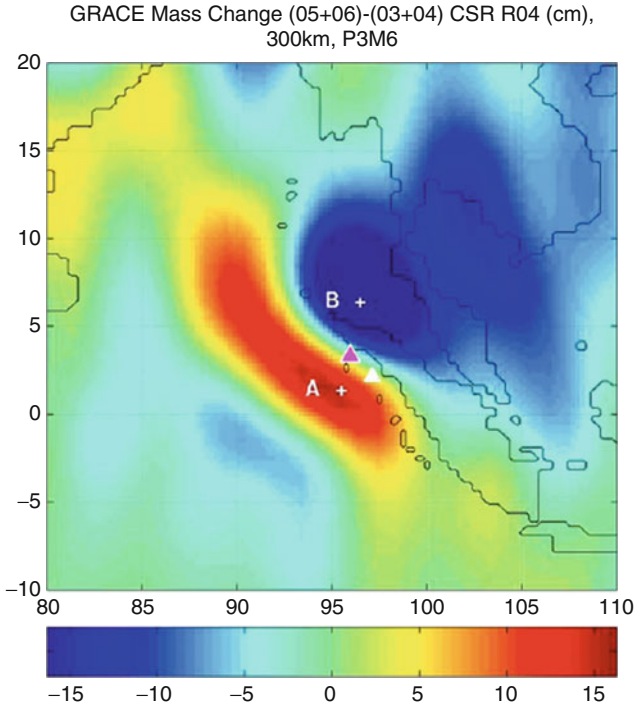
Geoid, Figure 3 The geoid height anomaly map of the world. The contour line interval is 10 m. Plate 1; Bowin, 2000. Reproduced by permission of American Geophysical Union. Copyright [2000], American Geophysical Union.

the Indian Ocean. This huge geoid "low" is caused by a mass deficiency in the mantle underneath the sea. Other short wavelength geoid anomalies exist along the oceanic trenches where one tectonic plate subsides underneath another plate, or along the mid-oceanic ridges where hot mantle material swells up to produce new oceanic crust. Geoid anomaly generally has a positive correlation with Free-air gravity anomaly. However, gravity anomaly is more sensitive to the shallower features whereas geoid anomaly is more sensitive to the deeper mass in homogeneities.

Study of the geoid anomaly helps to understand the distribution of large scale density inhomogeneity, the state of isostatic compensation, and also about the flexural property of the lithosphere. Global geoid anomaly data indicates that, when loaded (or unloaded) with a large dimension surface load (e.g., a mountain formation, or glaciation-deglaciation process) during a prolonged period, mantle flows and adjusts, and the surface load eventually attains hydrostatic equilibrium. In contrast, short-period or small-scale surface load is not isostatically compensated. These are explained by a viscoelastic lithosphere which is strong enough to support at a smaller spatial and temporal scale, but has a finite viscosity and flows like a fluid at a larger spatial and temporal domain. Study of spatial as well as temporal variation of the geoid anomaly can thus be a very useful tool to understand mechanical properties of the lithosphere. It may be noted that isostatically balanced mass distribution does not cause zero geoid anomaly. An isostatically compensated mountain range of 3 km height causes a positive geoid anomaly of ~16 m, where a 5 km deep ocean basin would result a negative geoid anomaly of about 9 m (Fowler, 1996). As geoid anomaly is more sensitive to lateral variation in lithospheric density structure, surface measurements of geoid anomaly can provide a very useful tool to study subcrustal configuration across a plate boundary (Banerjee et al., 1999).

GRACE and coseismic deformation

GRACE (Gravity Recovery and Climate Experiment) twin satellite system has made it possible to measure temporal variation of geoid height at a global scale with an accuracy of 2–3 mm (Tapley et al., 2004). The 2004 Sumatra-Andaman earthquake (Mw 9.2) provided an opportunity to measure coseismic as well as post-seismic gravity and mass variations using GRACE data (Han et al., 2006; Chen et al., 2007). Sudden change in lithospheric density structure as well as vertical movement could be detected in GRACE data (Figure 4). With this leap in space geodesy, remote sensing of Earth's Geoid provides another important tool to study tectonic processes.



Geoid, Figure 4 Temporal mass change in Sumatra-Andaman region during 2004 (Mw = 9.2) earthquake. Using GRACE data, it was possible to capture coseismic deformation of the December 26, 2004, Sumatra-Andaman event, as well as part of the coseismic deformation of the March 27, 2005, Nias earthquake. Figure 4; Chen et al., 2007. Reproduced by permission of American Geophysical Union. Copyright [2007], American Geophysical Union.

Summary

Geoid is the best available physical model of the earth. It is measurable using gravimetry and satellite altimetry data. Because geoid is a gravity equipotential surface, near-surface configuration of the earth's structure is reflected in geoid undulation. Tectonic forces that causes change in Earth's lithospheric configuration, are thus correlatable with the temporal and spatial changes of the geoid. Recent advances in satellite technology has enabled to measure and monitor geoid undulations with an unprecedented precision. As a result, many important Earth System processes like tectonic plate movement, surface deformation, sea level changes, etc are now being measured and monitored on a global scale. These have contributed significantly towards better understanding of the tectonics of the Earth.

Bibliography

Banerjee, P., Foulger, G. R., Satya, P., and Dabral, C. P., 1999. Geoid undulation modeling and interpretation at Ladakh, NW Himalaya using GPS and leveling data. *Journal of Geodesy*, **73**, 79–86.

Bowin, C., 2000. Mass anomaly structure of the Earth. *Reviews of Geophysics*, **38**(3), 355–387.

Chen, J. L., Wilson, C. R., Tapley, B. D., and Grand, S., 2007. GRACE detects coseismic and postseismic deformation from the Sumatra-Andaman earthquake. *Geophysical Research Letters*, **34**, L13302, doi:10.1029/2007GL030356.

Fowler, C. M. R., 1996. *The Solid Earth*. Cambridge: Cambridge University Press, pp. 214–216.

Han, S. C., Shum, C. K., Bevis, M., Ji, C., and Kuo, C.-Y., 2006. Crustal dilatation observed by GRACE after 2004 Sumatra-Andaman earthquake. *Science*, **313**(5787), 658–666, doi:10.1126/science.1128661.

Heiskanen, W. A., and Moritz, H., 1967. *Physical Geodesy*. San Francisco: W. H. Freeman. 364 pp.

http://earthobservatory.nasa.gov/Features/GRACE

Humphreys, E. D., and Coblentz, D. D., 2007. North American dynamics and western U.S. tectonics. *Reviews of Geophysics*, **45**(RG3001), 1–30.

Rapp, R. H., 1997. Use of potential coefficient models for geoid undulation determinations using a spherical harmonic representation of the height anomaly/geoid undulation difference. *Journal of Geodesy*, **71**(5), 282–289.

Stokes, G. G., 1849. On the variation of gravity at the surface of the Earth. *Transactions of the Cambridge Philosophical Society*, **8**, 672.

Tapley, B. D., Bettadpur, S., Ries, J., Thomson, P. F., and Watkins, M. M., 2004. GRACE measurements of mass variability in the Earth system. *Science*, **305**(5683), 503–505.

Torge, W., 1991. *Geodesy*. Berlin: Walter de Gruyter. 264 pp.

Turcotte, D. L., and Schubert, G., 2002. *Geodynamics*, 2nd edn. Cambridge: Cambridge University Press.

Cross-references

Geodesy, Figure of the Earth
Geoid Determination, Theory and Principles
Geoid Undulation, Interpretation
Geoid, Computational Method
Plate Driving Forces

GEOID DETERMINATION, THEORY AND PRINCIPLES

Michael G. Sideris
Department of Geomatics Engineering, University of Calgary, Calgary, AB, Canada

Definition

The *geoid* is the equipotential surface of the Earth's gravity field, which best approximates the mean sea level at rest. Its determination requires (1) knowledge of the Earth's gravity potential and (2) a way of geometrically representing the geoidal surface, which is done via its deviations – called geoid undulations N – from the surface of a known reference ellipsoid. The geometric representation is of particular importance given that the geoid is the reference surface, or datum, for the physically meaningful orthometric heights, often called heights above mean sea level.

Basic principles: the gravity, normal, and disturbing potentials

The Earth's *gravity potential* W comprises the gravitational potential V and the centrifugal potential Φ:

$$W = V + \Phi = k \iiint_v \frac{\rho}{l} \mathrm{d}v + \frac{1}{2}\omega^2\left(x^2 + y^2\right) \tag{1}$$

V is a Newtonian potential due to the attraction of the masses, expressed as an integral of the mass density ρ over the volume v; l is the Cartesian distance between the computation point and the mass element $\rho \mathrm{d}v$, and k is Newton's gravitational constant. The centrifugal potential Φ is a function of the Earth's rotational velocity ω and the distance $p = (x^2 + y^2)^{1/2}$ from the rotation axis, namely, $\Phi = \omega^2 p^2/2$.

Due to our insufficient knowledge of the density distribution, Equation 1 cannot be used directly to compute V and W. Nevertheless, a solution can be obtained by utilizing the fact that V is a harmonic function outside the masses, that is (a) it satisfies Laplace's equation

$$\nabla^2 V = \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = 0 \tag{2}$$

since the inverse distance in the integral of Equation 1 clearly satisfies $\nabla^2 l^{-1} = 0$ and (b) it vanishes as $l \to \infty$. *Stokes's theorem* postulates that V can be determined everywhere in free space from its values on the boundary surface that encloses all masses and thus knowledge of the interior density distribution is not necessary.

The above problem formulation is called *Dirichlet's problem*, or first boundary value problem (BVP) of potential theory (Hofmann-Wellenhof and Moritz, 2006). Its practical significance is rather limited since the potential is not a directly measurable quantity; it is important, however, for the expansion of V into a spherical harmonic series and for the relationship between V on the boundary surface and V in space (Poisson's integral). A more realistic BVP is *Neumann's problem*, or second BVP, in which the normal derivative $\partial V/\partial n$ is given on the boundary surface instead of V itself. This is because first-order derivatives of W constitute the *gravity vector*

$$\mathbf{g} = \nabla W = \left(\frac{\partial W}{\partial x}, \frac{\partial W}{\partial y}, \frac{\partial W}{\partial z}\right)^T \tag{3}$$

which is observable. It is, however, the third *BVP* that is particularly important in practice; it uses observables on the boundary surface that are a linear combination of V and $\partial V/\partial n$. We will see in the following that the determination of geoid undulations from gravity anomalies – which are derived from the measurements taken with a gravimeter – is such a problem.

The determination of W is a nonlinear problem and is further complicated by the fact that W itself is not harmonic since $\nabla^2 W = \nabla^2 \Phi = 2\omega^2$. Both limitations can be overcome by linearization and approximating the shape, size, and gravity field of the Earth by an equipotential ellipsoid of revolution with the same mass M and same rotational velocity ω as the Earth. In this case, the gravitational potential of the reference ellipsoid V_e can be computed analytically, and its centrifugal potential Φ_e is the same as Φ. Its gravity potential, termed *normal potential* U, is

$$U = V_e + \Phi_e = V_e + \Phi \tag{4}$$

and its gradient is the *normal gravity vector*

$$\boldsymbol{\gamma} = \nabla U = \left(\frac{\partial U}{\partial x}, \frac{\partial U}{\partial y}, \frac{\partial U}{\partial z}\right)^T \tag{5}$$

Both U and γ are analytical functions that can be computed from M, ω, and the semimajor and semiminor axes a, b of the reference ellipsoid. The difference between W and U at the same point is the *disturbing potential* T, which is obviously a harmonic function (since it is independent of the centrifugal potential, and since V_e, just like V, is harmonic being a function of the inverse distance):

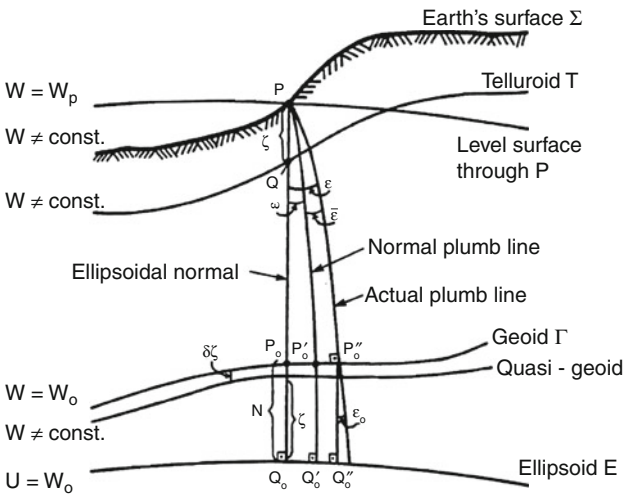
$$T = W - U = V - V_e \tag{6}$$

$$\nabla^2 T = 0 \tag{7}$$

Now since U is known, the problem of determining W has been reduced to the problem of determining T, which is a third BVP provided that an appropriate observable would be available on the boundary surface. This is indeed the case, as it will be shown in the following.

The Stokes and Molodensky geodetic BVPs

We select the geoidal surface Γ as the boundary surface that encloses all masses, we approximate Γ by the surface of the reference ellipsoid E, choose the normal potential U_o on E to be the same as the gravity potential W_o on the geoid, project a point P_o on the geoid on to a point Q_o on the ellipsoid along the ellipsoidal normal (see Figure 1), and define the geoid undulation N as the distance P_oQ_o and the *gravity anomaly* and *gravity disturbance* vectors



Geoid Determination, Theory and Principles,
Figure 1 Geometry of the Stokes BVP and the Molodensky BVP.

as $\mathbf{\Delta g} = \mathbf{g}_{P_o} - \boldsymbol{\gamma}_{Q_o}$ and $\mathbf{\delta g} = \mathbf{g}_{P_o} - \boldsymbol{\gamma}_{P_o} = \nabla T$, respectively, with magnitudes

$$\Delta g = g_{P_o} - \gamma_{Q_o} \tag{8}$$

$$\delta g = g_{P_o} - \gamma_{P_o} \tag{9}$$

We can now relate N to T, and Δg to T and its vertical derivative. Using Figure 1, only first-order terms in Taylor expansions, and assuming – because the vertical deflection angle ε is very small – that the vertical derivatives along the ellipsoidal normal, normal plumb line, and actual plumb line are all the same, we can write

$$\begin{aligned} T_{\mathrm{P_o}} = W_{\mathrm{P_o}} - U_{\mathrm{P_o}} &= W_{\mathrm{P_o}} - U_{\mathrm{Q_o}} - \left.\frac{\partial U}{\partial h}\right|_{\mathrm{Q_o}} N \\ &= W_{\mathrm{o}} - U_{\mathrm{o}} + \gamma N = \gamma N \end{aligned} \tag{10}$$

$$\begin{aligned} \delta g_{\mathrm{P_o}} = -\left.\frac{\partial T}{\partial h}\right|_{\mathrm{P_o}} &= g_{\mathrm{P_o}} - \gamma_{\mathrm{P_o}} = g_{\mathrm{P_o}} - \gamma_{\mathrm{Q_o}} - \left.\frac{\partial \gamma}{\partial h}\right|_{\mathrm{Q_o}} N \\ &= \Delta g - \frac{1}{\gamma}\left.\frac{\partial \gamma}{\partial h}\right|_{\mathrm{Q_o}} T \end{aligned} \tag{11}$$

Plumb lines are lines intersecting perpendicularly the surfaces of equal potential; that is, at every point, the gravity vector is tangent to the plumb line through that point.

The last two equations can now simply be written as

$$N = \frac{T}{\gamma} \tag{12}$$

$$\Delta g = \delta g + \frac{1}{\gamma}\frac{\partial \gamma}{\partial h} T = -\frac{\partial T}{\partial h} + \frac{1}{\gamma}\frac{\partial \gamma}{\partial h} T \tag{13}$$

Equation 12 is called *Bruns's equation* and provides N given T, while Equation 13 holds only on Γ and it is called the *fundamental equation of physical geodesy*. The normal gravity on an ellipsoid with semi-axes a, b and normal gravity values at the equator and pole γ_a, γ_b, respectively, can be computed from Somigliana's formula as a function of the point's latitude ϕ:

$$\gamma = \frac{a\gamma_a \cos^2\phi + b\gamma_b \sin^2\phi}{\left(a^2\cos^2\phi + b^2\sin^2\phi\right)^{1/2}} \tag{14}$$

Now *Stokes's BVP* can be defined as the problem of determining T, harmonic outside Γ, from measurements of Δg on the geoid that satisfy Equation 13. This is a third BVP of potential theory, mathematically described by the partial differential equation of Equation 7 with the boundary condition of Equation 13 and it provides T on and outside Γ. Because of the assumption that the geoid encloses all masses, the influence of masses outside Γ must be computed and removed mathematically from the measurements via various types of terrain reductions. These require knowledge of the density distribution of the topographic masses and are described in the section The Treatment of the Topography at the end of this chapter.

Since all gravity measurements are taken on the surface of the Earth $\sum$, a more realistic way of defining the problem is to select $\sum$ as the boundary surface rather than the geoid. This requires no terrain reductions but results in a mathematically more complicated BVP because $\sum$ is not an equipotential surface and the gravity vector is not perpendicular to it. The determination of the Earth's surface from gravity anomalies on it is an oblique derivative problem called Molodensky's problem (Molodensky et al., 1962).

In *Molodensky's BVP*, the Earth's surface is approximated by a surface called the *telluroid* T, W on $\sum$ is approximated by the normal potential U of T, and the gravity anomalies now are the difference between the gravity at the Earth's surface and the normal gravity at the telluroid, that is, $\Delta g = g_{\mathrm{P}} - \gamma_{\mathrm{Q}}$; see Figure 1. T is defined as the set of points Q above the ellipsoid that correspond to points P on $\sum$ having $U_{\mathrm{Q}} = W_{\mathrm{P}}$, $\phi_{\mathrm{Q}} = \phi_{\mathrm{P}}$, $\lambda_{\mathrm{Q}} = \lambda_{\mathrm{P}}$. This is called the Marussi telluroid and the ellipsoidal heights of its points are called normal heights H^*. The separation PQ between $\sum$ and T is called *height anomaly* ζ. The height anomaly is the solution to the Molodensky BVP, as defined by Equations 7 and 13, where now the boundary condition holds on the telluroid rather than the geoid and Δg refers to the Earth's surface rather than the geoid.

Analytical solutions to the geodetic BVPs

Since $\nabla^2 T = 0$, T and therefore N can be expressed as a spherical harmonic series (see *Gravity, Global Models*)

$$N = \frac{kM}{\gamma r}\sum_{n=2}^{\infty}\left(\frac{a}{r}\right)^n \sum_{m=0}^{n}\left[C_{nm}\cos m\lambda + S_{nm}\sin m\lambda\right]P_{nm}(\sin\phi) \tag{15}$$

the coefficients of which, S_{nm} and C_{nm}, can be determined from, for example, gravity anomalies on the geoid using the orthogonality property of spherical harmonics. P_{nm} are Legendre functions and r,ϕ,λ are the coordinates of the computation point. Applying Equation 15 to Equation 13, the spherical harmonic series for Δg is

$$\Delta g = -\frac{kM}{r^2}\sum_{n=2}^{\infty}\left(\frac{a}{r}\right)^n (n-1) \sum_{m=0}^{n}\left[C_{nm}\cos m\lambda + S_{nm}\sin m\lambda\right]P_{nm}(\sin\phi) \tag{16}$$

Another analytical solution of this problem is given by *Stokes's integral* (see *Gravity Field of the Earth*):

$$N = \frac{R}{4\pi\gamma}\iint_{\sigma}\Delta g S(\psi)\,d\sigma = \frac{1}{\gamma}\mathbf{S}\Delta g \tag{17}$$

where ψ is the spherical distance between data point and computation point, σ is the surface of a sphere of radius R (approximating the ellipsoid), and $\mathbf{S}$ is the Stokes integral operator. $S(\psi)$ is *Stokes's kernel function*

$$S(\psi) = \frac{1}{\sin(\psi/2)} - 6\sin\frac{\psi}{2} + 1 - 5\cos\psi - 3\cos\psi\ln\left(\sin\frac{\psi}{2} + \sin^2\frac{\psi}{2}\right) \tag{18}$$

$$\sin^2\frac{\psi}{2} = \sin^2\frac{\phi_P - \phi}{2} + \sin^2\frac{\lambda_P - \lambda}{2}\cos\phi_P\cos\phi \tag{19}$$

Molodensky's problem can be solved in various ways. We give here the final solution by the *method of analytical continuation to point level*; detailed derivations can be found in Moritz (1980). In this method, the gravity anomalies are reduced to the equipotential surface passing through the computation point P; see Figure 1. They are denoted by $\Delta g'$ and are obtained by the formula

$$\begin{aligned} \Delta g' &= \sum_{n=0}^{\infty} g_n, \\ g_n &= -\sum_{m=1}^{n} (h - h_P)^m \mathbf{L}_m g_{n-m}, \\ g_o &= \Delta g \end{aligned} \tag{20}$$

$\mathbf{L}$ is the *vertical derivative operator* expressed in terms of surface values as follows:

$$\begin{aligned} \mathbf{L}_m &= \frac{1}{m!}\mathbf{L}^m = \frac{1}{m!}\frac{\partial^m}{\partial h^m}, \\ \mathbf{Lf} &= -\frac{f}{R} + \frac{R^2}{2\pi}\iint_{\sigma}\frac{f - f_P}{l^3}\mathrm{d}\sigma \end{aligned} \tag{21}$$

where l is now the straight distance between the running point and the computational point P. Now, ζ at point P can be obtained by applying Stokes's integral operator $\mathbf{S}$ to $\Delta g'$:

$$\zeta = \frac{1}{\gamma}\mathbf{S}\Delta g' = \frac{1}{\gamma}\mathbf{S}\Delta g + \frac{1}{\gamma}\sum_{n=1}^{\infty}\mathbf{S}g_n = \sum_{n=0}^{\infty}\zeta_n \tag{22}$$

When the height anomaly is plotted above the ellipsoid, the resulting surface is called the *quasi-geoid*; see Figure 1. Since $h = H^* + \zeta \approx H + N$, the geoid undulation can be estimated from the height anomaly by the expression

$$\begin{aligned} N &= \zeta + \delta\zeta, \\ \delta\zeta &= H^* - H = \frac{\bar{g} - \bar{\gamma}}{\bar{\gamma}}H \approx \frac{\Delta g_B}{\gamma}H \end{aligned} \tag{23}$$

where $\bar{g}$ and $\bar{\gamma}$ are the mean values of gravity and normal gravity along the plumb line between the geoid and Earth surface and the normal plumb line between the ellipsoid and the telluroid, respectively, and $\Delta g_B = \Delta g - A_B + F$ is the Bouguer gravity anomaly at the computation point; the definitions of A_B and F are given after Equation 28. $H = \mathrm{PP''}_o$ and $H^* = \mathrm{QQ}_o$ are the *orthometric height* and *normal height*, respectively, of point P.

The operational solution to the geodetic BVPs

In contrast to the analytical solution which is model-driven, the operational solution is data-driven. It starts from the data and information about their noise and first and second moments (covariance functions) and estimates T as the "best" unbiased linear combination of the measurements. Besides being able to propagate the data noise into the results, it is also able to utilize as input, as well as predict, any functional of T. The harmonicity of T and the analytical relations between T and its functionals $f(T)$ are only used to derive the required covariance functions.

Mathematically, we want to estimate T, or more generally a signal vector $\mathbf{s}$ of functionals of T, from a vector of measurement $\mathbf{l}$, which can contain Δg or any other $f(T)$ such as, for example, deflections of the vertical, second-order gradients, geoid undulations, etc.

$$\hat{\mathbf{s}} = \mathbf{Hl} \tag{24}$$

In order to estimate the matrix $\mathbf{H}$, we require that the variances of the estimation errors $\mathbf{e} = \hat{\mathbf{s}} - \mathbf{s}$ be minimum or that the error covariance matrix $\mathbf{C}_{ee} = E\{\mathbf{ee}^T\}$ be minimum. By definition,

$$\begin{aligned} \mathbf{C}_{ee} &= E\{\mathbf{ee}^T\} \\ &= E\left\{(\mathbf{Hl} - \mathbf{s})(\mathbf{Hl} - \mathbf{s})^T\right\} \\ &= \mathbf{H}E\{\mathbf{ll}^T\}\mathbf{H}^T - E\{\mathbf{sl}^T\}\mathbf{H}^T - \mathbf{H}E\{\mathbf{ls}^T\} \\ &\quad + E\{\mathbf{ss}^T\} \\ &= \mathbf{HC}_{ll}\mathbf{H}^T - \mathbf{C}_{sl}\mathbf{H}^T - \mathbf{HC}_{ls} + \mathbf{C}_{ss} \\ &= \mathbf{C}_{ss} - \mathbf{C}_{sl}\mathbf{C}_{ll}{}^{-1}\mathbf{C}_{ls} \\ &\quad + \left(\mathbf{H} - \mathbf{C}_{sl}\mathbf{C}_{ll}{}^{-1}\right)\mathbf{C}_{ll}\left(\mathbf{H} - \mathbf{C}_{sl}\mathbf{C}_{ll}{}^{-1}\right)^T \end{aligned} \tag{25}$$

The last term in the right-hand side of the above equation is a nonnegative quadratic term and therefore $\mathbf{C}_{ee}$ can only be minimum when that term is zero, that is, when $\mathbf{H} = \mathbf{C}_{sl}\mathbf{C}_{ll}{}^{-1}$. Therefore,

$$\hat{\mathbf{s}} = \mathbf{C}_{sl}\mathbf{C}_{ll}^{-1}\mathbf{1} \tag{26}$$

$$\mathbf{C}_{ee} = \mathbf{C}_{ss} - \mathbf{C}_{sl}\mathbf{C}_{ll}^{-1}\mathbf{C}_{ls} \tag{27}$$

When the observations $\mathbf{l}$ contain noise $\mathbf{n}$, that is, consist of a signal component $\mathbf{t}$ and a noise component $\mathbf{n}$, then assuming that signal and noise are uncorrelated, the covariance matrices in the above equations become $\mathbf{C}_{ll} = \mathbf{C}_{tt} + \mathbf{C}_{nn}$, $\mathbf{C}_{ls} = \mathbf{C}_{ts}$, and $\mathbf{C}_{sl} = \mathbf{C}_{st} = \mathbf{C}_{ls}{}^T$. This method is generally known by the name of *least-squares collocation* (LSC). Knowledge of the covariance matrix of only one gravity field quantity is sufficient. From this, all other

covariance matrices can be derived by covariance propagation. A complete discussion of LSC can be found in Moritz (1980).

The treatment of the topography

The undulation of the geoid N can be obtained from the solution of Stokes's BVP provided that there are no masses outside the geoidal surface. There are many ways to mathematically account for the topographic masses, called *terrain reductions*, which preserve the total mass of the Earth but shift the masses above the geoid below it. *Helmert's second condensation reduction* will be briefly presented here as a representative from a number of possible terrain reductions. The masses above the geoid are basically condensed onto a layer on the geoid by (1) computing at point P and removing the attraction A_P of all masses above the geoid; (2) lowering station from P to P_o (see Figure 2) using the free-air reduction; and (3) restoring at P_o the attraction $A^c_{P_o}$ of the masses condensed on a layer on the geoid with density $\sigma = \rho H$. This procedure gives Δg on the geoid computed from the expression

$$\Delta g = \Delta g_P - A_P + F + A^c_{P_o} = \Delta g_P + F + \delta A \qquad (28)$$

where $F = -\partial g/\partial H \approx -\partial\gamma/\partial h$ is the free-air gravity gradient $\Delta(g_P + F)$ is the free-air gravity anomaly at P, and δA is the resulting attraction change.

The A and A^c terms in Equation 28 are obtained by applying Newton's integral in Equation 1 over the volume v of the relevant masses. For example, when the attraction A of all topographic masses above the geoid is computed, v is the volume of the masses between the geoidal surface and the surface of the topography. From Figure 2, it can be seen that $A = A_B - c$, where $A_B = 2\pi k\rho H$ (for ρ = constant) is the attraction of the infinite Bouguer plate through P called *Bugouer correction*, and c is the attraction of the residual topographic masses above and below the Bouguer plate, called the classical *terrain correction*.

δA is not the only change associated with this reduction. Due to the shifting of masses, the potential changes as well by an amount δT called the *indirect effect on the potential*:

$$\delta T = T_{P_o} - T^c_{P_o} \qquad (29)$$

where T_{P_o} is the potential of the topographic masses at P_o and $T^C_{P_o}$ is the potential of the condensed masses at P_o. The corresponding *indirect effect on the geoid* is $\delta N = \delta T/\gamma$. Due to this potential change, the use of Δg from Equation 28 into Equation 17 produces not the geoid but a surface called the *co-geoid*, depicted by the *co-geoidal height* N^c above the reference ellipsoid. Thus, before applying Stokes's equation, the gravity anomalies must be transformed from the geoid to the co-geoid by applying a small correction $\delta\Delta g$ called the *indirect effect on gravity*

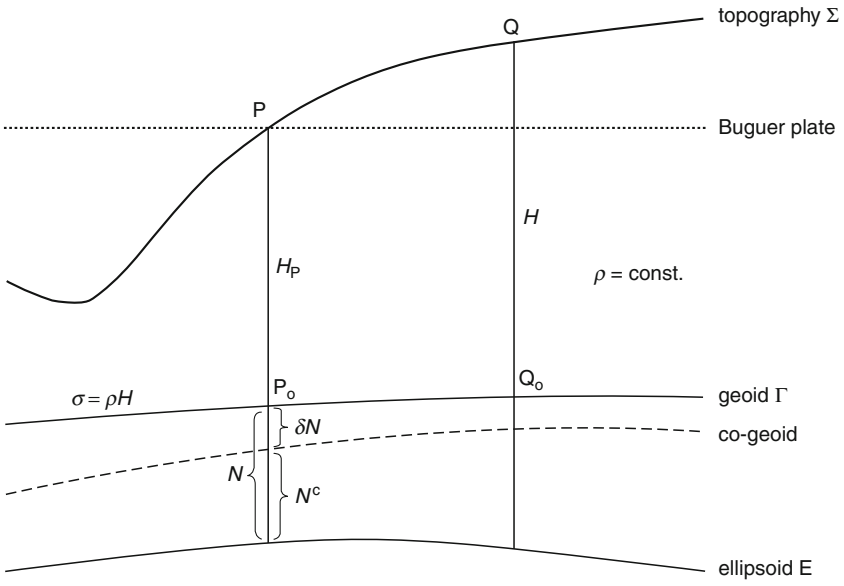
$$\delta\Delta g = -\frac{1}{\gamma}\frac{\partial\gamma}{\partial h}\delta T = -\frac{\partial\gamma}{\partial h}\delta N \qquad (30)$$

The final expression giving N can now be written as

$$N = \frac{1}{\gamma}\mathbf{S}(\Delta g + \delta A + \delta\Delta g) + \frac{1}{\gamma}\delta T = N^c + \delta N \qquad (31)$$

where N^c is the co-geoidal height and δN is the indirect effect on the geoid; see Figure 2.

Following Sideris (1990), δT and δA can formally be expressed using the planar approximation of the vertical derivative operator $\mathbf{L}$ of Equation 21, that is, without the



Geoid Determination, Theory and Principles, Figure 2 Actual and condensed topography.

term $-f/R$, which then becomes zero since $R \to \infty$ when the spherical surface is approximated by the tangent plane. The potential change is

$$\delta T = -\pi k \rho H_P^2 - 2\pi k \times \sum_{r=1}^{\infty} \frac{1}{(2r+1)!} \mathbf{L}^{2r-1}\left(\rho H^{2r+1}\right) \tag{32}$$

and the attraction change is equal to the classical terrain correction c:

$$\delta A = c = 2\pi k \sum_{r=1}^{\infty} \frac{1}{(2r)!} \mathbf{L}^{2r-1}\left[\rho (H - H_P)^{2r}\right] \tag{33}$$

These series expansions are valid for low slopes of the terrain, namely, for $(H - H_P)/l_o \leq 1$, where l_o denotes the planar distance between running and computational point.

Another commonly used terrain reduction is the *residual terrain model* (RTM) *reduction* (Forsberg and Tscherning, 1981). It computes at point P and removes only the high-frequency contributions of the topographic masses to gravity by, for example, computing the attraction of the masses between the actual topography and a very smooth "average topography" surface. The residual "geoid" produced by such reduced gravity anomalies is finally corrected by restoring the effect of the topographic masses that were removed. The RTM reduction yields a quasi-geoid rather than the geoid and, therefore, a final correction like the one in Equation 23 must be applied in the end in order to recover the actual geoid undulation.

The oceanic geoid and gravity from satellite altimetry

The only areas on Earth where, by definition, the geoid is almost directly observable are the oceans. From the difference between the ellipsoidal heights of altimetry satellites and the radar measurements to the surface of the sea, the ellipsoidal height h_s of the sea surface can be derived. h_s consists of the geoid undulation N and the sea surface topography (SST) H_s above the geoid. Thus, in principle, the geoid in the oceans can be derived from the expression

$$N = h_s - H_s \tag{34}$$

provided that a good model for the SST is available from oceanographic information. In addition, to minimize possible long-wavelength data biases, the *deflections of the vertical* are often derived numerically over the oceans as the slopes of the altimetry-derived geoid:

$$\xi = -\frac{\partial N}{r \partial \phi}, \eta = -\frac{\partial N}{r \cos\phi \partial \lambda} \tag{35}$$

Although theory exists for the solution of overdetermined and altimetry-gravimetry BVPs, where given are gravity anomalies on the continents and geoid from altimetry and/or gravity anomalies in the oceans, often in practice the oceanic geoid is also computed simply from Stokes's formula or from LSC using gravity data. Therefore, there is a need to obtain gravity anomalies in the oceans using as data altimetry-derived N and/or ξ,η.

Naturally, knowing N, one can use the following *inverse Stokes formula* to derive gravity anomalies in the oceans:

$$\Delta g_P = \frac{\gamma N_P}{R} - \frac{\gamma}{16\pi R} \iint_{\sigma} \frac{N - N_P}{\sin^3(\psi/2)} d\sigma \tag{36}$$

and then combine these gravity anomalies with the ones over land to produce a global set of gravity anomaly data which could be used in Stokes or LSC formulas to compute the global geoid. Note that the above equation has basically the form of the **L** operator acting on N, which, as expected, indicates that **L** and **S** are inverse operators. This can be proven easily, though rigorously, in planar approximation in the frequency domain; see details in *Geoid, Computational Method*.

Gravity anomalies in the oceans can also be computed from the deflections of the vertical, using the *inverse Vening Meinesz formula*

$$\Delta g_P = \frac{\gamma}{4\pi} \iint_{\sigma} (\xi \cos\alpha + \eta \sin\alpha) H(\psi) d\sigma \tag{37}$$

$$H(\psi) = \frac{\cos(\psi/2)}{2\sin(\psi/2)} \left(-\frac{1}{\sin(\psi/2)} + \frac{3 + 2\sin(\psi/2)}{1 + \sin(\psi/2)} \right) \tag{38}$$

where α is the azimuth of the line between the computation and the data point. A similar formula, termed the *deflection-geoid formula* (Hwang, 1998), also exists that gives N from ξ,η:

$$N_P = \frac{R}{4\pi} \iint_{\sigma} (\xi \cos\alpha + \eta \sin\alpha) C'(\psi) d\sigma \tag{39}$$

$$C'(\psi) = -\cot(\psi/2) + \frac{3}{2}\sin\psi \tag{40}$$

Another often used formula is the one that relates the vertical derivative of Δg to the horizontal derivatives of ξ and η. The spherical form of Laplace's equation is

$$\nabla^2 T = \frac{\partial^2 T}{\partial r^2} + \frac{2}{r}\frac{\partial T}{\partial r} + \frac{1}{r^2}\frac{\partial^2 T}{\partial \phi^2} - \frac{\tan\phi}{r^2}\frac{\partial T}{\partial \phi} + \frac{1}{r^2 \cos^2\phi}\frac{\partial^2 T}{\partial \lambda^2} = \frac{\partial \Delta g}{\partial r} + \gamma \frac{\partial \xi}{r \partial \phi} - \frac{\tan\phi}{r}\gamma\xi + \gamma \frac{1}{r\cos\phi}\frac{\partial \eta}{\partial \lambda} = 0 \tag{41}$$

If we neglect the third term on the right-hand side, we get

$$\frac{\partial \Delta g}{\partial r} = -\gamma\left(\frac{\partial \xi}{r\partial \phi} + \frac{1}{r\cos\phi}\frac{\partial \eta}{\partial \lambda}\right) \quad (42)$$

It is clear that since Δg, ξ, η, N, and their various derivatives are all functionals of T, they could all be used as input in LSC to predict both Δg and N in the oceans. Therefore, collocation could be used to either predict directly undulations and anomalies in the oceans from any gravity field quantity, or to densify or interpolate data necessary for the efficient evaluation of integrals like inverse Stokes or inverse Vening Meinesz. Details can be found in *Geoid, Computational Method*.

Summary

We have discussed briefly how the geoid is defined and how it can be estimated either analytically as the solution of the third BVP (or Molodensky's BVP) from gravity anomalies all over the Earth or as a "best" unbiased linear combination of all available gravimetric data using least-squares collocation. In either case, data have to be reduced to the geoid via an appropriate terrain reduction. The methodology used in the numerical determination of geoid undulations is described in the Chapter on the *Geoid, Computational Method*. The reader interested in obtaining a more comprehensive view of gravity field approximation is encouraged to also read the other chapters on geoid and gravity included in this volume. Finally, with dedicated gravity satellites like GRACE, the computation of temporal variations of the geoid and its application to other geosciences is now possible. This, however, is beyond the scope of this short entry and will not be discussed here.

Bibliography

Forsberg, R., and Tscherning, C. C., 1981. The use of height data in gravity field approximation by collocation. *Journal of Geophysical Research*, **86**(B9), 7843–7854.

Hofmann-Wellenhof, B., and Moritz, H., 2006. *Physical Geodesy*, 2nd edn. New York: Springer Wien.

Hwang, C., 1998. Inverse Vening Meinesz formula and deflection-geoid formula: applications to the predictions of gravity and geoid over the South China Sea. *Journal of Geodesy*, **72**, 304–312.

Molodensky, M. S., Eremeev, V. F., and Yurkina, M. I., 1962. *Methods for the Study of the External Gravity Field and Figure of the Earth*. Jerusalem: Israel Program of Scientific Translations (Russian original, 1960).

Moritz, H., 1980. *Advanced Physical Geodesy*. Herbert Wichmann Verlag: Karlsruhe (Reprint 2008 by School of Earth Sciences, Ohio State University, Columbus, Ohio).

Sideris, M. G., 1990. Rigorous gravimetric terrain modelling using Molodensky's operator. *Manuscripta Geodaetica*, **15**(2), 97–106.

Cross-references

Geodesy, Figure of the Earth
Geodesy, Physical
Geoid Undulation, Interpretation
Geoid, Computational Method
Gravimeters
Gravity Anomalies, Interpretation
Gravity Data, Advanced Processing
Gravity Field of the Earth
Gravity Field, Temporal Variations from Space Techniques
Gravity Field, Time Variations from Surface Measurements
Gravity Method, Satellite
Gravity, Data to Anomalies
Gravity, Global Models

GEOID UNDULATION, INTERPRETATION

Petr Vaníček
Department of Geodesy and Geomatics Engineering, University of New Brunswick, Fredericton, Canada

Synonyms

Geoidal height, aka geoid undulation; Direct gravimetric problem, aka forward modeling; Moho surface, aka compensation level for the Airy-Heiskanen isostatic hypothesis; Depth of Moho surface is the same as the thickness of the lithosphere; Schiaparelli's body, aka body of nil external potential; Stokes's integral truncated to the spherical distance of ψ_0, aka inner zone contribution to Stokes's integral; Spectral transfer function, aka admittance; Sea surface topography (SST), aka dynamic topography; Ocean bottom depth, aka bathymetry

Definition

Direct gravimetric problem is defined as follows: "Given the distribution of density within the Earth, derive the (external) gravity field the density generates."

The *inverse gravimetric problem* (GIP) reads: "Given the external gravity field of the Earth, derive the distribution of density implied by the field."

Introduction

The main goal of geophysics is to learn about the physical processes as well as properties of the material within the Earth. These are studied by means of measurements (data) collected on the surface of the Earth and above it and, to a very limited degree, also in very shallow probes inside the Earth. The discussion of the data to be collected and the techniques to be used to study the physical processes and material properties within the earth are what this book is all about.

The focus of this section is on gravity field, the geoid in particular, to the exclusion of all other, such as seismological, magnetic, electrical, heat flow data. The only attribute of the Earth that can be studied by means of gravity is the mass density variations as gravity is the acceleration caused by mass attraction. Three aspects of gravity field can be used for these studies: gravity acceleration, the deflection of the vertical, and gravity potential. Gravity acceleration, in the guise of gravity anomalies, are discussed in *Gravity Anomalies, Interpretation*, the

deflections of the vertical are introduced in *Geodesy, Physical* and for an in-depth view of the definition of geoid the reader is advised to consult *Geoid*.

One of the general properties of the gravity filed is that deep-seeded density anomalies have long wavelength signature in the gravity field while shallow density anomalies have short wavelength signature. Consequently, if one wishes to study shallow density features, one looks at the short wavelength part of gravity field and if one wishes to study deep-seeded features, one looks at the longer wavelength part. This is, of course, true only as a general principle as there exist also shallow features of large lateral extent that have long wavelength signatures.

When it comes to the interpretation of gravity field, the data of choice by exploration geophysicists, that is, people looking for ore, hydrocarbon, and other deposits, are gravity anomalies. The reason for this is that these data are bountiful on land, relatively cheap, and they contain all the short wavelength information. Observed deflections of the vertical as supplied by national geodetic offices are less popular as a source of information as they are not very accurate, and they are few and far apart. But we may live to see this situation change with the advent of Digital Zenith Cameras (Hirt et al., 2010).

Gravity potential data, such as the values of disturbing potential – divided by the value of gravity to convert them to the geoid – are much smoother than the first two kinds. This means that the short wavelength content is missing. We know that all three descriptors of gravity field we discuss here are functionally related: gravity acceleration at a point is a total gradient of the potential at that point, the deflection of the vertical at a point is a horizontal gradient of the equipotential surface passing through that point. Thus, there is nothing to stop us from using the potential data, the geoid, in such a way as to evaluate either the gravity anomalies or the deflections of the vertical and use these for the interpretation, is there?

That can be done of course, but this procedure cannot reconstitute the high frequency content as it has not been there to begin with. Thus, the conversion would not help and it is better to use the geoid as is; this is the thesis of the book (Vaníček and Christou, 1993). The direct use of the geoid has caught an interest of interpreters as the marine geoid data from satellite altimetry became quite reliable and freely available in the past 25 years.

Direct problem

Suppose we know the density distribution $\rho(\mathbf{r})$ within the Earth. Then we can determine the potential $V(\mathbf{r})$ outside the earth in two different ways. We can use the well-known Newton's integral to get the gravitational potential in a spatial form at any point external to the Earth as (Vaníček and Krakiwsky, 1986, Equation 20.48)

$$V(r,\Omega) = G \int_{\Omega_0} \int_{r=0}^{r_t(\Omega')} \frac{\rho(r',\Omega')}{|\mathbf{r}-\mathbf{r}'|}(r')^2 \mathrm{d}r'\mathrm{d}\Omega', \qquad (1)$$

where G is Newton's gravitational constant, r is the radial distance from the center of the Earth, Ω is the geocentric direction composed of colatitude θ and longitude λ, Ω_0 is the full geocentric solid angle, r_t is the distance from the center of the Earth to the topographical surface of the Earth.

Alternatively, the reciprocal distance $\frac{1}{|\mathbf{r}-\mathbf{r}'|}$ can be developed into a series of fully normalized spherical harmonic complex functions Y_{jm} to yield

$$V(r,\Omega) = \frac{GM}{a}\sum_{j=0}^{\infty}\left(\frac{a}{r}\right)^{j+1}\sum_{m=-j}^{j} A_{jm}Y_{jm}(\Omega), \qquad (2)$$

where M is the mass of the Earth, a is the major semiaxis of the best fitting Earth ellipsoid, and A_{jm} are the gravity potential coefficients. The potential coefficients can be evaluated from the known density $\rho(\mathbf{r}) = \rho(r, \Omega) = \rho(r, \theta, \lambda)$ within the Earth as

$$A_{jm} = \int_{\Omega_0}\int_{r=0}^{r_t(\Omega)} \rho(r,\Omega)G_{jm}(r,\Omega)\mathrm{d}r\mathrm{d}\Omega, \qquad (3)$$

where the integration kernel G_{jm} is given by Martinec in Vaníček and Christou (1993, Chap. 7, Equations 13 and 14)

$$G_{jm}(r,\Omega) = \frac{4\pi}{Ma^j(2j+1)}r^{j+2}Y_{jm}*(\Omega). \qquad (4)$$

Here * denotes a complex conjugate. This is the spherical harmonic form of potential V, a dual form to the spatial form (Equation 1 above) – these two forms can be transformed into each other, back and forth any time. The relation between potential V and the geoid undulation N is explained in *Geodesy, Physical*.

The direct approach, also known as forward modeling, can be used for solving some geophysical problems. Notably, it can be used to determine the depth of Mohorovicic's surface, the surface that divides the astenosphere (density about 3.2 g cm^{-3}) from the lithosphere (density about 2.67 g cm^{-3}). The Moho surface is also the compensation level for the Airy-Heiskanen isostatic hypothesis. The depth of the Moho surface, or the lithospheric (crustal) thickness derived this way, assuming the most realistic astenospheric density and described by a spherical harmonic series of a limited degree and order is shown on a web page http://geology.com/articles/mohorovicic-discontinuity.shtml. The displayed map was compiled by US Geological Survey in 1999.

Inverse problem

The determination of density $\rho(\mathbf{r})$ as a function of position $\mathbf{r}$ from any of the gravity field attributes, that is, gravity anomaly $\Delta g(\mathbf{r})$, the components of the deflection of the vertical $(\xi(\mathbf{r}),\eta(\mathbf{r}))$, or the geoidal height $N(\mathrm{r})$, is known as the inverse gravity problem (GIP). The non-uniqueness of GIP was known already to Schiaparelli, who in

1875–1876 was speaking of "gravitationally transparent bodies," specific distributions of density that do not generate any gravitational attraction outside the Earth. A complete mathematical description of a Schiaparelli's body was formulated by Vaníček and Kleusberg (1985) as follows: for arbitrary values of coefficients $c^{(i)}_{\ell jk}$ the following "body":

$$\theta \in <-\frac{\pi}{2},\frac{\pi}{2}>, \lambda \in <0, 2\pi>, r \in <0, 1>:$$
$$\tilde{\rho}(\mathbf{r}) = \sum_{i=0}^{\infty} \sum_{\substack{\ell=0 \\ \ell \neq j}}^{i} \sum_{j=0}^{\infty} \sum_{k=0}^{j} c^{(i)}_{\ell jk} \phi_{ijk}(\mathbf{r}), \tag{5}$$

where

$$\forall i,j = 0, 1, ..., k \leq j : \phi_{ijk}(\mathbf{r}) = J_i(3, 3, r)[Y^c_{jk}(\Omega), Y^s_{jk}(\Omega)], \tag{6}$$

J_i are normalized Jacobi polynomials (Abramowitz and Stegun, 1964) and Y^c_{jk}, Y^s_{jk} are the cosine and sine parts of spherical harmonic functions, has a nil external potential outside the ball of radius 1. The consequence of this behavior is that if one tries to interpret the external gravity potential in terms of internal density distribution, any distribution that conforms to Equation 5 can be randomly and legitimately added to the solution.

If the GIP does not have a unique solution, does it make any sense to even think about using gravity in studying the density distribution? True, it is not possible to take the gravity field outside the Earth and invert it uniquely into the density distribution within the Earth. But the external gravity field, particularly the geoid in our context here, can be used together with other data to get useful information on density.

Truncated geoid

There is a variety of techniques used in geophysical exploration to interpret gravity information obtained on the surface of the Earth in terms of anomalous density. One interesting approach uses the truncated geoid to determine the depth of a spherical, i.e., ball-like, density anomaly, where by "truncated geoid" we understand the following integral:

$$N^{\psi_0}(P) = \frac{R}{4\pi\gamma} \int_{C_{\psi_0}} \Delta g(Q) S(P, Q) \mathrm{d}\sigma, \tag{7}$$

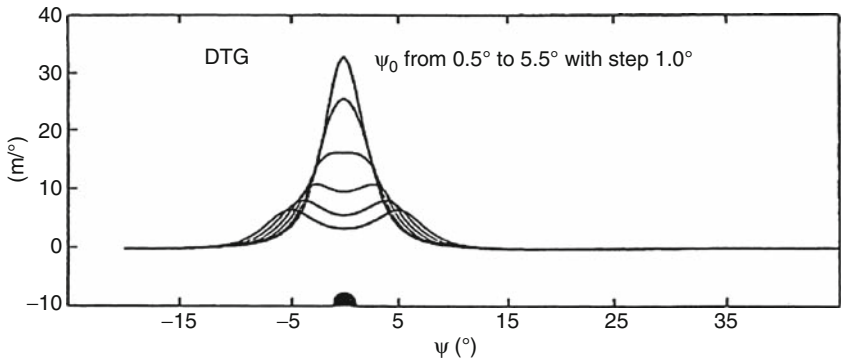
which represents the contribution of gravity to the geoid within the spherical cap C of a radius ψ_0. It can be viewed as the Stokes integral over the whole Earth, truncated to the spherical distance ψ_0, where P is the point of interest, Q is the dummy point in the integration, $S(P,Q) = S(\psi)$ is the Stokes spherical kernel and $\mathrm{d}\sigma$ is the surface element of the integration sphere of radius R.

It is the change of the truncated geoid with respect to the truncation parameter ψ_0,

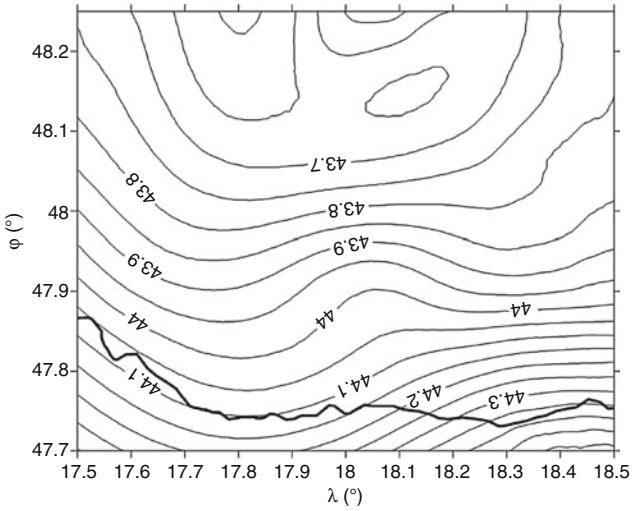
$$\frac{\mathrm{d}N^{\psi_0}}{\mathrm{d}\psi_0} = \lim_{\Delta\psi_0 \to 0} \frac{N^{\psi_0+\Delta\psi_0} - N^{\psi_0}}{\Delta\psi_0}, \tag{8}$$

that is of real interest. At the location of the density anomaly, the sequence of surfaces or profiles $\frac{\mathrm{d}N^{\psi_0}}{\mathrm{d}\psi_0}$ displays a dimple for a specific value of ψ_0 – see Figure 1 (from Vajda and Vaníček, 1999). This specific value of ψ_0 is almost linearly related to the depth of the anomaly.

In Figure 2, geoid from an area in Southern Slovakia is shown. Just northeast of the geographical center of the area, there is a local geoid low, known as "Kolarovo anomaly," that has been intriguing geophysicists for some time. It has been postulated by several authors that the center of mass of the density anomaly responsible for



Geoid Undulation, Interpretation, Figure 1 Sequence of profiles of the change of the truncated geoid with respect to the truncation parameter ψ_0 that shows the onset of the dimple for a value of $\psi_0 = 2.5°$.



Geoid Undulation, Interpretation, Figure 2 Geoid in the area of "Kolarovo anomaly" in Southern Slovakia (Courtesy of J. Janak, Slovak University of Technology).

the shown pattern is about 9.5 km bellow the Earth surface. The application of the truncated geoid technique shows an estimate of 8.7 km (Vajda et al., 2002).

Some other applications

In addition to the above-discussed applications, geoid can be used in many other ways; one of those applications is studying tectonic forces. Interested reader may read about the involved concepts in *Gravity, Global Models*. Bowin, in his earlier studies (1991), postulated that the 4–10° geoid undulations are consistent with the present zones of subducted tectonic slabs. His postulate has more recently been confirmed by Ricard et al. (2006). These authors also agree with Bowin's speculation (1983) that only the deepest Earth's mass anomalies cause geoid undulations of 2–10°.

Casenave in Vaníček and Christou (1993, Ch. 13) discusses the use of the geoid in studying different isostatic compensation models using the spectral transfer function, aka admittance in the "theory of systems" parlance. She studies the mechanical behavior of oceanic tectonic plates (flexure under seamount loading) and thermal evolution of oceanic lithosphere. Ricard et al. (2006) show that the elasticity of the lithosphere affects the geoid between 200° and 250°. Casenave also argues that the signatures of oceanic hot spots in the geoid are in the medium wavelength of $n \in (5, 20)$ and in 1–10 m in range.

It may be worth mentioning that meteorite craters are often clearly discernible in detailed geoid maps. The Manicouagan crater in Eastern Canada (location: N 51° 23′ and W 68° 42′) makes a very clear imprint on the geoidal map.

The last applications of the geoid we want to show here concern physical oceanography and the oceans in general. A very good overview of this topic by Nerem and Koblinsky is in Vaníček and Christou (1993, Chap. 16). Geoid, being an equipotential (level) surface, should be represented by the mean sea level if seawater were homogeneous, which, of course, it is not. As a result, the mean sea level is not in equilibrium and currents develop. Some time after the initiation of this motion, these currents reach a steady state, called geostrophic, and swirl around the globe in a perpetual motion attenuated only by friction.

The equation that describes the geostrophic motion is ascribed to Navier and Stokes, and it reads as follows:

$$\boldsymbol{V} + (\boldsymbol{V} \bullet \nabla)\boldsymbol{V} + 2\boldsymbol{\Omega} \times \boldsymbol{V} = \frac{-\nabla p}{\rho} + \boldsymbol{g} + \boldsymbol{F}, \qquad (9)$$

where $\boldsymbol{V}$ is the flow velocity vector, $\boldsymbol{\Omega}$ is the rotation vector of the Earth, p is the hydrostatic pressure, ρ is the seawater density, $\boldsymbol{g}$ is the gravitation vector, and $\boldsymbol{F}$ is the vector representing frictional forces (*ibid.* Equation 1). This equation, together with measurements taken at sea, is then used to derive the height of the sea surface above the geoid, which is called the "dynamic topography."

As discussed above, the instantaneous sea surface has been monitored by a technique called satellite altimetry for years. This surface represents, of course, the sum of the geoidal height and the dynamic topography. Subtracting the geoid from the measured sea surface yields the dynamic topography that can be easily converted into information about currents. Conversely, subtracting the dynamic topography from the measured surface yields the geoid. Interested reader can refer to following URL http://www.aviso.oceanobs.com/en/data/products/auxiliary-products/index.html for more detailed information.

Last but not least, the geoid obtained from satellite altimetry can be used to calculate the ocean bottom depth, aka bathymetry (Sandwell et al., 2006). The principle used in these calculations is the same as the one used in the determination of the Moho surface. Acknowledging that the ocean bottom is a surface that divides the crust (of an average density of 2.67 g cm^{-3}) from the layer of water (of an average density of 1.03 g cm^{-3}), forward modeling it then used to yield the ocean bottom surface.

Summary

The geoidal height (geoid undulation) is a quantity that can be quite readily interpreted in terms of density distribution within the Earth. The main application is in studying the processes that take place deep within the Earth, but even shallower density anomalies have an impact on a geoid map. Nowadays, when sea surface height

measured by satellite altimetry is freely available, marine geoid has become a particularly handy tool to use.

Bibliography

Abramowitz, M., and Stegun, I. A., 1964. *Handbook of Mathematical Functions*. St. Mineola, NY: Dover reprint.

Bowin, C., 1983. Depth of principal mass anomalies contributing to the Earth's geoidal undulations and gravity anomalies. *Marine Geodesy*, **7**, 61.

Bowin, C., 1991. Speculations on a mechanism for maintenance of absolute plate motion, AGU abstract, *Eos, Transactions, American Geophysical Union*, **72**(44), 244.

Hirt, C., Burki, B., Somieski, A., and Seeber, G., 2010. Modern determination of vertical deflections using digital zenith cameras. *Journal of Surveying Engineering*, **136**(1), 1–12.

Ricard, Y., Chambat, F., and Lithgow-Bertelloni, C., 2006. Gravity observations and 3D structure of the Earth. *Comptes Rendus Geosciences*, **338**, 992–1001.

Sandwell, D. T., Smith, W. H. F., Gille, S., Kappel, E., Jayne, S., Soofi, K., Coakley, B., and Geli, L., 2006. Bathymetry from space: rationale and requirements for a new, high-resolution altimetric mission. *Comptes Rendus Geosciences*, **338**, 1049–1062.

Schiaparelli, G. V., 1875–1876. *Lezzioni di Meccanica Celeste (Appunti)*. Italy: University of Pavia.

Vajda, P., and Vaníček, P., 1999. Truncated geoid and gravity inversion for one point-mass anomaly. *Journal of Geodesy*, **73**, 58–66.

Vajda, P., Bielik, M., and Pohanka, V., 2002. Testing the application of the truncation filtering methodology in interpreting real gravity data: the Kolárovo gravity anomaly, *Contributions to Geophysics and Geodesy*, **32**(1), 57–66.

Vaníček, P., and Kleusberg, A., 1985. What an external gravitational potential can really tell us about mass distribution. *Bollettino di Geofisica Teorica ed Applicata*, Vol. XXCII, No. 108, December, pp. 243–250.

Vaníček, P., and Krakiwsky, E. J., 1986. *Geodesy: The Concepts*, 2nd edn. Amsterdam: North Holland.

Vaníček, P., and Christou, N. T., 1993. *Geoid and its Geophysical Interpretations*. Boca Raton: CRC Press.

Cross-references

Earth, Density Distribution
Geodesy, Physical
Geoid
Geoid Determination, Theory and Principles
Gravity Anomalies, Interpretation
Gravity Field of the Earth

GEOID, COMPUTATIONAL METHOD

Michael G. Sideris
Department of Geomatics Engineering, University of Calgary, Calgary, AB, Canada

Introduction

In this entry, we are presenting numerical methods for estimating geoid undulations and other relevant quantities, based on the theory presented in the article on *Geoid Determination, Theory and Principles*. We will refer to formulas in that article by the equation number followed by Sideris (see *Geoid Determination, Theory and Principles*). For example, Stokes's equation will be referred to as Equation 17 of the article *Geoid Determination, Theory and Principles*.

The remove-compute-restore technique

In practice, we can evaluate Stokes's equation by numerical integration using only discrete data; therefore, short-wavelength information will not be properly represented in the solution (it will be aliased). In addition, we use data only in a certain region, and therefore the long-wavelength contributions of the gravity field will not be present in the results and must be computed in another way. These problems can be remedied by using a global geopotential model (GM) and heights, *H,* in a digital elevation model (DEM) to obtain the long and short wavelengths of the field, respectively. This leads to the standard *remove-compute-restore technique*, in which the long and short wavelengths are first removed from the gravity anomalies in a preprocessing step, then the residual anomalies Δg_r are used to compute residual undulations $N_{\Delta gr}$, and then the long- and short-wavelength contributions are restored in a post-processing step. The computation is based on the following formula:

$$N = N_{\mathrm{GM}} + N_{\Delta g_r} + N_{\mathrm{H}}, \quad \Delta g_r = \Delta g - \Delta g_{\mathrm{GM}} - \Delta g_{\mathrm{H}} \tag{1}$$

where Δg_{H} denotes the terrain contributions and N_{H} the corresponding indirect effect on the geoid; see Figure 1. It should also be noted that such residual anomalies are typically smaller and smoother, and therefore appropriate for use in interpolation and in the creation of gridded values by, e.g., least-squares collocation.

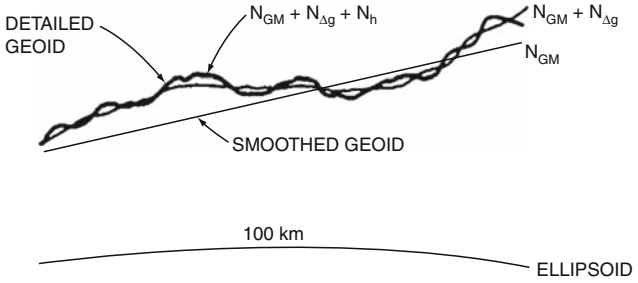
Computation of the GM-contributions

The geopotential model parts of Δg and N are usually computed by the spherical approximation of Equations 15 and 16 in Sideris, *Geoid Determination, Theory and Principles* from the coefficients of a GM of maximum degree $n_{\max}$, which is a function of the resolution of the data that were used to compute the coefficients C_{nm}, S_{nm}:

$$N_{\mathrm{GM_P}} = R \sum_{n=2}^{n_{\max}} \sum_{m=0}^{n} [C_{nm} \cos(m\lambda_{\mathrm{P}}) + S_{nm} \sin(m\lambda_{\mathrm{P}})] P_{nm}(\sin \phi_{\mathrm{P}}) \tag{2}$$

$$\Delta g_{\mathrm{GM_P}} = G \sum_{n=2}^{n_{\max}} (n-1) \sum_{m=0}^{n} [C_{nm} \cos(m\lambda_{\mathrm{P}}) + S_{nm} \sin(m\lambda_{\mathrm{P}})] P_{nm}(\sin \phi_{\mathrm{P}}) \tag{3}$$

G and R are the mean gravity and mean radius of the sphere, and ϕ_{P}, λ_{P} are the spherical latitude and longitude



Geoid, Computational Method, Figure 1 Contributions of different data to regional geoid determination.

of the computation point. This representation causes an omission or truncation error, but there is also a commission error because the data used to estimate the coefficients, and therefore also the coefficients themselves, do contain errors. Naturally, the spherical approximation itself causes an additional error, and therefore Equations 15 and 16 in Sideris, *Geoid Determination, Theory and Principles* are preferable to Equations 2 and 3.

A comprehensive discussion on the determination and proper use of GM coefficients can be found in *Spherical Harmonic Analysis Applied to Potential Fields*. Information on the high-resolution EGM08 model is given in Pavlis et al. (2008).

Computation of the Δ*g*-contribution

The contribution of gravity anomalies can be computed by use of either the Stokes integral or least-squares collocation (LSC). For simplicity in the notation, we will often write just N and Δg instead of $N_{\Delta g_r}$ and Δg_r in the equations, respectively.

Evaluation of Stokes's integral. The Stokes integral can be evaluated either on the sphere or on the plane by numerical integration using data on circular sectors (ring integration method) or on regular grids. Since gravity data are typically available on grids, and because Stokes's integral is a convolution integral on the plane and along parallels on the sphere, we will discuss here the integral's evaluation on regular grids by use of the very efficient *fast Fourier transform* (FFT) *technique* (Bracewell, 1978).

On the sphere, using Equations 17–19 in Sideris, *Geoid Determination, Theory and Principles*, we can write Stokes's integral explicitly as

$$N_P = \frac{R}{4\pi\gamma} \iint_{\sigma} \Delta g S(\phi_p, \phi, \lambda - \lambda_p) \cos\phi \, d\phi d\lambda \tag{4}$$

which is a convolution between $\Delta g \cos\phi$ and the Stokes kernel S only along parallels. For Δg_r given on $M \times N$ points on a spherical grid with spacings $\Delta\phi$, $\Delta\lambda$, Equation 4 can be written for every parallel of constant latitude ϕ_l as

$$\begin{aligned} N_{\Delta g_r}(\phi_l, \lambda_k) &= \frac{R}{4\pi\gamma} \sum_{j=0}^{N-1} \left[\sum_{i=0}^{M-1} \Delta g_r(\phi_j, \lambda_i) \cos\varphi_j \right. \\ &\quad \left. S(\phi_l, \phi_j, \lambda_k - \lambda_i) \Delta\lambda \right] \Delta\phi \\ &= \frac{R}{4\pi\gamma} \sum_{j=0}^{N-1} [(\Delta g_r(\phi_j, \lambda_k) \cos\phi_j) * \\ &\quad S(\phi_l, \phi_j, \lambda_k)] \Delta\phi, \\ &\quad \phi_l = \phi_1, \phi_2, \ldots, \phi_N \end{aligned} \tag{5}$$

The brackets in Equation 5 contain a one-dimensional (1D) discrete convolution with respect to l, i.e., along a parallel. Therefore, the above discrete convolution, denoted by $*$, can be evaluated by the 1D FFT (Starng Van Hees, 1990; Haagmans et al., 1993) as follows:

$$\begin{aligned} N_{\Delta g_r}(\phi_l, \lambda_k) = \frac{R}{4\pi\gamma} F^{-1} \left\{ \sum_{j=0}^{N-1} F\left\{ \Delta g_r(\phi_j, \lambda_k) \cos\phi_j \right\} \right. \\ \left. F\left\{ S(\phi_l, \phi_j, \lambda_k \right\} \right\}, \; \phi_l = \phi_1, \phi_2, \ldots, \phi_N \end{aligned} \tag{6}$$

where F denotes the direct Fourier transform and F^{-1} the inverse Fourier transform (Bracewell, 1978). Although Equation 6 is rigorous, its evaluation is not as efficient as it could be, as it still requires numerical summation along meridians. By employing various approximations for the Stokes kernel function, such as, e.g., $\cos\phi_P \cos\phi = \cos^2\bar{\phi}$, where $\bar{\phi}$ is the mean latitude of the computation area, it is possible to transform Equation 6 into a two-dimensional (2D) convolution. In this case, accuracy is traded for efficiency. Detailed formulas can be found in Forsberg and Sideris (1993).

In the planar approximation, which can be employed when the data is available in a limited area E with extent of no more than several degrees in each direction, the Stokes kernel can be approximated by its first term $1/\sin(\psi/2)$. Then the kernel and the differential areal can be converted to their planar equivalents, i.e., $S(\psi) \approx 1/\sin(\psi/2) \approx 2/\psi \approx 2R/l$, and $R d\sigma = dxdy$, yielding the following *planar form of Stokes's integral* (Schwarz et al., 1990):

$$N_P = \frac{1}{2\pi\gamma} \iint_E \frac{\Delta g}{l} dxdy = \frac{1}{2\pi\gamma} \Delta g_P * l_{N_P} \tag{7}$$

where l is the planar distance between data and computation point and l_N is the planar form of Stokes's kernel:

$$l_{\mathrm{N}} = \left(x^2 + y^2\right)^{-1/2} \tag{8}$$

On a grid with spacings Δx, Δy, the above 2D convolution integral takes the discrete form

$$\begin{aligned} N_{\Delta g_{\mathrm{r}}}(x_k, y_l) &= \frac{1}{2\pi\gamma} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \Delta g_{\mathrm{r}}(x_i, y_j) \\ &\quad l_N(x_k - x_i, y_l - y_j) \Delta x \Delta y \\ &= \frac{1}{2\pi\gamma} \Delta g_{\mathrm{r}}(x_k, y_l) * \; l_N(x_k, y_l) \end{aligned} \tag{9}$$

$$\begin{aligned} & l_{\mathrm{N}}(x_k - x_i, y_l - y_j) \\ &= \begin{cases} [(x_k - x_i)^2 + (y_l - y_j)^2]^{-1/2}, & x_k \neq x_i \;\; or \;\; y_l \neq y_j \\ 0, & x_k = x_i \;\; and \;\; y_l = y_j \end{cases} \end{aligned} \tag{10}$$

Geoid undulations can then be evaluated by the 2D FFT as follows:

$$N_{\Delta g_{\mathrm{r}}}(x_k, y_l) = \frac{1}{2\pi\gamma} F^{-1}\left\{ F\{\Delta g_{\mathrm{r}}(x_k, y_l)\} F\{l_{\mathrm{N}}(x_k, y_l)\} \right\} \tag{11a}$$

or, in more compact form (which we will use for all convolutions from this point on),

$$N_{\mathrm{P}} = \frac{1}{2\pi\gamma} F^{-1}\left\{ F\{\Delta g_{\mathrm{P}}\} F\{l_{N_{\mathrm{P}}}\} \right\} \tag{11b}$$

Because $S(\psi)$ and l_{N} are singular at the origin, the contribution of the gravity anomaly of the computation point must be evaluated separately and added to the results of the convolutions. On the sphere, Hofmann-Wellenhof and Moritz (2006) give this contribution approximately as

$$\mathrm{d}N_{\Delta g_{\mathrm{r}}}(\varphi_{\mathrm{P}}, \lambda_{\mathrm{P}}) \approx \gamma^{-1} s_{\mathrm{o}} \Delta g_{\mathrm{r}}(\varphi_{\mathrm{P}}, \lambda_{\mathrm{P}}) \tag{12}$$

where s_{o} denotes the radius of a small area around the computation point in which Δg_{r} can be considered constant. On the plane, the corresponding equation would be

$$\mathrm{d}N_{\Delta g_{\mathrm{r}}}(x_{\mathrm{P}}, y_{\mathrm{P}}) \approx \gamma^{-1} \sqrt{\frac{\Delta x \Delta y}{\pi}} \Delta g_{\mathrm{r}}(x_{\mathrm{P}}, y_{\mathrm{P}}) \tag{13}$$

Equation 11 employs the numerical evaluation of the spectrum of Stokes's kernel and is the one that should be used in practice (with proper zero padding of the data) if we want the results from FFT to be identical to those from numerical integration; see Li and Sideris (1993). This is in fact true for any of the other convolution integrals that we are discussing here. Nevertheless, it is well known from Fourier theory that the inverse distance function l_{N} has an analytical spectrum, which is the inverse frequency q^{-1} in the frequency domain. Using u,v,w to denote the frequencies corresponding to x,y,z, this spectrum is

$$\begin{aligned} F\{l_{\mathrm{N}}(x, y)\} &= F\left\{ \left(x^2 + y^2\right)^{-1/2} \right\} \\ &= \left(u^2 + v^2\right)^{-1/2} = q^{-1} \end{aligned} \tag{14}$$

and therefore Equation 7 can be evaluated as follows:

$$N_{\mathrm{P}} = \frac{1}{2\pi\gamma} F^{-1}\left\{ F\{\Delta g_{\mathrm{P}}\} q^{-1} \right\} \tag{15}$$

This representation clearly shows that Stokes's integral is a smoothing, mainly low-frequency-dependent operator, and therefore geoid undulations are more sensitive to the long wavelengths of the gravity field.

The FFT method can use either mean or point data on a grid and computes simultaneously the geoidal heights at all grid points; see Sideris and Tziavos (1988) for the formulas with mean data. A large area can thus be covered very fast and with minimum cost. An interpolation technique such as LSC must be employed for predicting geoidal heights at points not on the grid and/or to grid irregular gravity anomalies. With gridded data, the FFT technique is by far the fastest of all available methods. Criticism that only the FFT suffers from leakage and aliasing effects are not valid since these effects depend on data density and coverage and thus affect all methods.

Given gridded gravity and topography data, the Molodensky solution of Equations 20–22 in Sideris, *Geoid Determination, Theory and Principles* can also be efficiently evaluated by FFT. Details can be found in Sideris (1987) and Schwarz et al. (1990).

Computation by least-squares collocation. A solution can also be obtained using least-squares collocation (LSC). For example, in the simple case that we only want to predict residual geoid undulations from residual gravity anomalies, the Equations 26 and 27 in Sideris, *Geoid Determination, Theory and Principles* reduce to

$$\hat{\mathbf{N}}_{\Delta g} = \mathbf{C}_{\mathrm{N}\Delta g}\left(\mathbf{C}_{\Delta g \Delta g} + \mathbf{C}_{nn}\right)^{-1} \mathbf{\Delta g} \tag{16}$$

$$\mathbf{C}_{ee} = \mathbf{C}_{NN} - \mathbf{C}_{N\Delta g}\left(\mathbf{C}_{\Delta g \Delta g} + \mathbf{C}_{nn}\right)^{-1} \mathbf{C}_{\Delta g N} \tag{17}$$

Given the covariance matrix for T, $\mathbf{C}_{\mathrm{TT}}$, the other covariance matrices can be derived by covariance propagation using Equations 12 and 13 in Sideris, *Geoid Determination, Theory and Principles* as follows:

$$\begin{aligned} \mathbf{C}_{NN} &= \frac{1}{\gamma^2} \mathbf{C}_{TT}, \\ \mathbf{C}_{N\Delta g} &= \frac{1}{\gamma}\left(-\frac{\partial \mathbf{C}_{TT}}{\partial h} + \frac{1}{\gamma}\frac{\partial \gamma}{\partial h} \mathbf{C}_{TT} \right), \\ \mathbf{C}_{\Delta g \Delta g} &= \left(-\frac{\partial}{\partial h} + \frac{1}{\gamma}\frac{\partial \gamma}{\partial h} \right)\left(-\frac{\partial \mathbf{C}_{TT}}{\partial h} + \frac{1}{\gamma}\frac{\partial \gamma}{\partial h} \mathbf{C}_{TT} \right) \end{aligned} \tag{18}$$

LSC's drawback is the need for inverting large covariance matrices or simply solving large systems of linear equations, which in practice often prevents the utilization of all given data, and thus prevents the application of LSC over large regions. To circumvent this problem, if gridded data are available, LSC can be applied in the frequency domain. This is actually related directly to the use of FFT formulas with noisy, heterogeneous data. It will in fact be shown below that the FFT method can use heterogeneous data, provided that they are given on a grid, and can produce error estimates, provided that the power spectral density (PSD) functions, which are the Fourier transform of the covariance functions, of the data and their noise are known. In this case, the technique is equivalent to frequency-domain collocation.

Computation by frequency-domain collocation. To illustrate how error propagation can be used with Stokes's integral, assume that the anomalies have errors n with known PSD P_{nn}. In this case, Stokes's convolution of, e.g., Equation 7, becomes

$$N = (\Delta g + n) * s + e \tag{19}$$

where $s = l_N/(2\pi\gamma)$ and e is the error of the estimated undulations.

Making use of the definition of the PSD function P_{xy} for two signals x and y as

$$P_{xy} = F\{x\}F\{y\}^* \tag{20}$$

where the superscript $*$ denotes complex conjugate, and assuming no correlation between the signal and noise, the PSD of N and the cross-PSD of Δg and N can be derived as follows:

$$P_{NN} = P_{ee} + S(P_{\Delta g\Delta g} + P_{nn})S^* \tag{21}$$

$$P_{\Delta gN} = P_{\mathrm{ee}} + S(P_{\Delta g\Delta g} + P_{nn}) \tag{22}$$

where $S = F\{s\}$ and $P_{\Delta g\Delta g}$ is the PSD of the gravity anomalies. From the above equations, the spectrum of Stokes's kernel and the output-noise PSD can be obtained as

$$S = \frac{P_{N\Delta g}}{P_{\Delta g\Delta g} + P_{nn}} \tag{23}$$

$$P_{\mathrm{ee}} = P_{NN} - P_{N\Delta g}(P_{\Delta g\Delta g} + P_{nn})^{-1}P_{N\Delta g} \tag{24}$$

Now Equation 19 can be evaluated in the frequency domain as follows:

$$F\{N\} = P_{N\Delta g}(P_{\Delta g\Delta g} + P_{nn})^{-1}F\{\Delta g\} \tag{25}$$

One can see immediately that Equations 24 and 25 are, formally, the frequency-domain equivalents of the LSC Equations 16 and 17. In fact, they can also be obtained starting from the general form of the equation for P_{ee} and estimating that $S = F\{s\}$ which makes P_{ee} minimum; such a derivation can be found in Sideris (1996). What makes them much more efficient, though, is that they require no matrix inversion as the -1 superscript in Equations 24 and 25 implies straightforward division at each frequency. Notice also that for noiseless data ($P_{nn} = 0$), Equation 24 yields to $P_{\mathrm{ee}} = 0$, and Equation 23 becomes $S = P_{N\Delta g} / P_{\Delta g\Delta g} = F\{N\}/F\{\Delta g\} = (2\pi\gamma q)^{-1}$. Note, however, that for the PSDs to be the Fourier transform of the covariance functions (CVs) used in collocation, these CVs have to be stationary, which is not the case in practice with the noise CVs. Thus, the FFT method, although it is much more efficient than LSC, has to approximate nonstationary noise covariance functions (which are easily handled by LSC) by stationary ones; for details, see Sansò and Sideris (1997), Kotsakis and Sideris (2001).

Computation of the terrain contributions

The terrain effect on Δg and the indirect effect on N are computed from approximations to Equations 32 and 33 in Sideris, *Geoid Determination, Theory and Principles*. Keeping only the terms for $r = 1$, these equations take the following form:

$$\begin{aligned} -\Delta g_{H_P} = c_P &= \pi k \mathbf{L}(\rho H - \rho H_P)^2 \\ &= \pi k\left[\mathbf{L}(\rho H^2) - 2H_P\mathbf{L}(\rho H)\right] \end{aligned} \tag{26}$$

$$\delta N_{H_P} = -\frac{\pi k\rho}{\gamma}H_P^2 - \frac{\pi k}{3\gamma}\mathbf{L}(\rho H^3) \tag{27}$$

With height data given in small area E around the computation point P, the planar approximation of the $\mathbf{L}$ operator can be used. In this case, the above equation becomes (Sideris, 1990)

$$\begin{aligned} c_P &= \frac{1}{2}k\iint_E \frac{\rho H^2 - \rho H_P^2}{l^3}\mathrm{d}x\mathrm{d}y \\ &\quad - H_P k\iint_E \frac{\rho H - \rho H_P}{l^3}\mathrm{d}x\mathrm{d}y \\ &= \frac{1}{2}k\left[(\rho_P H_P^2) * l_P^{-3} - H_P^2(\rho_P * l_P^{-3})\right] \\ &\quad - H_P k[(\rho_P H_P) * l_P^{-3} - H_P(\rho_P * l_P^{-3})] \end{aligned} \tag{28}$$

$$\begin{aligned} \delta N_{H_P} &= -\frac{\pi k\rho}{\gamma}H_P^2 - \frac{k}{6\gamma}\iint_E \frac{\rho H^3 - \rho H_P^3}{l^3}\mathrm{d}x\mathrm{d}y \\ &= -\frac{\pi k\rho}{\gamma}H_P^2 - \frac{k}{6\gamma}[(\rho_P H_P^3) * l_P^{-3} - H_P^3(\rho_P * l_P^{-3})] \end{aligned} \tag{29}$$

where $l_P^{-3} = (x_P^2 + y_P^2)^{-3/2}$. The above convolutions can be evaluated by FFT as follows:

$$\begin{aligned} c_P = &\frac{1}{2}k[F^{-1}\{F\{\rho_P H_P^2\}F\{l_P^{-3}\}\} \\ &-H_P^2 F^{-1}\{F\{\rho_P\}F\{l_P^{-3}\}\}] \\ &-H_P k[F^{-1}\{F\{\rho_P H_P\}F\{l_P^{-3}\}\} \\ &-H_P F^{-1}\{F\{\rho_P\}F\{l_P^{-3}\}] \end{aligned} \tag{30}$$

$$\begin{aligned} \delta N_{H_P} = &-\frac{\pi k\rho}{\gamma}H_P^2 - \frac{k}{6\gamma}\left[F^{-1}\left\{F\{\rho_P H_P^3\}F\{l_P^{-3}\}\right\}\right. \\ &\left.-H_P^3 F^{-1}\left\{F\{\rho_P\}F\{l_P^{-3}\}\right\}\right] \end{aligned} \tag{31}$$

From the definition of the vertical derivative operator **L** in Equation 21 in Sideris, *Geoid Determination, Theory and Principles* and the derivative property of the Fourier transform (Bracewell, 1978), we can see that

$$F\{\mathbf{L}f\} = F\left\{\frac{\partial f}{\partial z}\right\} = 2\pi i w F\{f\} = -2\pi q F\{f\} \tag{32}$$

where i is the imaginary unit. We have also used the relationship $iw = -q$, which is a straightforward result of Laplace's equation in the frequency domain:

$$\begin{aligned} F\{\nabla^2 f\} = &[(i2\pi u)^2 + (i2\pi v)^2 \\ &+ (i2\pi v)^2]F\{f\} = 0 \end{aligned} \tag{33}$$

Now Equations 26 and 27 can also be evaluated by FFT using the analytical spectrum of the **L** operator as follows:

$$\begin{aligned} c_P = &\pi k(F^{-1}\{-2\pi q F\{\rho_P H_P^2\}\} \\ &-2H_P F^{-1}\{-2\pi q F\{\rho_P H_P\}\}) \end{aligned} \tag{34}$$

$$\delta N_{H_P} = -\frac{\pi k}{\gamma}\rho_P H_P^2 - \frac{\pi k}{3\gamma}F^{-1}\{-2\pi q F\{\rho_P H_P^3\}\} \tag{35}$$

Such analytical representations of kernel functions, though not recommended for use in limited areas (Li and Sideris, 1993), are often used in satellite altimetry applications, as discussed in the section that follows.

Evaluation of satellite altimetry contributions

Like Stokes's formula, the inverse Stokes, inverse Vening Meinesz and deflection-geoid formulas employ kernels that are functions of the spherical distance ψ between the data and computation points; see Equation 19 in Sideris, *Geoid Determination, Theory and Principles*. Therefore, when gridded data is available, they can be efficiently evaluated on the sphere by either the rigorous 1D FFT or the approximate 2D FFT formulas, completely analogous to Equations 4–9 used for Stokes's integral. They will not be discussed further here, and the interested reader is referred to Hwang (1998) for detailed equations. Instead, we will concentrate here on the planar approximations with analytical kernel spectra.

On the plane, employing the same approximation for the sine and cosine of ψ as we did for Stokes's formula, Equations 36–40 in Sideris, *Geoid Determination, Theory and Principles* simplify to

$$\Delta g_P = -\frac{\gamma}{2\pi}\iint_E \frac{N - N_P}{l^3}\mathrm{d}x\mathrm{d}y = -\gamma \mathbf{L}N = -\gamma\frac{\partial N}{\partial z} \tag{36}$$

$$\begin{aligned} \Delta g_P &= -\frac{\gamma}{2\pi}\iint_E \left(\xi\frac{y - y_P}{l^3} + \eta\frac{x - x_P}{l^3}\right)\mathrm{d}x\mathrm{d}y \\ &= -\frac{\gamma}{2\pi}\left(\xi_P * l_{\xi_P} + \eta_P * l_{\eta_P}\right) \end{aligned} \tag{37}$$

$$\begin{aligned} l_\xi &= y/(x^2 + y^2)^{3/2} = -\partial l_N/\partial y \\ l_\eta &= x/(x^2 + y^2)^{3/2} = -\partial l_N/\partial x \end{aligned} \tag{38}$$

$$\begin{aligned} N_P &= -\frac{1}{2\pi}\iint_E (\xi\frac{y - y_P}{l^2} + \eta\frac{x - x_P}{l^2})\mathrm{d}x\mathrm{d}x \\ &= -\frac{1}{2\pi}(\xi_P * l_{x_P} + \eta_P * l_{y_P}) \end{aligned} \tag{39}$$

$$\begin{aligned} l_y &= y/(x^2 + y^2) \\ l_x &= x/(x^2 + y^2) \end{aligned} \tag{40}$$

By making use of the derivative property of the Fourier transform, their evaluation by FFT is done as follows:

$$\Delta g_P = -\gamma F^{-1}\left\{2\pi q F\{N_P\}\right\} \tag{41}$$

$$\Delta g_P = \gamma F^{-1}\left\{\frac{i}{q}\left(vF\{\xi_P\} + uF\{\eta_P\}\right)\right\} \tag{42}$$

$$N_P = F^{-1}\left\{\frac{i}{2\pi q^2}\left(vF\{\xi_P\} + uF\{\eta_P\}\right)\right\} \tag{43}$$

The planar form of Equation 42 in Sideris, *Geoid Determination, Theory and Principles*, namely

$$\frac{\partial \Delta g}{\partial z} = -\gamma\left(\frac{\partial \xi}{\partial y} + \frac{\partial \eta}{\partial x}\right) \tag{44}$$

also reduces to Equation 42 in the frequency domain, unless one is interested not in Δg but in its vertical derivative, which then can be computed from the expression

$$\left.\frac{\partial \Delta g}{\partial \xi}\right|_P = 2\pi\gamma F^{-1}\{ivF\{\xi_P\} + iuF\{\eta_P\}\} \tag{45}$$

Finally, Equation 35 in Sideris, *Geoid Determination, Theory and Principles* that gives the deflections from altimetry-derived undulations is in planar approximation

$$\xi = -\frac{\partial N}{\partial y}, \quad \eta = -\frac{\partial N}{\partial x} \tag{46}$$

and therefore its FFT-evaluation can be done via the expression

$$\begin{aligned} \xi_P &= -2\pi F^{-1}\{ivF\{N_P\}\}, \\ \eta_P &= -2\pi F^{-1}\{iuF\{N_P\}\} \end{aligned} \tag{47}$$

Summary

In this entry, we have outlined the steps in the remove-compute-restore technique that is employed in the numerical computations of geoid undulations. We have shown that most of the equations expressing the terrain and gravity effects on the geoid are convolution integrals, and have described in detail how they can be efficiently evaluated by FFT given gridded data on the sphere or on the plane. In addition, computations by least-squares collocation have been outlined both in the space and in the frequency domain, in which the formula is equivalent to Wiener filtering and provides both noise filtering and accuracy estimates of the geoid through propagation of the data noise to the results. The entry closes with the evaluation of various formulas used in gravity and geoid determination in the oceans using data from satellite altimetry.

Bibliography

Bracewell, R., 1978. *The Fourier Transform and Its Application*, 2nd edn. New York: McGraw-Hill, p. 444, revised.

Forsberg, R., and Sideris, M. G., 1993. Geoid computations by the multi-banding spherical FFT approach. *Manuscripta Geodaetica*, **18**, 82–90.

Haagmans, R., de Min, E., and van Gelderen, M., 1993. Fast evaluation of convolution integrals on the sphere using ID FFT, and a comparison with existing methods for Stokes' integral. *Manuscripta Geodaetica*, **18**, 227–241.

Hofmann-Wellenhof, B., and Moritz, H., 2006. *Physical Geodesy*, 2nd edn. New York: Springer Wien.

Hwang, C., 1998. Inverse Vening Meinesz formula and deflection-geoid formula: applications to the predictions of gravity and geoid over the South China Sea. *Journal of Geodesy*, **72**, 304–312.

Jekeli, C., 2010. Gravity field of the Earth. In *Encyclopedia of Solid Earth Geophysics*. New York: Springer, this Vol., Chap. 19.

Kotsakis, C., and Sideris, M. G., 2001. A modified Wiener-type filter for geodetic estimation problems with non-stationary noise. *Journal of Geodesy*, **75**(12), 647–660.

Li, Y. C., and Sideris, M. G., 1993. Gravity field convolutions without windowing and edge effects. *Manuscripta Geodaetica*, **67**, 107–118.

Moritz, H., 1980. *Advanced Physical Geodesy*. Karlsruhe: Wichmann (reprinted 2008; Columbus, OH: School of Earth Sciences, Ohio State University).

Pavlis, N. K., Holmes, S. A., Kenyon, S. C., and Factor, J. K., 2008. An Earth gravitational model to degree 2160: EGM2008. Paper presented at the 2008 General Assembly of the European Geosciences Union, Vienna, Austria, April 13–18.

Sansò, F., and Sideris, M. G., 1997. On the similarities and differences between systems theory and least-squares collocation in physical geodesy. *Bolletino di Geodesia e Scienze Affini*, **LVI**(2), 173–206.

Schwarz, K. P., Sideris, M. G., and Forsberg, R., 1990. The use of FFT in physical geodesy. *Geophysical Journal International*, **100**, 485–514.

Sideris, M. G., 1987. Spectral methods for the numerical solution of Molodensky's problem. UCSE Report No. 20024, Department of Surveying Engineering, The University of Calgary, Calgary, Alberta.

Sideris, M. G., 1990. Rigorous gravimetric terrain modelling using Molodensky's operator. *Manuscripta Geodaetica*, **15**, 97–106.

Sideris, M. G., 1996. On the use of heterogeneous noisy data in spectral gravity field modelling methods. *Journal of Geodesy*, **70**(8), 470–479.

Sideris, M. G., and Tziavos, I. N., 1988. FFT-evaluation and applications of gravity-field convolution integrals with mean and point data. *Bulletin Geodesique*, **62**, 521–540.

Strang van Hees, G., 1990. Stokes' formula using fast Fourier techniques. *Manuscripta Geodaetica*, **15**, 235–239.

Cross-references

Geodesy, Figure of the Earth
Geodesy, Physical
Geoid Determination, Theory and Principles
Geoid Undulation, Interpretation
Gravimeters
Gravity Anomalies, Interpretation
Gravity Data, Advanced Processing
Gravity Field of the Earth
Gravity Method, Satellite
Gravity, Data to Anomalies
Gravity, Global Models

GEOMAGNETIC EXCURSIONS

Martha Schwartz
Department of Earth Sciences, University of Southern California, Los Angeles, CA, USA

Synonyms

Excursions were initially observed as "tiny wiggles" in marine magnetic anomalies (LaBreque et al., 1977) and were termed "cryptochrons" by Cande and Kent (1992a, b).

Definition

Geomagnetic Excursions are periods of unusually high-amplitude secular variation (see *Geomagnetic Field, Secular Variation*). They are usually defined as short-term deviations of the Virtual Geomagnetic Dipole (VGP) more than 40–45° from the geographic poles (see *Geomagnetic Field, Global Pattern*).

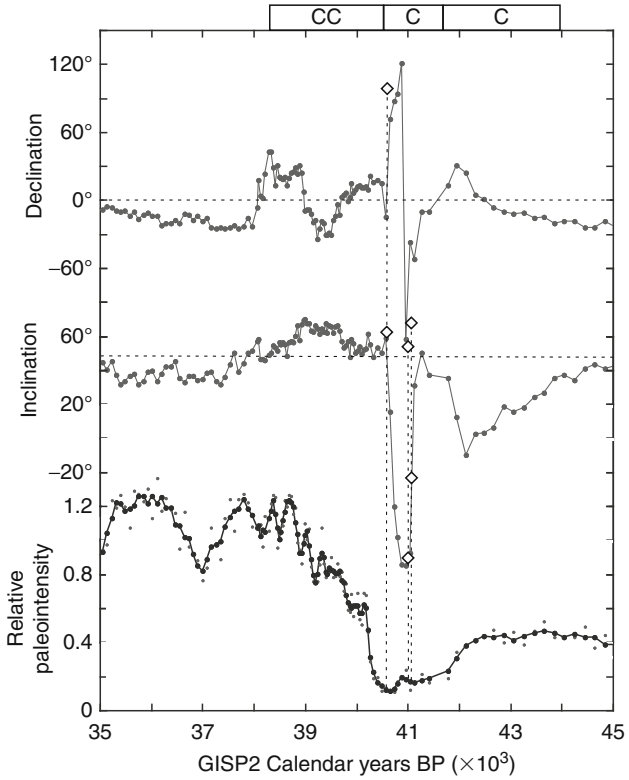
Geographical and temporal extent

Excursions last for up to a few thousand years during times of otherwise stable field polarity. The unusual directions, which sometimes represent regionally complete reversals (see *Geomagnetic Field, Polarity Reversals*),

are always accompanied by globally correlative low levels of field intensity (see *Paleomagnetic Field Intensity*). Low intensity starts earlier than the anomalous directions and outlasts them. There is some emerging evidence that excursions are global in extent. However, uncertainties in age control limit the ability to establish synchronicity among directional records recovered from different regions. Inadequate paleomagnetic recording may also cause excursions to be missed at certain localities. Lava flows do not provide a continuous record of field behavior, and sediment records of short-term directional field behavior may be smeared due to low sedimentation rates (see *Geomagnetic Field, Measurement Techniques*).

The frequency and duration of geomagnetic excursions are also open issues. Answers depend on recovering further high-quality replicated records, as well as better understanding of their regional versus global extent. There is reproducible evidence for 17 time intervals within the Brunhes Chron with at least regionally verifiable magnetic field excursions (Lund et al., 2006). Two well-recorded examples of excursions are the Laschamp Excursion, first noted in French lavas (Bonhommet and Babkine, 1967) which occurred about 40 ka, and the Blake Event (first observed by Smith and Foster [1969]) from about 120 ka.

The Figure 1, reproduced from Lund et al. (2005), is a high-resolution record of the Laschamp Excursion, as recovered from the Blake Outer Ridge in the North Atlantic Ocean.



Geomagnetic Excursions, Figure 1 Paleomagnetic secular variation and relative paleointensity (small dots are original data, large dots are 3-pt running average) records from Blake Outer Ridge core JPC14 for the time interval of the Laschamp Excursion. NRM after 20 mT demagnetization was used outside the excursion interval and NRM characteristic remanences were used in the excursion interval. These records show the same distinctive interval of low paleointensity seen in other cores (Figures 2, 3). But it contains a much more dramatic pattern of directional variability that we consider to be a more accurate and high-resolution record of the Laschamp Excursion as seen on the Blake Outer Ridge. The dashed lines indicate three time intervals where directions in this record are almost identical to the Laschamp Excursion directions seen in Europe (open diamonds). These three intervals also have similar paleointensity lows. Boxes at top indicate intervals of clockwise (C) or counterclockwise (CC) looping (circularity) in the directional data.

Field configuration during excursions

Complete understanding of the mechanism causing excursions depends on improved mapping of the global field geometry during excursional intervals. Roberts (2008) summarized three suggested excursionary field configurations. (1) The field remains largely dipolar but wanders away from the rotational axis. In this case VGP's would be consistent regardless of sampling site. (2) The ratio of non-dipole and dipole field components becomes very high. In this case VGPs would depend strongly on the sampling site. (3) Relatively localized sources in the outer-core produce regional excursions (see *Geomagnetic Field, Theory*).

Several numerical dynamo models are able to simulate magnetic field excursions for certain choices of boundary conditions. But model numerical simulations are limited in spatial or temporal resolution due to presently limited computer technology (Glatzmaier, 1999).

Origin

The most widely accepted explanation for the origin of geomagnetic excursions comes from Gubbins (1999). According to this theory, the outer core reverses briefly but the solid inner core, which has a magnetic diffusion time on the order of 3,000 years, does not reverse. In the case in which the inner core did have time to turn around, the event would be recorded as a full field reversal (see *Geomagnetic Field, Polarity Reversals*).

Summary

Geomagnetic excursions are brief periods of anomalous directions coupled with low field intensity. Their spatial extent, duration, and causative mechanism are under study. A review paper by Laj and Channel (see bibliography) provides an excellent summary of the research to date.

Bibliography

Bonhommet, N., and Babkine, J., 1967. Sur la presence d'aimentation inverse dans la Chaine des Puys. *Comptes Rendus Hebdomadiares des Seances de l'Academy des Science Series B*, **264**, 92–94.
Cande, S. C., and Kent, D. V., 1992a. A new geomagnetic polarity timescale for the late cretaceous and cenozoic. *Journal of Geophysical Research*, **97**, 13917–13951.
Cande, S. C., and Kent, D. V., 1992b. Ultrahigh resolution marine magnetic anomaly profiles: a record of continuous paleointensity variations? *Journal of Geophysical Research*, **97**, 15075–15083.
Gubbins, D., 1999. The distinction between geomagnetic excursions and reversals. *Geophysical Journal International*, **137**, F1–F4.
Glatzmaier, G. A., 1999. The role of the Earth's mantle in controlling the frequency of geomagnetic reversals. *Nature*, **401**, 885–990.
LaBreque, J. L., Kent, D. V., and Cande, S. C., 1977. Revised magnetic polarity time-scale for the late cretaceous and cenozoic time. *Geology*, **5**, 330–335.
Laj, C., and Channel, J. E. T., 2007. Geomagnetic excursions. In Kono, M. (ed.), *Treatise on Geophysics, Geomagnetism*. Amsterdam: Elsevier, Vol. 5, pp. 373–416.
Lund, S. P., Schwartz, M., Keigwin, L., and Johnson, T., 2005. Deep-sea sediment records of the Laschamp geomagnetic field excursion (~41,000 calendar years before present). *Journal of Geophysical Research*, **110**, B04101, doi:10.1029/2003JB002943.
Lund, S., Stoner, J. S., Channell, J. E. T., and Acton, G., 2006. Brunhes paleomagnetic field variability recorded in ocean drilling program cores. *Physics of the Earth and Planetary Interiors*, **156**, 194–205.
Roberts, A., 2008. Geomagnetic excursions: knowns and unknowns. *Geophysical Research Letters*, **35**, L17307.
Smith, J. D., and Foster, J. H., 1969. Geomagnetic reversal in the Brunhes normal polarity epoch. *Science*, **163**, 565–567.

Cross-references

Geomagnetic Field, Global Pattern
Geomagnetic Field, Measurement Techniques
Geomagnetic Field, Polarity Reversals
Geomagnetic Field, Theory
Paleomagnetic Field Intensity

GEOMAGNETIC FIELD, GLOBAL PATTERN

Susan Macmillan
British Geological Survey Murchison House, Edinburgh, Scotland, UK

Introduction

The geomagnetic field is generated in the fluid outer core region of the Earth by electrical currents flowing in the slowly moving molten iron. In addition to sources in the Earth's core, the geomagnetic field observable on the Earth's surface has sources in the crust and in the ionosphere and magnetosphere. The signal from the core dominates, accounting for over 95% of the field at the Earth's surface. The geomagnetic field varies on a range of scales, both temporal and spatial; the description of the variations made here concentrates on the recent spatial and temporal variations of the field with origins in the Earth's core that can be surmised from observations made over the last four centuries.

Observations

The geomagnetic field is a vector quantity, having magnitude as well as direction. The direction of the geomagnetic field has been directly observed and recorded for at least 400 years beginning with observations of the angle in the horizontal plane between true north and the magnetic field vector, known as declination or magnetic variation. In 1634, it was realized by Henry Gellibrand that declination in London was changing with time when he compared his observations with those made by Edmund Gunter 12 years earlier. Since then, this important discovery of secular variation has ensured regular observations of the magnetic field through time. Measurements of declination were important for navigation across the oceans, and a source of early observations is ships' logbooks (Jackson et al., 2000). Before long it was also realized that other elements of the geomagnetic field were of interest, in particular the angle of dip of the magnetic field vector from the horizontal known as inclination, and this was also measured. However, it was not till 1832 when Carl Freidrich Gauss established his technique for measuring absolute intensity did we have accurate measurements of the magnitude of the geomagnetic field.

As demands on accuracy and interest in the geomagnetic field increased, permanent observatories began to be established. Since the 1840s the number of observatories around the world has slowly increased but has now dropped to about 160 in recent times. The advent of the proton precession magnetometer and the fluxgate magnetometer in the twentieth century considerably eased the automation of observatories during this time. However, regular manual absolute observations, nowadays of declination and inclination using a fluxgate theodolite, are necessary to maintain accurate baseline control over long periods of time. There are also networks of repeat stations providing data over extensive areas from which country-wide magnetic charts can be derived. More recently, magnetic measurements made by satellites have become important in determining the pattern of the geomagnetic field. In particular, vector data from Magsat (1979–1980), Ørsted (1999-current), and CHAMP (2000–2010) have all been utilized in the production of recent spherical harmonic models of the geomagnetic field.

Spherical harmonic analysis

Till the 1950s, magnetic charts at a given epoch for use in navigation depended on manually drawing contours through the observations (Malin, 1971). Although spherical harmonic analysis had been developed in 1839 by Carl Freidrich Gauss, it was not routinely used to fit mathematical models to the observations till the advent of computers. Before computers spherical harmonic

analysis required that the data be in the form of values of one or more of the orthogonal components X, Y, and Z of the geomagnetic field (magnetic intensities in the northerly, easterly, and vertically down directions) at points regularly spaced in latitude and longitude (Barraclough, 1978). This involved the intermediate stage of interpolating values from manually drawn contour charts.

The mathematical and physical basis for spherical harmonic analysis is now given. In a source-free region, the Earth's magnetic field B is the negative gradient of a magnetic potential V that satisfies Laplace's equation:

$$B = -\nabla V \quad \text{where} \quad \nabla^2 V = 0$$

A solution to Laplace's equation for the magnetic potential arising from sources inside the Earth at a given epoch is given by:

$$V(r,\theta,\lambda) = a\sum_{n=1}^{n_{max}} \left(\frac{a}{r}\right)^{n+1} \sum_{m=0}^{n} \left(g_n^m \cos m\lambda + h_n^m \sin m\lambda\right) P_n^m(\theta)$$

In this equation, r, θ, λ are geocentric coordinates (r is the distance from the center of the Earth, θ is the colatitude, i.e., 90° - latitude, and λ is the longitude), a is a reference radius for the Earth (6371.2 km), g_n^m and h_n^m are the spherical harmonic Gauss coefficients of degree n and order m, $P_n^m(\theta)$ are the Schmidt semi-normalized Legendre functions and n_{max} is the maximum degree of the spherical harmonic expansion. If the observations of the magnetic field are spread over time, the magnetic potential and corresponding Gauss coefficients are also dependent on time with splines being commonly used.

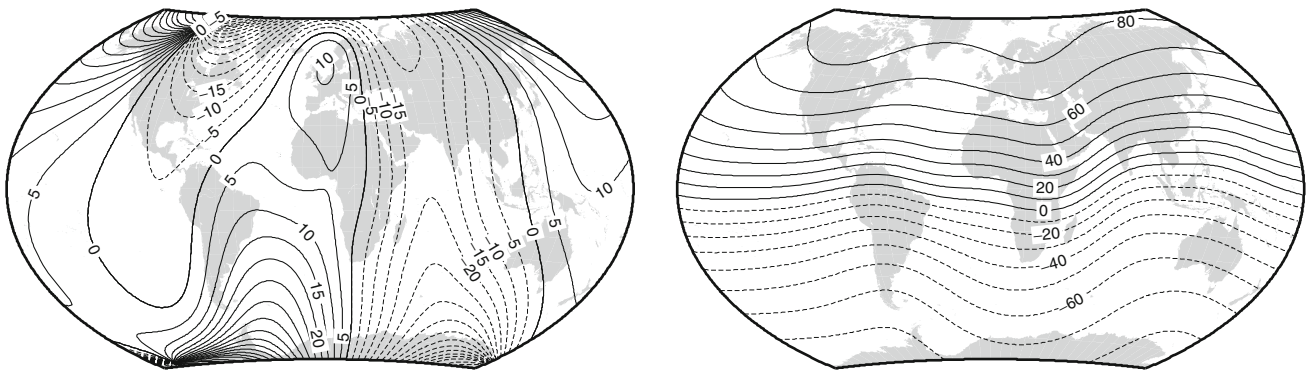
Maps of the geomagnetic field

It is worth noting here that the first published map of the geomagnetic field is that of declination made in 1701 by Edmond Halley. It was based on declination observations made during a magnetic survey expedition under naval command and covered the Atlantic Ocean (Clark and Barraclough, 2001).

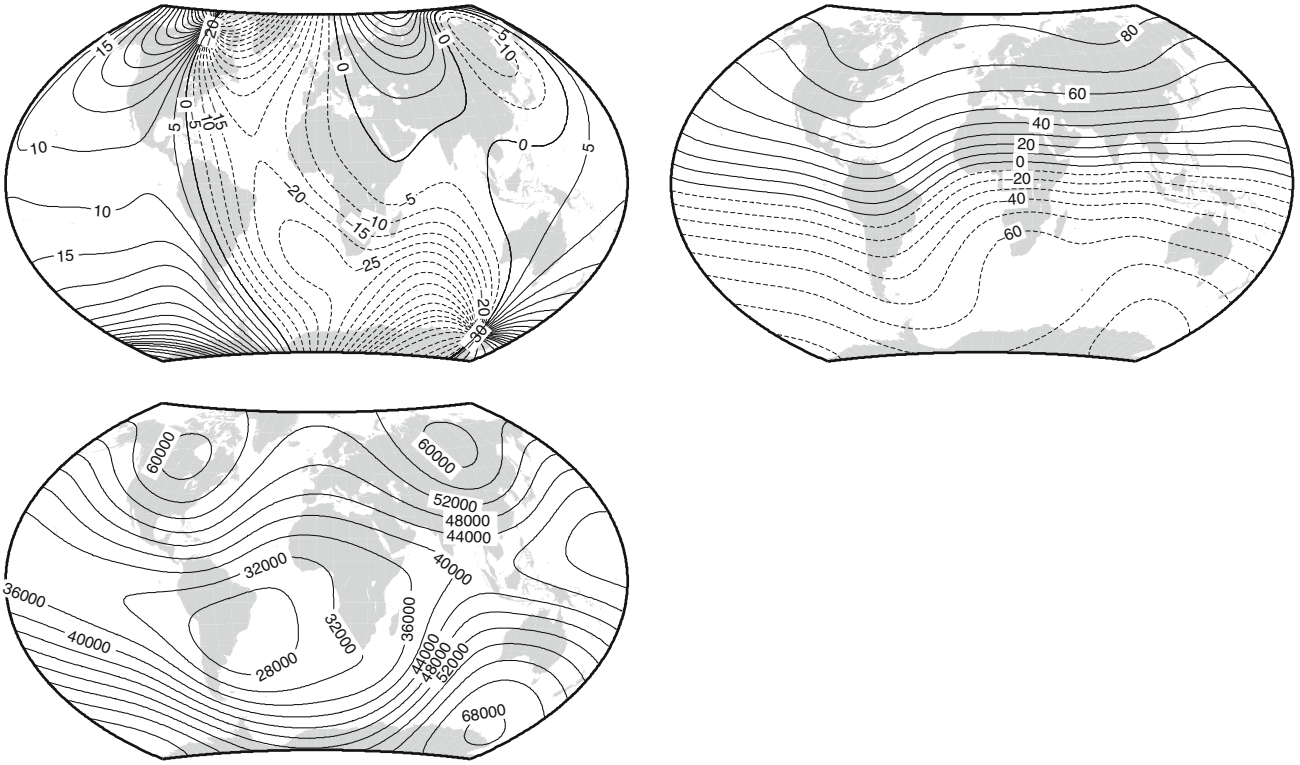
The maps that we show here are based on spherical harmonic models. Nowadays, there are many such models of the geomagnetic field and here we look at two examples. One is the *gufm1* model which is derived from four centuries of magnetic field observations (Jackson et al., 2000) and extends to spherical harmonic degree 14 with B-splines being used to fit the variations in time from 1590 to 1990. It has smoothness imposed in both the spatial and temporal domains. The other model is the International Geomagnetic Reference Field (Macmillan and Finlay, 2011) produced under the auspices of the International Association of Geomagnetism and Aeronomy (IAGA). It extends to spherical harmonic degree 10 till 1995 and thereafter extends to degree 13. The variations in time from 1900 to 2015 in the IGRF are assumed to be piecewise linear and there is effectively no smoothness imposed in either the spatial or temporal domains. The final IGRF coefficients are made up of weighted averages of candidate sets of coefficients from (mostly) unregularized models.

Using the *gufm1* model, plots of declination and inclination at 1600, declination, inclination, and total intensity at 1950 are shown in Figures 1 and 2. Using the eleventh generation IGRF, plots of the magnetic elements at 2010 and their rates of change are shown in Figures 3–9. The rates of change plots are derived from a predictive secular variation model which extends to spherical harmonic degree 8. Up-to-date maps, and on-line calculators are available from a number of websites, for example, http://www.geomag.bgs.ac.uk/navigation.html and http://www.ngdc.noaa.gov/geomag/WMM/DoDWMM.shtml.

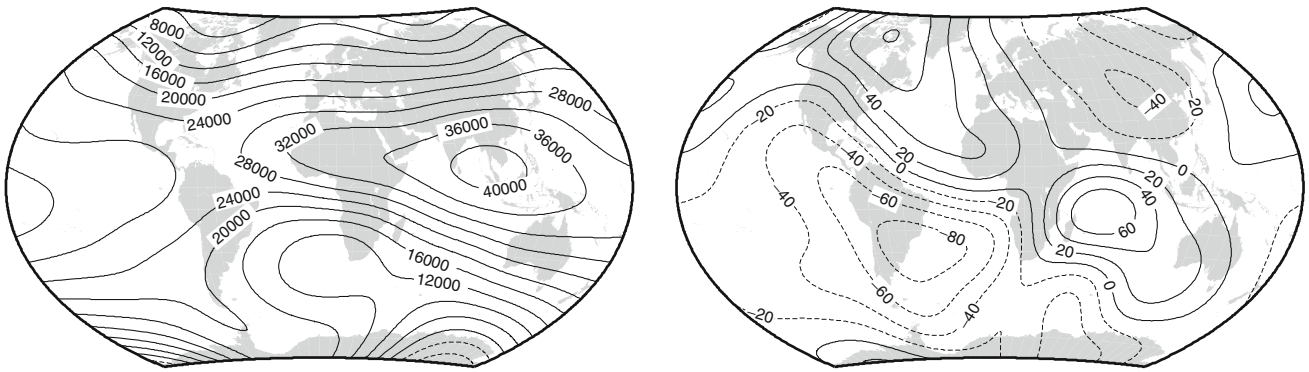
Features of note on these maps are the dip equator and dip poles in the maps of inclination (Figures 1, 2 and 8), and the area of weak magnetic field in the South Atlantic, and strong magnetic field near the poles in Figures 2 and 9. The dip equator is where the magnetic field is horizontal (inclination = 0°) and a current system is set up in the upper atmosphere called the equatorial electrojet. The dip poles (also called magnetic poles) are locations where the magnetic field is, on average, vertical. Another set of poles is the dipole poles or geomagnetic poles. Their positions can be derived from the degree 1 spherical harmonic coefficients (g_1^0, g_1^1, and h_1^1). Associated with the



Geomagnetic Field, Global Pattern, Figure 1 Declination and inclination at 1600.0 (degrees) computed from the *gufm1* model.



Geomagnetic Field, Global Pattern, Figure 2 Declination, inclination (degrees), and total intensity (nT) at 1950.0 computed from the *gufm1* model.



Geomagnetic Field, Global Pattern, Figure 3 Northerly intensity (nT) at 2010.0 and its rate of change (nT/year) for 2010.0–2015.0 computed from IGRF-11.

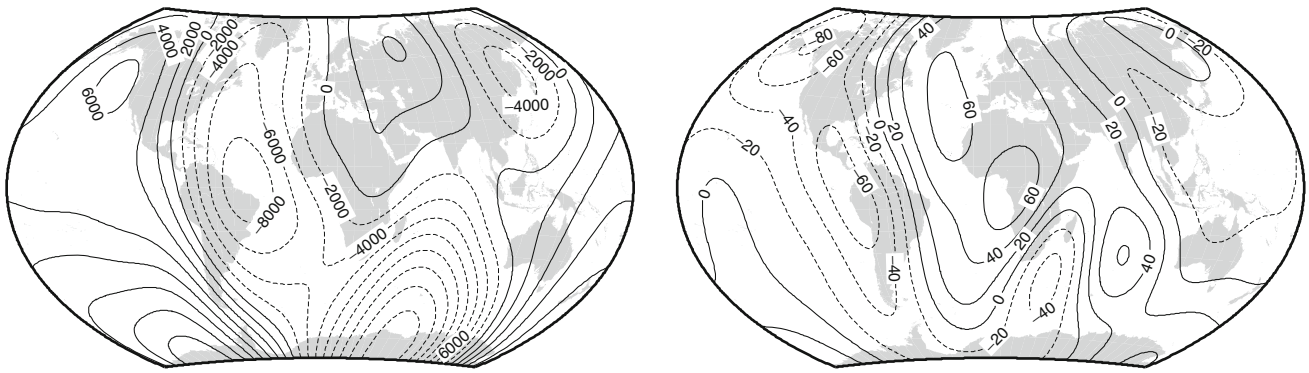
geomagnetic poles, in approximately oval-shaped loci, are the auroral electrojets in the upper atmosphere. The electrojets, particularly the auroral electrojets, generate rapid time-varying magnetic fields of significant amplitude.

Recent changes in the geomagnetic field

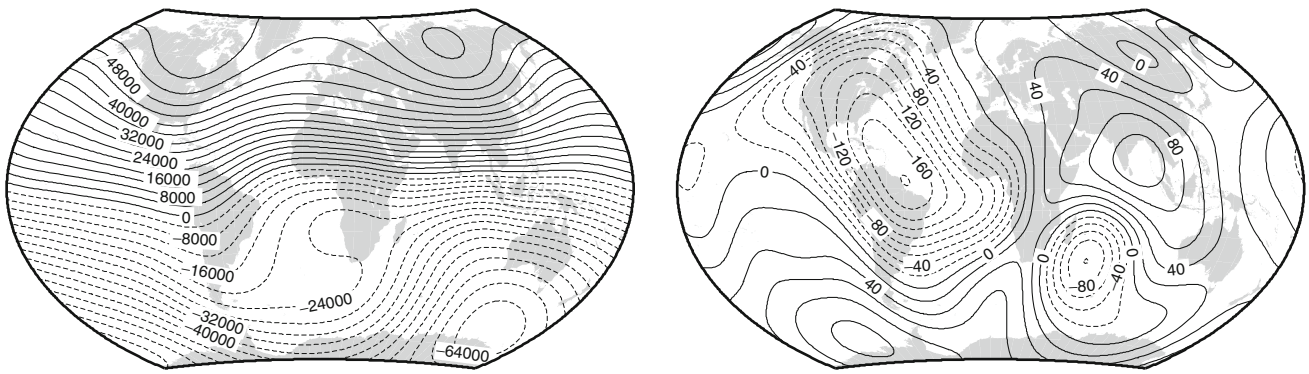
From a comparison of the declination maps in Figures 2 and 7 it can be seen that the zero contour (the agonic line) is moving slowly westwards with time. This westwards movement of the magnetic field pattern at the Earth's surface, particularly prevalent in the Atlantic hemisphere, is related to the motion of fluid at the core surface slowly westwards, dragging with it the magnetic field lines.

Using IGRF-11 to compute the root mean square magnetic field vector at the Earth's surface arising from all spherical harmonic terms ($n \leq 10$), the centred dipole terms ($n = 1$) and the non-dipole terms ($1 < n \leq 10$), it can be seen in Figure 10 that since 1900, the geomagnetic

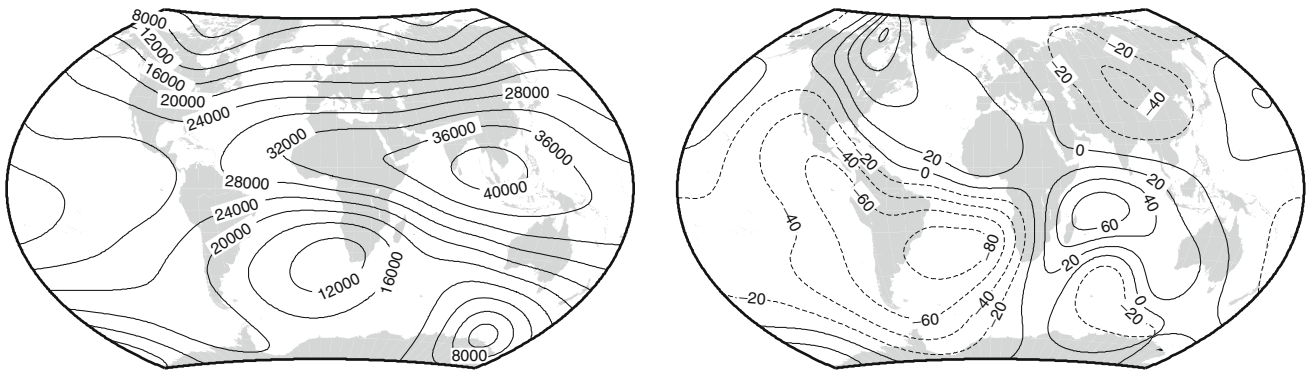
field is weakening overall by becoming less dipolar. However the non-dipolar part is strengthening. This may have consequences for the trajectories of energetic charged particles that enter the Earth's magnetosphere. One manifestation of this is the deepening, and westwards movement, of the South Atlantic Anomaly, the region where the geomagnetic field is weaker than elsewhere (see Figures 2 and 9). Energetic charged particles are able to penetrate closer to the Earth and cause a radiation hazard for satellites passing through this region.



Geomagnetic Field, Global Pattern, Figure 4 Easterly intensity (nT) at 2010.0 and its rate of change (nT/year) for 2010.0–2015.0 computed from IGRF-11.

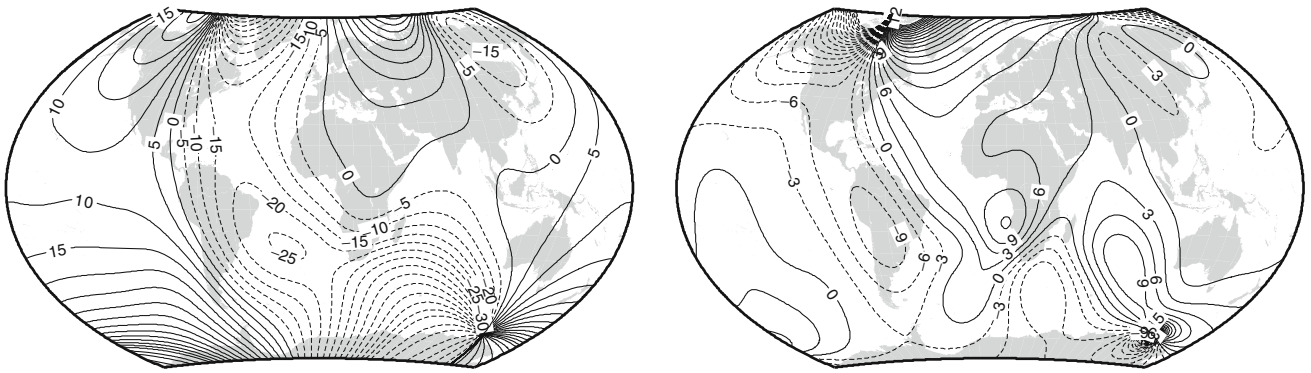


Geomagnetic Field, Global Pattern, Figure 5 Vertical intensity (nT) at 2010.0 and its rate of change (nT/year) for 2010.0–2015.0 computed from IGRF-11.

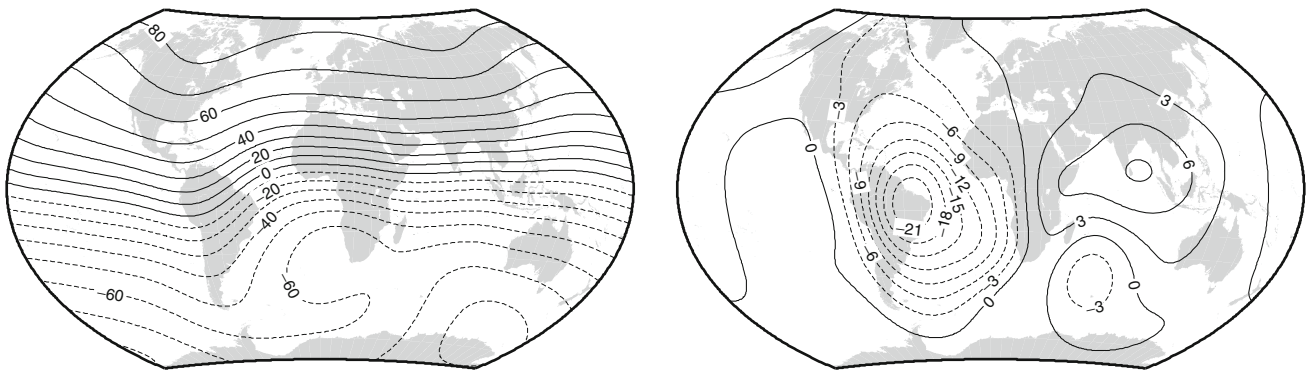


Geomagnetic Field, Global Pattern, Figure 6 Horizontal intensity (nT) at 2010.0 and its rate of change (nT/year) for 2010.0–2015.0 computed from IGRF-11.

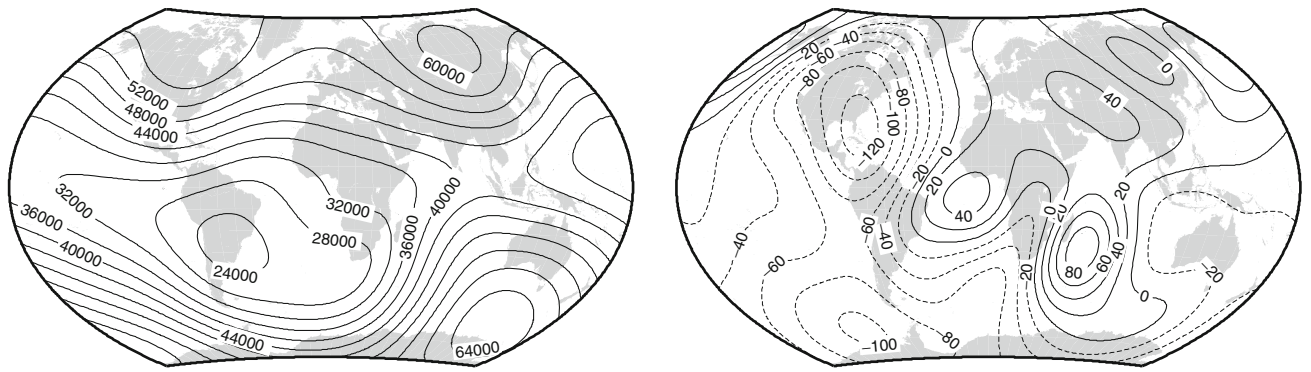
Superimposed on these gradual changes of the magnetic field are the so-called geomagnetic jerks. They are relatively abrupt (duration of months) changes in the second time derivative, or secular acceleration, of the magnetic field. The first observed geomagnetic jerk was that around 1969, and since the late nineteenth century when direct and continuous measurements of the Earth's magnetic field have been available, geomagnetic jerks



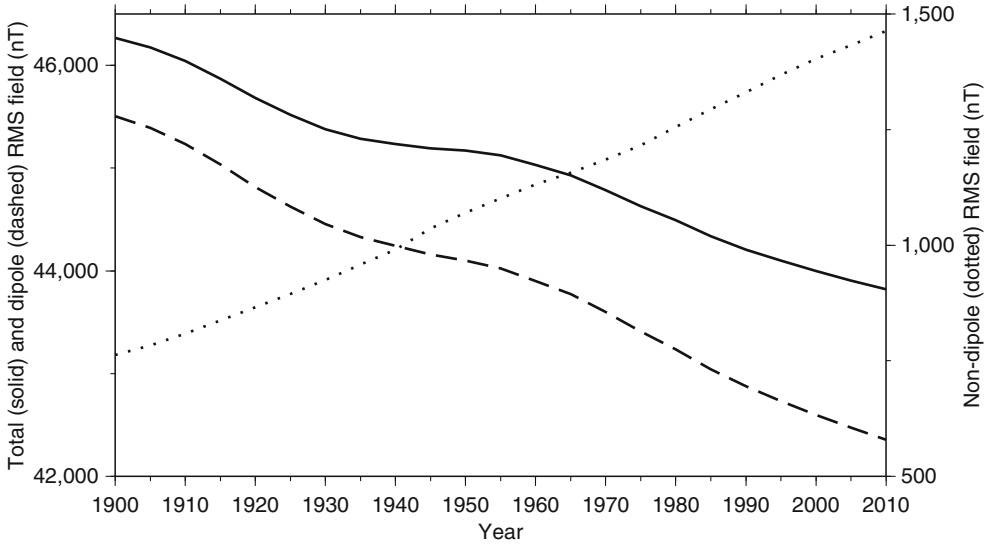
Geomagnetic Field, Global Pattern, Figure 7 Declination (degrees) at 2010.0 and its rate of change (arc-minutes/year) for 2010.0–2015.0 computed from IGRF-11.



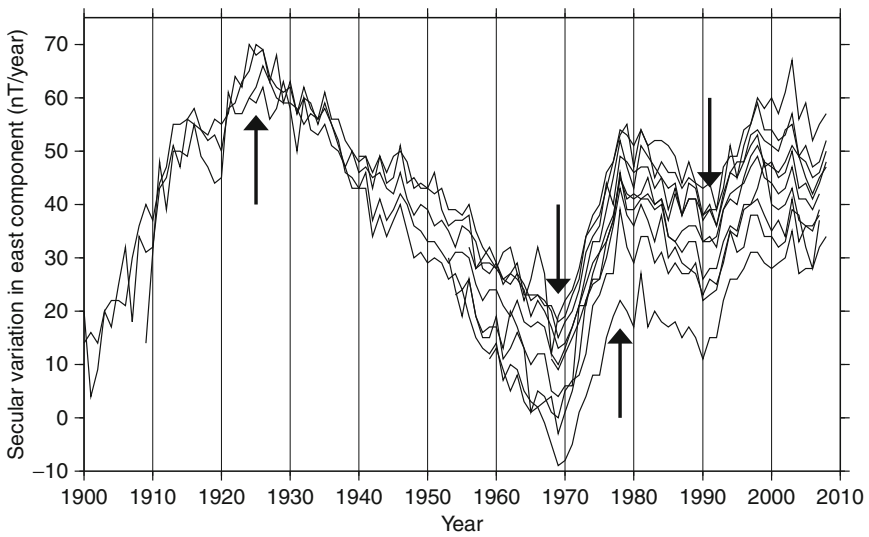
Geomagnetic Field, Global Pattern, Figure 8 Inclination (degrees) at 2010.0 and its rate of change (arc-minutes/year) for 2010.0–2015.0 computed from IGRF-11.



Geomagnetic Field, Global Pattern, Figure 9 Total intensity (nT) at 2010.0 and its rate of change (nT/year) for 2010.0–2015.0 computed from IGRF 11.



Geomagnetic Field, Global Pattern, Figure 10 The decline of the whole, and dipolar part of the geomagnetic field at the Earth's surface and the growth of the non-dipolar part since 1900, computed from IGRF-11.



Geomagnetic Field, Global Pattern, Figure 11 Geomagnetic jerks as seen in the secular variation of the east component of the geomagnetic field observed at European observatories (IAGA codes of the observatories used are LRV, DOB, NUR, LER, ESK, HAD, DOU, BDV, FUR, CLF). Times of jerks are shown by arrows.

have also been observed to occur around 1925, 1978, 1991, and 1999. These jerks are most readily observed in the first time derivative of the easterly intensity at European observatories (Figure 11).

Various analysis techniques have been applied to jerks to investigate specific aspects of their temporal and spatial characteristics. Using wavelets, the 1969 and 1978 jerks have been shown to have different arrival times at the Earth's surface, with the northern hemisphere leading the southern hemisphere by about 2 years (Alexandrescu et al., 1996).

Understanding their origin is important, not only because they result from interesting dynamical processes in the core and may help determine the conductivity of the mantle, but also for improving time-dependent models of the geomagnetic field and for the strictly practical purpose of forecasting its future behaviour, for example, as used in navigation.

Bibliography

Alexandrescu, M., Gibert, D., Hulot, G., Le Mouël, J.-L., and Saracco, G., 1996. Worldwide wavelet analysis of geomagnetic jerks. *Journal of Geophysical Research*, **101**(B10), 21,975–21,994.

Barraclough, D. R., 1978. Spherical harmonic models of the geomagnetic field. *Geomagnetic Bulletin Institute of Geological Sciences*, No. 8.

Clark, T., and Barraclough, D., 2001. The first magnetic chart. *Astronomy and Geophysics*, **42**, 1.23–1.25.

Jackson, A., Art, R. T. J., and Matthew, R. W., 2000. Four centuries of geomagnetic secular variation from historical records. *Philosophical Transactions of the Royal Society of London Series A-Mathematical Physical and Engineering Sciences*, **358**, 957–990.

Macmillan, S., and Finlay C., 2011. The International Geomagnetic Reference Field. In Mandea, M., Korte, M. (eds.), *Geomagnetic Observations and Models*. IAGA Special Sopron Book Series 5, DOI 10.1007/978-90-481-9858-0_10.

Malin, S. R. C., 1971. British World Magnetic Charts. In Zmuda, A. J. (ed.), World Magnetic Survey 1957–1969, *IAGA Bulletin* No. 28.

Cross-references

Geomagnetic Field, IGRF
Geomagnetic Field, Measurement Techniques
Geomagnetic Field, Secular Variation
Geomagnetic Field, Theory
Magnetic Methods, Satellite
Magnetic, Global Anomaly Map

GEOMAGNETIC FIELD, IGRF

Aude Chambodut
Laboratoire de Géodynamique Globale: Champs de Potentiels et Déformations Actives (CNRS UMR 7516), Ecole et Observatoire des Sciences de la Terre (UDS), Strasbourg Cédex, France

Definition

The International Geomagnetic Reference Field (IGRF) is a series of mathematical models of the Earth's main field and its annual rate of change (secular variation).

Each generation of the IGRF comprises three constituent models:

1. The IGRF model that is a prediction of the state of the main field at an epoch t (considering that the epoch of IGRF model elaboration is approximately between [*t-3 years*] and [*t-1 years*]).
2. The secular variation model that is also a prediction of the annual variation rate of the main field between epoch t and ($t + 5$ *years*).
3. The Definitive Geomagnetic Reference Field model (DGRF), which is the revision of the previous generation IGRF model 5 years before the epoch (*t-5 years*).

The constituent models are designated provisional or definitive. Once a constituent model is designated definitive it is not revised in subsequent generations of the IGRF.

Datasets

As it can be seen in Table 1, the periodicity of IGRF's elaboration varies with time. In addition, upgrades are regularly incorporated. For example, the truncation degree/order of the main field spherical harmonic analysis was extended from 10 up to 13 after 1995 (IAGA, Division V, Working Group 8, 2003). Similarly, the precision of the coefficients was extended from 1 nT to 0.1 nT after 1995 (IAGA, Division V, Working Group 8, 2003) and then to 0.01 nT after 2000 (IAGA, Working Group V-MOD, 2010).

These two changes were made possible by technological advances. Indeed, before 1979, only near-surface data (permanent observatories, repeat stations, land, airborne and marine surveys; see *Geomagnetic Field, Measurement Techniques*; *Magnetic Methods, Airborne*) were used (Peddie, 1983; Zmuda, 1971) and, consequently, the spatial distribution of data over the globe remained poor. The International Decade of Geopotential Field Research allowed modelers to take advantage of the excellent spatial coverage and high quality of measurements made by near-polar satellites (see *Magnetic Methods, Satellite*). The POGO satellites (1965–1971), Magsat (1979–1980), POGS (1990–1993), Oersted (1999–), SAC-C (2000–), and CHAMP (2000–), have all been utilized in the production of the IGRF.

Today, data from permanent magnetic observatories are still the most important source of information about time-varying fields, especially for secular variation constituent models (see *Geomagnetic Field, Secular Variation*) while magnetic satellite missions allow the acquisition of homogeneously distributed data over the entire globe during a short period of time (as compared to magnetic observatories time series). The two datasets remain highly complementary.

Basic mathematical expression

Each mathematical model consists in a set of coefficients, called Gauss coefficients, associated to solid spherical harmonic functions.

Indeed, in source-free regions (the spherical shell bounded by the Earth's surface and the base of the ionosphere at an altitude around 100 km), the main field, generated by fluid motions in the Earth's electrically conducting liquid outer core, may be expressed as the negative gradient of a scalar potential V, which is represented by a truncated (up to degree N) series expansion (see *Geomagnetic Field, Theory*):

$$V(r,\theta,\varphi) = R\sum_{n=1}^{N}\left(\frac{R}{r}\right)^{n+1}\sum_{m=0}^{n}\left(g_n^m\cos(m\varphi) + h_n^m\sin(m\varphi)\right)P_n^m(\theta)$$

where r, θ, φ are geocentric coordinates according to a reference coordinate system.

Geomagnetic Field, IGRF, Table 1 Summary of IGRF history

Full name	Short name	Valid for	Definitive for
IGRF 11th generation (revised 2009)	IGRF-11	1900–2015	1945–2005
IGRF 10th generation (revised 2004)	IGRF-10	1900–2010	1945–2000
IGRF 9th generation (revised 2003)	IGRF-9	1900–2005	1945–2000
IGRF 8th generation (revised 1999)	IGRF-8	1900–2005	1945–1990
IGRF 7th generation (revised 1995)	IGRF-7	1900–2000	1945–1990
IGRF 6th generation (revised 1991)	IGRF-6	1945–1995	1945–1985
IGRF 5th generation (revised 1987)	IGRF-5	1945–1990	1945–1980
IGRF 4th generation (revised 1985)	IGRF-4	1945–1990	1965–1980
IGRF 3rd generation (revised 1981)	IGRF-3	1965–1985	1965–1975
IGRF 2nd generation (revised 1975)	IGRF-2	1955–1980	–
IGRF 1st generation (revised 1969)	IGRF-1	1955–1975	–

r is the distance from the center of the Earth, θ is the colatitude (i.e., 90° – latitude), and φ is the longitude, R is a reference radius (Earth mean radius) set to 6371.2 km; ${g^n}_m$ and h^n are the Gauss internal coefficients, and ${P^n}_m(\theta)$ are the Schmidt semi-normalized associated Legendre functions of degree n and order m.

The IGRF coefficients are time dependent. In the series of IGRF mathematical models, the variations are assumed to be linear on 5-year intervals. For the upcoming 5-year epoch, the rate of change is given by predictive secular variation coefficients (see *Geomagnetic Field, Secular Variation*).

IGRF evolving

One of the first suggestions to develop an International Geomagnetic Reference Field model was made in 1954 at the International Association of Geomagnetism and Aeronomy (IAGA) Scientific General Assembly in Toronto. The aim in developing such reference models was to have a common basis for calculating the main geomagnetic field in order to homogenize the results of academical/fundamental studies (on fields as diverse as the crustal and external fields) and to provide the industrial world (Meyers and Davis, 1990) with a well-identified description of the Earth's core magnetic field.

Thus, an internationally coordinated effort was carried out under the aegis of IAGA relying on cooperation between magnetic field modelers, institutes, and agencies responsible for collecting and publishing geomagnetic field data.

Since 1969, date of the first adoption of an IGRF (IGRF-1 valid for 1955–1975), the IGRF has evolved into a series of mathematical models of the Earth's main field (Langel, 1987) and its annual rate of change at 5-year intervals covering the period 1900–2015 (Table 1).

New candidate models are carefully produced and widely documented as are their evaluation procedures. The final decision is made by the IAGA Working Group-V Mod. Most of the time, the elected DGRF and IGRF models of the current generation are derived by taking means (sometimes weighted) of the coefficients of selected candidate models. Indeed, one particular selected candidate model cannot be chosen preferably to another as the various datasets, selection criteria of the data, applied corrections and even modeling methods have their own strengths and weakness. The robustness of IGRF lies in the fact that even if there are significant differences between the candidate models in every part of the development chain (datasets, data selections, algorithms, regularizations), the results are fairly consistent at the end.

General Information about the IGRF and coefficients of the 11th generation are available at: http://www.ngdc.noaa.gov/IAGA/vmod/igrf.html.

Additional figures and programs to compute the field components for a specific location and time are available at: http://wdc.kugi.kyoto-u.ac.jp/igrf/index.html.

Bibliography

International Association of Geomagnetism and Aeronomy (IAGA) Division V, Working Group 8, 2003. The 9th generation international geomagnetic reference field. *Earth Planets Space*, **55**, i–ii.

International Association of Geomagnetism and Aeronomy (IAGA), Working Group V-MOD, 2010. International Geomagnetic Reference Field: The Eleventh Generation. *Geophysical Journal International*, accepted.

Langel, R. A., 1987. The main field. In Jacobs, J. A. (ed.), *Geomagnetism*. London: Academic, Vol. 1, pp. 249–512.

Meyers, H., and Davis, W. M., 1990. A profile of the geomagnetic model users and abusers. *Journal of Geomagnetism and Geoelectricity*, **42**, 1079–1085.

Peddie, N. W., 1983. International geomagnetic reference field – its evolution and the difference in total field intensity between new and old models for 1965–1980. *Geophysics*, **48**, 1691–1696.

Zmuda, A. J., 1971. The international geomagnetic reference field: introduction. *Bulletin International Association of Geomagnetism and Aeronomy*, **28**, 148–152.

Cross-references

Geomagnetic Field, Measurement Techniques
Geomagnetic Field, Secular Variation
Geomagnetic Field, Theory
Magnetic Methods, Airborne
Magnetic Methods, Satellite

GEOMAGNETIC FIELD, MEASUREMENT TECHNIQUES

Mioara Mandea[1], Anca Isac[2]
[1]Université Paris Diderot – Institut de Physique du Globe de Paris, Géophysique spatiale et planétaire, Paris Cedex 13, France
[2]Geological Institute of Romania, Bucharest, Romania

Introduction

Continuous monitoring, modeling, and interpretation of the geomagnetic field and its spatial and temporal variations are the prerequisites to understand our magnetic planet. Based on geomagnetic models (see *Geomagnetic Field, Global Pattern*), covering very short to longer temporal scales, it makes it possible to understand core dynamics (see *Core Dynamo*), lithospheric magnetic signals (see *Magnetic, Global Anomaly Map*; *Magnetic Anomalies, Interpretation*), and the geomagnetic environment. This represents an important challenge in understanding the System Earth. Clearly, geomagnetic measurements from ground and space represent the main ingredient in modeling and understanding the Earth's magnetic field (Mandea and Purucker, 2005, *Geomagnetic Field, IGRF*). Furthermore, magnetic measurements are also used in connection with different geophysical surveys (ground and/or airborne) in detecting mineralization and/or geological and tectonic structures. They are also used to detect archeological sites, shipwrecks, and other buried or submerged objects. Moreover, magnetic techniques are used at drilling sites for navigating well-bores, or in monitoring the geomagnetically induced currents that can result in damage to transformers, a potentially major expense to power companies. Finally, prediction of space weather is crucial considering its potential impact on man-made technologies on Earth and in space (satellites and spacecraft, electricity power grids, pipelines, radio and telephone communications, geophysical exploration, etc.).

Two types of measurements are needed to characterize the geomagnetic field: scalar and vector. A scalar measurement gives a very accurate value of the total strength of the magnetic field, however, a full characterization of the magnetic field vector is not reached. A vector measurement has the capability to provide the components of the magnetic field in particular directions, relative to the spatial orientation of the sensors. These measurements are relative. Both kinds of measurements are needed for an accurate and full description of the magnetic field.

Scalar measurements are based on the transformation of a frequency – a very accurate measurement – into intensity of the magnetic field. In the spectral domain, the measurement is indeed the location of the center of the spectral line. The relative accuracy of this determination is very high, something as 10^{-12} being common in physics. For magnetic measurements a relative accuracy of order of 10^{-7} is reached, due to the rather broad spectral line. In a proton precession magnetometer the universal physical constant is the gyromagnetic ratio of the proton (42.58 MHz/T), and the basic unit is time (inverse of frequency). In a proton magnetometer, protons in a liquid (such as water or kerosene) are first polarized; when the polarizing field is switched off, the protons precess around the magnetic field direction with a frequency that is proportional to the magnetic field intensity. Overhauser magnetometers are typically installed at magnetic observatories, but they are also used for some other ground observations or on board of the Ørsted and CHAMP magnetic satellites.

In contrast, vector measurements made with a fluxgate magnetometer are based on the nonlinearity of the magnetization of soft magnetic materials. The very sensitive part of this instrument is an easily saturable core made of a material with high permeability. This core is wrapped by two windings: an excitation coil and a pick-up coil. An alternating excitation current is passed through the excitation coil, driving the core through an alternating cycle of magnetic saturation. This changing field induces an electrical current in the pick-up coil, and this output current can be accurately measured. If the core is exposed to an external magnetic field, it is more easily saturated in alignment with that field and less easily saturated in opposition to it. The signal in the pick-up coil has not only the excitation frequency, but also the other harmonics. The second harmonic is particularly sensitive to the background magnetic field. Often, the current in the pick-up coil is integrated, yielding an output analog voltage, proportional to the magnetic field. The fluxgate magnetometers are subject to an instrument drift. To minimize these drift contributions in the final data different approaches are used to calibrate the fluxgate magnetometers, on one hand for the ground measurements and on the other hand for the satellite measurements.

The instruments used to measure the magnetic field on ground or in space are based on the same physical concepts, and they are basically the same. Of course, for space missions constraints are imposed by weight and consuming energy. More details on magnetometers used in ground-based observations are given in the section Earth's Surface Magnetic Field Measurements. Considering the different and more complex approaches in processing data provided by satellite missions, we review them in the section Near-Earth Magnetic Field Measurements. Finally, in the section Summary we offer a perspective on promising avenues in measuring the magnetic field.

Earth's surface magnetic field measurements

Historically, magnetic observatories were established to monitor the secular change of the Earth's magnetic field, and this remains one of their major functions (see *Geomagnetic Field, Secular Variation*). This generally involves *absolute measurements* that are sufficient in number to monitor instrumental drift of fluxgate magnetometers giving *variation measurements* of the three field components, in arbitrary units (see *Magnetometers*).

A scalar measurement of the field intensity obtained commonly by a proton magnetometer is absolute: as noticed before, it depends only upon our knowledge of a physical constant and a measurement of frequency. The measurements achieved with such an instrument have a great accuracy (in excess of 10 ppm). Figure 1a shows an Overhauser magnetometer.

Scalar magnetometers make measurements of the strength of the magnetic field only, and provide no information about its direction. It is also possible to make an absolute measurement of direction of the geomagnetic field, by measuring the angle I (*inclination*) with respect to the horizontal plane, and the angle D (*declination*) in the horizontal plane between the magnetic North and true North (the last one determined by reference to a fixed mark of known azimuth, obtained astronomically or by using a gyro). A fluxgate-theodolite (DI-flux), as shown in Figure 1b, used commonly to measured D and I, requires manual operation. The DI-flux consists of a nonmagnetic theodolite and a single-axis fluxgate sensor mounted on top of a telescope. The DI-flux is considered to be an absolute instrument, which means that the angles measured by the instrument do not deviate from the true values D and I. This is achieved by using an observation procedure that eliminates unknown parameters such as sensor offset, collimation angles, and theodolite errors. For example, collimation errors between the fluxgate sensor and the optical axis of the theodolite, and within the theodolite, are minimized by taking readings from four telescope positions. With the fluxgate sensor operating in the null-field mode, the stability and sensitivity of the sensor and its electronics are maximized. In a land-based observatory, such absolute measurements are typically performed once/twice a week, manually, and are used to monitor the drift of the fluxgate variometers.

A vector measurement is made with a fluxgate magnetometer that operates with minimal manual intervention. These instruments require temperature-controlled environments and installation on extremely stable platforms (however, some modern systems are suspended and

Geomagnetic Field, Measurement Techniques, Figure 1 Geomagnetic observatory sensor package: proton magnetometer (**a**), fluxgate magnetometer (**b**), theodolite with fluxgate coupled to the telescope (**c**). The new prototype of Geomagnetic AUtomated SyStem (GAUSS) is also shown (**d**).

therefore compensate for platform tilt). Even with these precautions they can still be drifting subject to sources from both within the instrument (temperature effects) and also the stability of the instrument mounting. Because these measurements are not absolute, they are referred to as variation measurements, they have to be calibrated against standard instruments, and then the instruments are known as *variometers*. One of the most widely used variometers is the FGE fluxgate manufactured by the Danish Meteorological Institute, Denmark (see Figure 1c). The sensor unit consists of three orthogonally mounted sensors on a marble cube. In order to improve long-term stability, these sensors have compensation coils wound on quartz tubes, resulting in a sensor drift of only a few nT per year. The marble cube is suspended by two strips of crossed phosphor-bronze working as a cardanic suspension to compensate for pillar tilting that might cause baseline drift. The box containing the electronics is almost magnetic free, and is placed several meters from the sensor to further minimize its effect on the readings. The magnitude and directional response of these instruments needs to be calibrated against accurately known sources.

All modern land-based magnetic observatories use similar instrumentation and process the data in a comparable way to produce similar data products. For a full description, see the INTERMAGNET Web site (www.intermagnet.org). The fundamental measurements recorded are on average 1-min values of the vector components and of scalar intensity. From the 1-min data, hourly, daily, monthly, and annual mean values are produced.

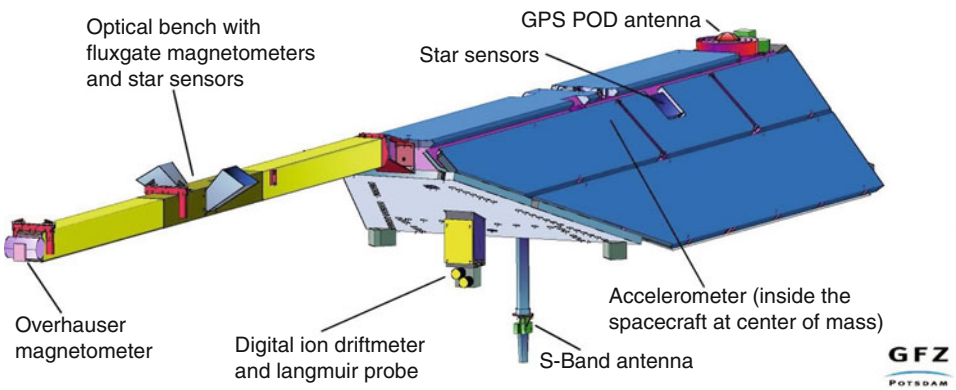
Finally, let us note that data processing consists in a weekly determination of baseline values from absolute measurements. The processing has to include the determination of all systematic errors, and the quality of measurements depends strongly on the experience and accuracy of the observer. Because of this subjective parameter in the final data products and because regular absolute measurements lack in remote areas, several attempts have been made to operate absolute vector field instruments automatically. Among them, one can note a DI-theodolite automatization (van Loo and Rasson, 2006) or a proton vector magnetometer (Auster et al., 2006) for discrete absolute measurements. Efforts have been done to develop a Geomagnetic AUtomated SyStem (GAUSS), instrument-based on rotating a three-component fluxgate magnetometer about well-known axes (Auster et al., 2007). Figure 1d shows this instrument, as mounted in Niemegk magnetic observatory (Germany). Only low mechanical requirements are necessary compared to the standard method with a DI-flux theodolite, but all demands on an absolute measurement are satisfied. The new system aims to be a reliable working automated absolute geomagnetic instrument worldwide.

Near-Earth magnetic field measurements

As stated previously, instruments used in measuring the magnetic field from space are based on the same techniques as the ones on ground, however, the approach to process the magnetic satellite measurements is more complex and different from the one used for magnetic observatories. For this reason, the main part of this contribution is dedicated to measurements processing from Overhauser and fluxgate magnetometers, and the examples refer mainly to the CHAMP satellite (www.gfz-potsdam.de) (Mandea et al., 2010). The instruments on board the CHAMP satellite are shown in Figure 2; those needed to measure the magnetic field are situated on the 4-m boom, far from the satellite body and its magnetic disturbances (see also *Magnetic Methods, Satellite*).

Overhauser magnetometer data processing

An Overhauser magnetometer (OVM), as the one that has been used on Ørsted or CHAMP satellites, is regarded as the magnetic reference instrument. Its output frequency is directly proportional, through the gyromagnetic ratio, to the ambient magnetic field magnitude. Generally, this instrument samples the magnetic field at a rate of 1 Hz



Geomagnetic Field, Measurement Techniques, Figure 2 The CHAMP satellite. The main magnetic instruments are installed on the 4-m boom, the Overhauser magnetometer (at the very end), and the two fluxgate magnetometers and the star cameras (on the optical bench).

with a resolution of 0.1 nT. In order to make the readings a traceable standard, the output frequency is compared to a reference oscillator.

Although the OVM readings are highly accurate, they do not simply reflect the Earth's magnetic field. There are magnetic influences from the spacecraft and the instruments that have to be accounted for. The major processing steps for correcting the data provided by such an instrument are (1) time correction, (2) correction of internal oscillation, (3) correction of additive fields, and (4) correction of multiplicative effects.

1. The input data come with a time stamp, accounting for the time of transmission from the instrument to the onboard data handling system. For the data interpretation it is important to know precisely the epoch at which the reading is valid. Since the satellite is a fast moving platform, readings have to be time-tagged with millisecond precision. The relevant time difference has been determined in ground tests. The epoch of the OVM readings is defined as the reference time to which the next level data of all the other instruments, relevant for the magnetic field products, are resampled.
2. During the next processing step the applied scale factor to the magnetic field readings is checked and corrected. For determining the actual frequency of the internal oscillator the cycles are counted within a gate of 60-GPS s. To convert the Larmor frequency into magnetic field strength, the gyro-magnetic ratio is applied, as recommended by the International Association of Geomagnetism and Aeronomy (IAGA), during the IUGG General Assembly in Vienna, 1991. With this convention, for example, the CHAMP scalar magnetic field readings can be traced back uniquely to the two internationally maintained standards, GPS-second and gyro-magnetic ratio.
3. The remaining corrections concern the disturbances produced by the satellite, and all fields adding to the true ambient are considered together. Contributions come from the remanent magnetization of the spacecraft, from electric currents in the systems, and also from the directional anisotropy of the sensor. These extra fields cannot be simply subtracted. Their influence on the field magnitude has to be determined by finding the projection of the disturbance on the total field vector. For example, the remanent magnetic field of the spacecraft has to be determined in a preflight system test. Since the OVM is mounted at the tip of a long boom, the remaining effect is only of the order of 0.5 nT. There are some stray fields, generated by electric currents, for example, in the solar panels that make effects of less than 0.5 nT. Current readings from the housekeeping data are also needed for this correction. Comparably large are the contributions from the three orthogonal magneto-torquers, which are used for controlling the spacecraft attitude. This contribution can be predicted quite reliably when the currents through the torquer coils are measured.
4. In the last step corrections associated with perturbations proportional to the ambient magnetic field are considered. For CHAMP, these are the permeability of the spacecraft and the cross-talk from the fluxgate magnetometers onto the Overhauser magnetometer. Based on the fairly small sizes of the various correction terms described above one can conclude that the field magnitudes provided are rather reliable, with an absolute accuracy of better than 0.5 nT.

Fluxgate magnetometer data processing

A fluxgate magnetometer (FGM) measures the three components of the magnetic field. These vector field measurements are performed at a rate of 50 Hz with a resolution of 0.1 nT. This higher rate is justified by the significantly larger variability of vector components compared to the fluctuations of the field magnitude. The FGM characteristics are expected to change in response to environmental influences or with time. In order to ensure reliable readings of the vector field components from a multiyear mission, the calibration parameters have to be updated at regular intervals (for CHAMP this is done every 15 days). The in-flight calibration is based on a direct comparison of the FGM readings with the OVM data. The main steps are: (1) time correction, (2) preliminary scaling of data, (3) correction for sensor temperature, (4) correction for torque fields, (5) rescaling of data with actual calibration parameters, and (6) decimation to 1 Hz data.

1. The FGM calibration starts with a proper dating of readings where the delay of the time stamp with regard to the exact epoch of measurement is considered.
2. In the following step the raw data are converted to physical units with the help of a preliminary set of parameters. The measurements are expected to have a bias and need to be scaled. Firstly, the offset vector is subtracted from the FGM readings, measured in engineering units. Thereafter, the results are scaled into nT. For CHAMP, linear, quadratic, and cubic terms are taken into account.
3. Corrections of the environmental influences are crucial. The sensor changes its geometry with the ambient temperature. This has an influence on the scaling factor. Other parameters, bias and sensor orientation, are not affected by temperature. In preflight test the temperature coefficients have been determined, and they amount to about 30 ppm/K for all axes. This effect is corrected using the temperature measurements at the sensor place.
4. During the next step the magnetic fields produced by the spacecraft are considered. For the FGM, there is no need to correct for constant or slowly varying influences, such as the remanent and induced magnetic field of the spacecraft: these effects are accounted for in the FGM calibration parameters. However, disturbances varying over short time have to be corrected directly. An example for that is the magnetic field generated by the current flowing through the torquer coils.

5. FGM data corrected to this level are used for the scalar calibration. This calibration against OVM data results in an improvement of the nine FGM processing parameters (three scale factors, three offset values, three angles between sensors).
6. The processing scheme is applied to the full 50 Hz data set. In order to make vector data consistent with the scalar data, they are resampled to 1 Hz. This resampling is not accomplished by a simple average over the FGM reading within a second. A linear fit to the 100 values centered on the target time is preferred. The new value is then computed from the derived function at the epoch of the related OVM reading. This procedure is performed individually for all three components.

The new data sets are accessible to users. These fully calibrated vector data useful for certain applications, are given in the FGM sensor frame. A transformation is needed to obtain data into the commonly used NEC frame (a local Cartesian frame having its origin at the position of the satellite with the three components pointing to geographic north, east, and to the center of the Earth). A number of rotations have to be performed for this coordinate transformation. The rotation angles from the FGM sensor system to the star camera (ASC) are determined before launch. The rotation into the celestial frame, ICRF, is based on the attitude data. For the rotation from the ICRF into the Earth-fixed, ITRF frame, the Earth's current rotation parameters are used (as provided by the IERS service), where the satellite position is taken from the precise orbit determination. The final rotation into the NEC frame requires just the satellite position. Magnetic field vector data in the NEC frame are the prime data source for all modeling efforts and for various other applications.

In-flight scalar calibration

When processing the FGM magnetic field readings for a multiyear mission, in-flight calibration plays an important role. For that reason the approach used for CHAMP is summarized here.

During the calibration process the readings of the OVM are used as a reference. By comparing field magnitude values with those from FGM, the nine principle parameters of the vector instrument can be determined in a nonlinear inversion. The basic idea is that the OVM and FGM should provide the same values for the magnetic field strength, and any difference can be explained by an improvement of the nine basic FGM parameters. These parameters are expected to be constant over at least one day. A linear expression relates the processed OVM data to the magnetic field components from the FGM through the nine unknowns. In practice, a system of equations is set up making use of all 86,400 daily measurements, from which the nine FGM parameters are determined by least squares. Averages over 15 days are used in the FGM processing for the final scaling of the vector data. After this last processing step the root mean square (rms) value of the difference between OVM and FGM varies from 0.1 to 0.2 nT. This can be regarded as a verification of the used calibration approach.

After more than 10 years of being in orbit, CHAMP satellite gives an interesting example about how the prime FGM parameters have varied over the mission period. There are systematic differences between the variations of the scale factors and the other parameters. The scale values show a monotonic increase over time. The trend is the same for all three components and it can reasonably well be approximated by a logarithmic function. Over the mission time life, the scale factor has changed by seven parts in 10.000, corresponding to an error in magnetic field of some 40 nT. The other parameters show no significant long-term trend, but exhibit periodic variations in phase with the orbit local time. Amplitudes of the offset variations are generally smaller than 0.5 nT. Regarding the sensor stability, variations are confined to angles of 0.001°, which correspond to 3.6 arc s. For these six parameters the in-flight calibration confirms the high stability of the FGM instrument on CHAMP. The synchronization of the deduced variations with the orbital local time suggests that other, not corrected perturbations of the measurements leak into these parameters. However, this does not affect the validity of the final data.

Summary

Here, we concentrate primarily, but not exclusively, on geomagnetic instruments and techniques used in a magnetic observatory settings or in a near-Earth magnetic observation. We have briefly described the basic instruments, their mode of operation, their relative and/or absolute accuracy, and the related sources of error. Currently, the magnetometry is a mature discipline, so something equivalent to the invention of the proton magnetometer would not be achieved very soon. It is more likely that some instrumental advances will be made toward the aim to stabilize vector magnetometers. Advances will be made in power reduction, data storage, and telemetry with the goals of increased automation and cutting costs. The coming age of the automated absolute instruments makes it possible to consider once again the possibility of a true underwater observatory.

The future needs of the scientific and commercial users will include ground- and space-based measurements. Increasing resolution and timing accuracy will be crucial. However, to achieve these requirements will demand more efforts from instrument makers and spacecraft designers.

The Earth's magnetic field will remain under observation with the ESA's forthcoming Swarm mission (Friis-Christensen et al., 2009). Three satellites will be launched in 2011 and will measure the magnetic field and its variations far more accurately than ever before. However, a comprehensive separation and understanding of the internal and external processes contributing to the Earth's magnetic fields is possible only by joint analysis of

satellite and ground-based data, with all difficulties arising in combining such different datasets. Continuous space-borne and ground-based monitoring of the magnetic field aims to address such needs.

Bibliography

Auster, V., Hillenmaier, O., Kroth, R., and Weidemann, M., 2006. Proton magnetometer development. *XIIth IAGA Workshop on Geomagnetic Observatory Instruments, Data Acquisition and Processing*, Abstract volume, **56**.

Auster, H. U., Mandea, M., Hemshorn, A., Pulz, E., and Korte, M., 2007. Automation of absolute measurement of the geomagnetic field. *Earth Planets Space*, **59**, 20131014.

Friis-Christensen, E., LŁuhr, H., Hulot, G., Haagmans, R., and Purucker, M., 2009. Geomagnetic research from space. *Eos, Transactions, AGU,* **90**(25), 213–215.

Lühr, H., Korte, M., and Mandea, M., 2009. The recent Geomagnetic Field and its Variations. In Glassmeier, K.-H., Soffel, H., and Negendank, J. (eds.), *Geomagnetic Field Variations*. Springer, ISBN 978-3-540-76938-5, pp. 25–64.

Mandea, M., and Purucker, M., 2005. Observing, modeling, and interpreting magnetic fields of the solid earth. *Surveys in Geophysics*, **26**, 415–459, doi:10.1007/s10712-005-3857-x.

Mandea, M., Holschneider, M., Lesur, V., and Lühr, H., 2010. The Earth's Magnetic Field at the CHAMP Satellite Epoch, in System Earth via Geodetic-Geophysical Space Techniques. In Flechtner, F., Gruber, T., Güntner, A., Mandea, M., Rothacher, M., Schöne, T., Wickert, J. (eds.), Berlin: Springer, pp. 475–526.

Van Loo, S. A., and Rasson, J. L., 2006. Presentation of the prototype of an automatic DI-Flux. *XIIth IAGA Workshop on Geomagnetic Observatory Instruments, Data Acquisition and Processing*, Abstract volume, **21**.

Cross-references

Core Dynamo
Geomagnetic Field, Global Pattern
Geomagnetic Field, IGRF
Geomagnetic Field, Secular Variation
Magnetic Anomalies, Interpretation
Magnetic Methods, Satellite
Magnetic, Global Anomaly Map
Magnetometers

GEOMAGNETIC FIELD, POLARITY REVERSALS

Carlo Laj
Laboratoire des Sciences du Climat, Unité mixte CEA-CNRS-UVSQ, Gif-sur-Yvette, France

Introduction: The discovery of geomagnetic reversals

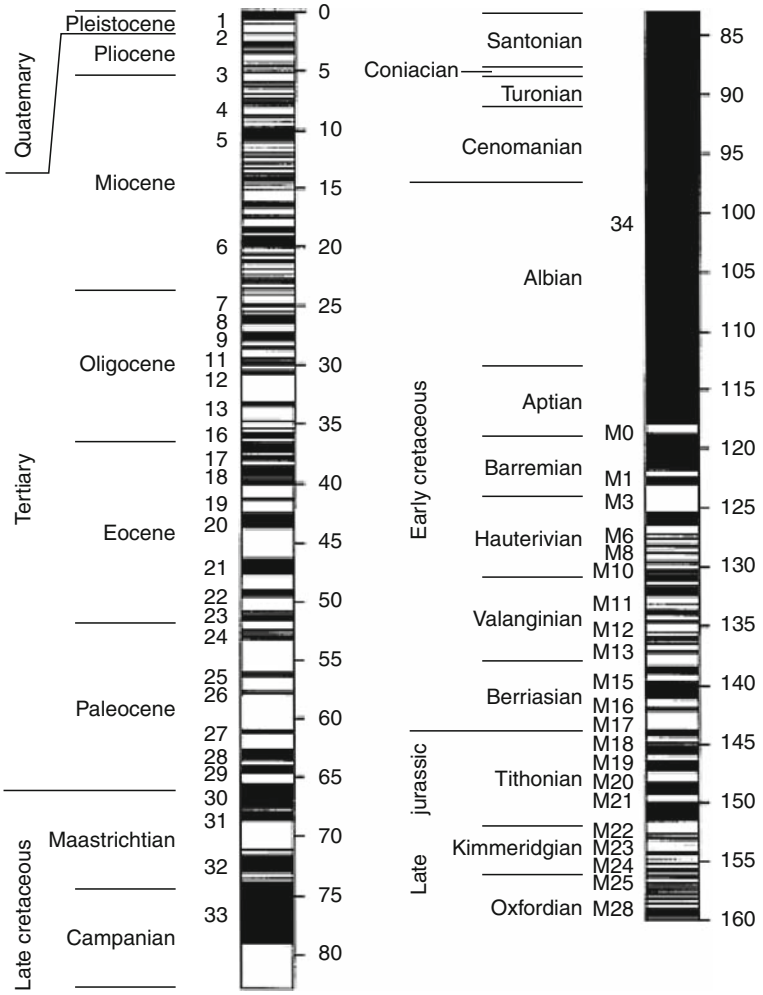
Bernard Brunhes (1906) was the first to measure magnetization directions in rocks that were approximately antiparallel to the present Earth's field. Brunhes (1906) recorded magnetizations in baked sedimentary rocks that were aligned with reverse magnetization directions in overlying Miocene lavas from central France (Puy de Dome). In so doing, Brunhes (1906) made first use of a field test for primary thermal remanent magnetization (TRM) that is now referred to as the "baked contact" test. Matuyama (1929) was the first to attribute reverse magnetizations in (volcanic) rocks from Japan and China to reversal of geomagnetic polarity, and to differentiate mainly Pleistocene lavas from mainly Pliocene lavas based on the polarity of the magnetization. In this respect, Matuyama (1929) was the first person to use the sequence of geomagnetic reversals as a means of ordering rock sequences.

The reality of geomagnetic reversals was then progressively established with the studies of Hospers (1951, 1953) in Iceland, and Roche (1950, 1951, 1956) in the Massif Central of France. The work of Hospers on Icelandic lavas was augmented by Rutten and Wensink (1960) and Wensink (1966) who subdivided Pliocene-Pleistocene lavas in Iceland into three polarity zones from young to old: N-R-N. Magnetic remanence measurements on basaltic lavas combined with K/Ar dating, pioneered by Cox et al. (1963) and McDougall and Tarling (1963a, 1963b, 1964), resulted in the beginning of development of the modern geomagnetic polarity timescale (GPTS). These studies, and those that followed in the mid-1960s, established that rocks of the same age carry the same magnetization polarity, at least for the last few million years. The basalt sampling sites were scattered over the globe. Polarity zones were linked by their K/Ar ages, and were usually not in stratigraphic superposition. Doell and Dalrymple (1966) designated the long intervals of geomagnetic polarity of the last 5 Myrs as magnetic epochs, and named them after pioneers of geomagnetism (Brunhes, Matuyama, Gauss, and Gilbert).

Then, the discovery of marine magnetic anomalies confirmed seafloor spreading (Vine and Matthews, 1963), and the GPTS was extended to older times (Vine, 1966; Heirtzler et al., 1968; Lowrie and Alvarez, 1981). Since then, the succession of polarity intervals has been extensively studied and used to construct magnetostratigraphic timescales linking biostratigraphies, isotope stratigraphies, and absolute ages (see Opdyke and Channell, 1996,"Magnetic stratigraphy", for a review).

The geomagnetic polarity timescale

The fit of the land-derived polarity timescale, from paleomagnetic and K/Ar studies of exposed basalts, with the polarity record emerging from marine magnetic anomalies (MMA) (Vine and Matthews, 1963; Vine, 1966; Pitman and Heirtzler, 1966; Heirtzler et al., 1968) resulted in a convincing argument for synchronous global geomagnetic polarity reversals, thereby attributing them to the main axial dipole. This intense research effort has culminated with the work of Cande and Kent (1995) based on the Marine Magnetic Anomalies record, using the South Atlantic as the fundamental template with inserts from faster spreading centers in the Indian and Pacific oceans. Cande and Kent (1995) adopted the astrochronological age estimates for the Pliocene-Pleistocene polarity reversals and fixed the age tie-point at the Cretaceous-Tertiary boundary at 65 Ma (Figure 1).



Geomagnetic Field, Polarity Reversals, Figure 1 The geomagnetic polarity timescale for the last 160 My. Numbers indicate the marine magnetic anomalies.

The changing frequency over time and the duration of geomagnetic reversals

Examination of the geomagnetic polarity timescale reveals some of the major characteristics of geomagnetic reversals. First, a clear characteristic is the widely different duration of periods of stable polarity separating geomagnetic reversals. Long periods of time during which there were no reversals have been clearly identified, such as the Cretaceous Long Normal Chron (also called the Cretaceous Superchron, about 118–183 million years ago) or the Kiaman Superchron, which lasted approximately from the late Carboniferous to the end of the Permian.

The frequency of reversals decreased prior to the Cretaceous Superchron, and slowly increased over the 80 million years from the end of this Superchron to the present. It has been suggested (Mazaud et al., 1983; Mazaud and Laj, 1991) that in this last period the frequency does not increase randomly but presents a dominant frequency corresponding to a period of approximately 15 million years. This suggestion has, however, been received with some skepticism (McFadden, 1987). The most recent analyses seem to converge toward a nonstationary Poisson process or at least a gamma process with k only slightly greater than 1 (Lowrie and Kent, 2004).

Another characteristic, revealed from the very beginning by the studies of reversals, is that the Earth's magnetic field strength drops to low levels during polarity reversals. There has been some debate on whether this drop is symmetrical or not on each side of the reversal. Valet and Meynadier (1993) suggested that the dipole intensity slowly decreases over a long period of time preceding a reversal and rapidly increases immediately after (the so-called sawtooth hypothesis). These authors also suggested that there is a positive correlation between the magnitude of the increase in intensity and the length of the subsequent polarity interval.

The sawtooth pattern, however, is not present in many sedimentary records (Tauxe and Shackleton, 1994; Tauxe

and Hartl, 1997; Laj et al., 1996; Channell and Kleiven, 2000; Channell et al., 2009). Interpretations of the sawtooth pattern, other than true geomagnetic behavior, have been given, either assuming sedimentary postdepositional artifacts (Mazaud, 1996) or the presence of a cumulative viscous component not removed by alternating field demagnetization (Kok and Tauxe, 1996a, b). Moreover, McFadden and Merrill (1998) have strongly questioned the suggestion that the duration of polarity intervals could be determined by the amplitude of the field recovery after the reversal. The fast recovery of the field intensity after the reversal does not appear either in a recent reconstruction of the Brunhes–Matuyama transition, based on four independent records, where the pre- and post-transitional intensity records appear symmetric about the reversal (Leonhardt and Fabian, 2007). In summary, the experimental evidence for the sawtooth hypothesis is controversial.

Also clearly apparent from the geomagnetic polarity timescale is the extremely short duration of polarity reversals compared to periods of stable polarity. The time it takes for the reversal to happen has been roughly estimated from the early works (Harrison and Somayajulu, 1966; Cox et al., 1963, 1964; Cox, 1969) to be between 10^3 and 10^4 years, with some estimates up to 28,000 years. In a recent article, Clement (2004) presents an analysis of available sedimentary records and shows that these records yield an average estimate of 7,000 years for the duration of a reversal. Moreover, the duration varies with site latitude, with a shorter duration observed at low-latitude sites. This short duration has been one of the most delicate factors to overcome in studying the detailed structure of reversals.

The morphology of geomagnetic reversals

Because of its importance to dynamo theory, the morphology of the geomagnetic field during reversals and whether or not it displays statistical asymmetries has been a subject of continuous research for many years. Records of geomagnetic field changes during a polarity transition have, however, been difficult to obtain, not only because of their short duration mentioned above, but also because the magnetizations of the rocks is weak due to the low intensity of the transitional field. It is only with the advent of sensitive magnetometers that the number of published transitional records has rapidly increased.

In probably the most famous of the early records, Hillhouse and Cox (1976) showed that the Virtual Geomagnetic Pole (VGP) (The Virtual Geomagnetic Pole is defined as the pole of the geocentered axial dipolar field that would give the observed direction of the field at the observation site. If the VGPs observed at several sites at a given time coincide, then the field has a dipolar structure.) path for the Brunhes–Matuyama transition recorded in sediments from Lake Tecopa (California) was different from the one previously obtained by Niitsuma (1971) from the Boso Peninsula in Japan, thereby providing evidence that the transitional field is not dipolar. Although in a more recent article Valet et al. (1988a) showed that the AF demagnetization used by Hillhouse and Cox (1976) did not completely remove a strong overprint, the conclusion that the transitional field was not dipolar was maintained.

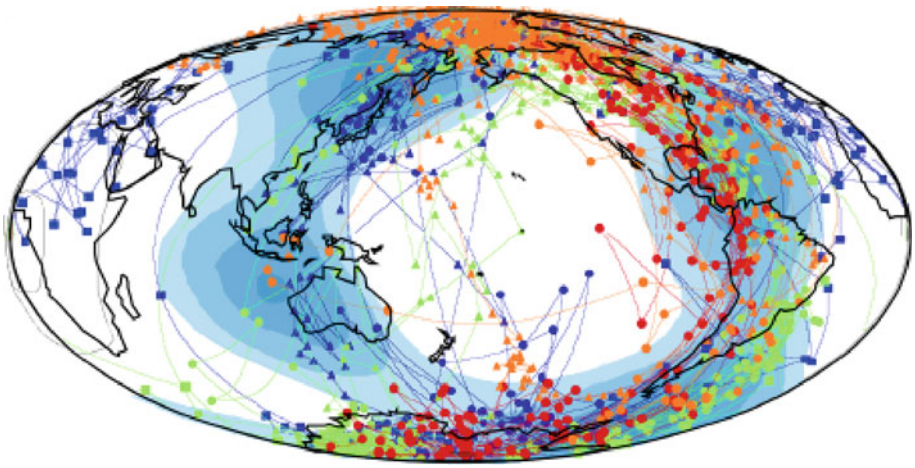
Another famous record was obtained from a Miocene lava sequence at the Steens Mountain (Mankinen et al., 1985; Prévot et al., 1985). The record displays several phases with large swings, one of which occurring at the extremely rapid rate of $50 \pm 21^\circ$ per year, while the intensity change was $6{,}700 \pm 2{,}700$ μT per year, about 10–50 times larger than the maximum rate of change observed for the non-dipole field in historical records. These amazing rates of change are still a matter of discussion among geophysicists.

Other transitional records were obtained from lava sequences in French Polynesia (Roperch and Chauvin, 1987; Roperch and Duncan, 1990; Chauvin et al., 1990). These records appear to indicate that a zonal field dominates the beginning of the transition, then non-dipolar terms emerge without any preferred axisymmetry, consistent with an earlier suggestion by Hoffman (1982).

Transitional records from sequences of relatively high accumulation rate marls in Western Crete (Greece), were reported in a series of papers by Valet and Laj (1981, 1984) and by Valet et al. (1983, 1986, 1988b). The results from multiple records indicated that the reversal process observed for four sequential transitions remained unchanged for about 1.3 Ma. In one record, obtained from a particularly high sedimentation rate section, fast directional fluctuations, reminiscent of the Steens Mountain record, were observed whose amplitude significantly increases during the transition. Slight smoothing of the data reveals that almost periodic fluctuations occurred on a longer timescale during the transition (Valet et al., 1986).

Two landmark papers were published almost simultaneously and independently in 1991. Clement (1991) examined records of the Brunhes–Matuyama transition obtained from the northern, equatorial, and southern latitudes in the Atlantic sector. The VGP paths from the mid-latitudes in the northern and southern hemispheres are nearly coincident and extend over the Americas. Paths from the equatorial site track almost antipodal to the other paths. These results suggest that during the reversal VGPs tend to fall along two longitudinal bands.

From a compilation of all the available sedimentary records from several sites and spanning about 12 million years, Laj et al. (1991) suggested that reversals and excursions exhibit geometric regularity: the VGP paths of these transitions tend to fall along either American or Asian-Australian longitudes, that is, the same as for the Brunhes–Matuyama reversal (Figure 2). As remarked by the authors a persistent geometry over 12 Ma has important geomagnetic consequences : since the duration



Geomagnetic Field, Polarity Reversals, Figure 2 The compilation of sedimentary records for the last 12 My (Laj et al., 1991) indicating the existence of preferred longitudinal bands for transitional VGPs. The blue shaded zones are characterized by fast seismic velocities in the lower mantle.

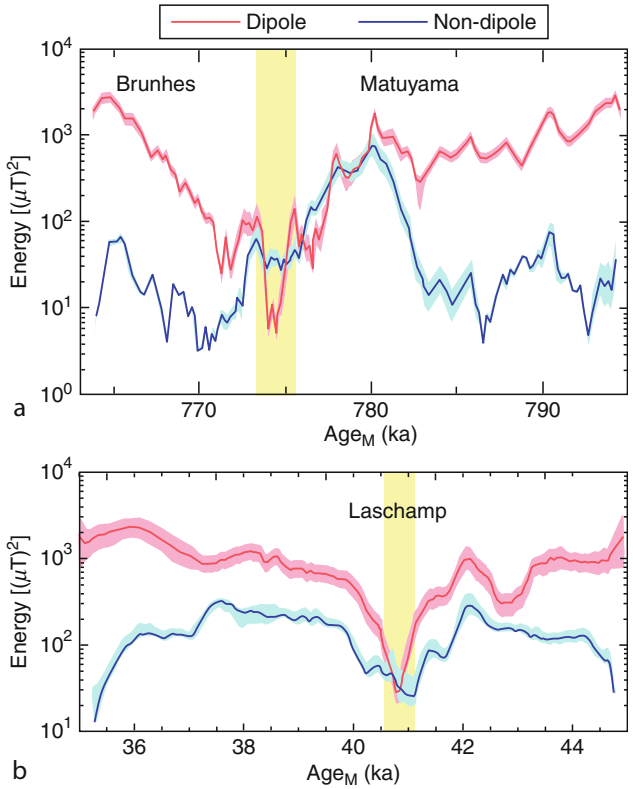
between transitions is longer than both the core's convective and magnetic diffusion timescales, the most plausible mechanism by which the core could retain a memory of previous transitions, particularly preferred longitudinal bands, is through some form of core-mantle coupling. This argument was reinforced by Laj et al. (1991) who noted the apparent correlation between preferred longitudes and lower mantle seismic velocity variations, possibly arising from thermal heterogeneities.

The suggestion of preferred longitudinal bands for transitional VGPs has, however, been met with some skepticism, some authors questioning the adequacy of the geographical distribution of the sites (Valet et al., 1992), or the reliability of transitional records from sediments (Langereis et al., 1992; Barton and McFadden, 1996). Statistical analyses of the distribution of path have not given a clear answer (Weeks et al., 1992; McFadden et al., 1993; Quidelleur and Valet, 1994). An initial analysis of transitional records from lavas (which are devoid of many of the problems of sedimentary magnetizations) appeared to indicate that transitional VGPs are statistically axi-symmetric (Prévot and Camps, 1993). However, a more recent analysis by Love (1998, 2000) using a mathematical approach in which the records are normalized taking into account differences in recording density of the paleomagnetic directions, has concluded that the volcanic transitional VGPs indeed show a significant tendency to fall along American and Asian-Australian longitudes, consistent with the sedimentary data. The existence of preferred longitudinal bands for transitional VGPs must, therefore, be considered as a realistic suggestion.

Alongside with the observations of records of geomagnetic reversals, phenomenological models have been developed to account for at least the main characteristics of the records. As early as 1976, Hillhouse and Cox (1976) suggested that if the usual non-dipole drift remained unchanged during the reversal, then one should observe large longitudinal swings in the reversal PGV paths. As this was not observed, these authors proposed that an invariant component dominates the field when the usual dipole has vanished (the so-called standing field hypothesis). Subsequently, Hoffman and Fuller (1978) and Hoffman (1979) proposed a model in which reversals originate in particular regions of the core and progressively extend to other regions (the so-called flooding model). None of these models was, however, capable of predicting VGP paths from particular sites on the Earth.

A significant step forward has been made recently, with the development by Leonhardt and Fabian (2007) of a Bayesian inversion method devised to reconstruct the spherical harmonic contents of the geomagnetic field during a reversal from paleomagnetic data. The validity of this inversion technique is proven, in particular, by iteratively combining four geographically distributed high quality paleomagnetic records of the Brunhes–Matuyama reversal into a single scenario without assuming an a priori common age model. The obtained results successfully describe most independent Brunhes–Matuyama transitional records.

Leonhardt and Fabian (2007) were able to discuss many questions about the transitional field, which had been controversially discussed (see, e.g., Merrill et al. (1996)) so far: primarily the question of contribution of dipole versus non-dipole fields during the reversal process. The results of their modeling indicates a strong increase of non-dipolar energy a few kyrs before the reversal. The non-dipole energy prevails on the dipole some 5 kyrs before the dipole minimum, after which both the non-dipole and dipolar energies decrease. Following the minimum in dipolar energy, the non-dipolar energy



Geomagnetic Field, Polarity Reversals, Figure 3 (**a**) Evolution of the dipole and non-dipole energies during the Matuyama–Brunhes polarity transition. (**b**) Evolution of the dipole and non-dipole energies during the Laschamp excursion.

drops further, while the dipolar energy increases first sharply then progressively recovers (Figure 3a). The model also shows that, during the reversal, the inverse flux penetrates the tangent cylinder (the imaginary cylinder aligned with the Earth's rotation axis and tangent to the inner core).

Dynamo mechanisms and reversals

For many years, scientific opinion has been divided on what causes geomagnetic reversals. Initial suggestions have been made that geomagnetic reversals are not due to spontaneous processes in the Earth's dynamo, but rather that they are triggered by external factors (such as arrival of continental slabs near the core-mantle boundary as a result of plate tectonics, or even mantle-core shear forces linked to external impacts).

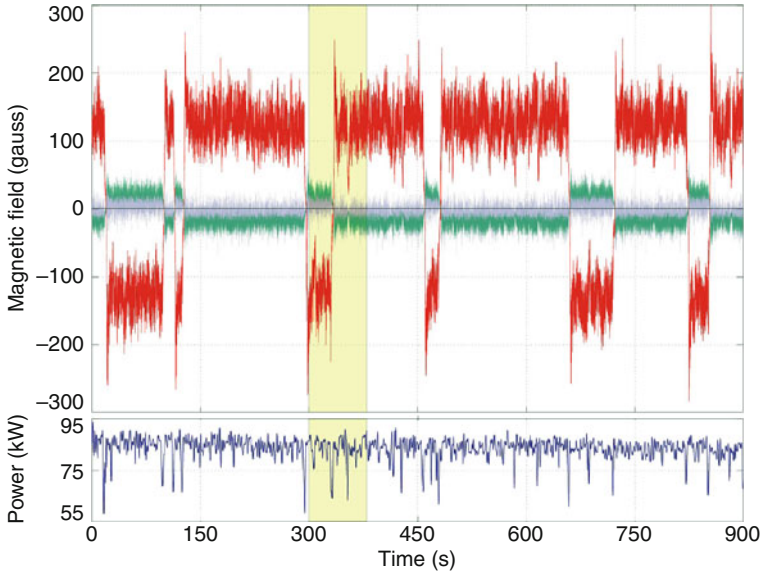
Today, most geophysicists believe that reversals are inherent aspects of the dynamo, which generates the geomagnetic field. Recent developments in numerical dynamo models, which take advantage of modern supercomputers, and also an experiment with a laboratory dynamo, point strongly that this is the only sustainable mechanism. In these models the equations describing the time-dependent thermal, compositional, velocity and magnetic field are solved simultaneously in the three-dimensional spherical geometry appropriate to the core. There is no a priori prescription for reversals, so, if they happen, they are an intrinsic property of the mechanisms of the geodynamo.

The first of these numerical models aiming to simulate Earth-like characteristics was published by Glatzmaier and Roberts (1995). This three-dimensional, self-consistent model was able to maintain a magnetic field for over 40,000 years. The model, considers a finitely conducting solid inner core, in agreement with findings by Hollerbach and Jones (1993) that a solid inner core tends to stabilize the geodynamo. Over the period of 40,000 year the model undergoes several polarity excursions and finally a reversal is observed. The model, therefore, shares some similarities with real reversals of the geomagnetic field, and therefore may provide insight into the mechanisms of geomagnetic reversals.

In this respect, Glatzmaier et al. (1999) and Coe et al. (2000) have run numerical simulations of the Glatzmaier–Roberts dynamo using a variety of thermal boundary conditions. The model displays a range of behavior that resembles records of real reversals of the Earth's magnetic field. Reversals occurred during both the homogeneous and tomographic simulation (in this last case the heat flux at the Core-Mantle Boundary is patterned after a seismic velocity model from tomographic investigation of the lower mantle). All reversals appear spontaneously, with no external triggering required. The tomographic model of Glatzmaier et al. (1999) also appears to offer some support for the hypothesis that VGPs during reversals correlate with areas of higher than average heat flux.

The question of how well the geodynamo model represents the Earth's field is of course fundamental. Indeed, if the Earth's values for the radius and the rotation of the core are used, then, owing to computation limitations, viscous diffusivity values at least three to four orders of magnitude larger than those of the Earth must be used. So, while some modelers have argued that the field produced by models should be fairly realistic, other modelers are less confident that the results can be directly extrapolated to the Earth, even at the large scales. In the author's opinion, numerical models have very largely improved our understanding of the mechanisms inherent to the production of the Earth's magnetic field.

Finally, a benchmark experiment showing how magnetic field reversals can be generated without any external triggering in a turbulent conducting field has been obtained recently in a laboratory environment, the so-called VKS experiment (this acronym stands for Von Karman Sodium) (Berhanu et al. (2007)). In 2000 experiments in Riga (Gailitis et al., 2001) and Karlsruhe (Stieglitz and Müller, 2001) showed that fluid dynamo could be generated by fluid sodium flows, when a favorable geometry is imposed to the flow. In the more recent VKS experiment, a turbulent flow is produced in a copper cylinder filled with liquid sodium whose flow



Geomagnetic Field, Polarity Reversals, Figure 4 Evolution of the magnetic field measured inside the flow vessel during the VKS experiment. The main component (red) is the azimuthal one. Note that all components decay to zero at a reversal. The bottom graph shows the evolution of the power driving the flow (Adapted from Berhanu et al., 2007).

is created by two counter-rotating impellers. When the speed of the impellers exceeds a critical value, the flow spontaneously generates a magnetic field that is self-sustained by the flow itself. This field undergoes large fluctuations, originating from the high level of turbulence. The most spectacular result was observed when the two impellers were counter-rotated at different speeds: in this case the temporal evolution of the field is characterized by a series of erratic reversals of polarity. These reveal remarkable similarities with the Earth's magnetic field, in the sense that the duration of the transitions is very short compared to the periods of stable polarity. In addition, excursions are also present, during which the field decays to very low values before growing again without a change in polarity (Figure 4). Although similarity to the Earth's magnetic field should not be overstressed, this laboratory experiment clearly shows that reversals can appear spontaneously, as a consequence of the complexity of the fluid flow, without any necessity for an external triggering.

Outlook: The future of geomagnetic reversals

The geomagnetic dipole has decreased in recent historical times at a rate of 5% per century, which is about 10–12 times the rate of free Ohmic decay of the geomagnetic dipole field in the core. Additional evidence for this rapid decrease is the poleward migration of large patches of reverse magnetic field at the core-mantle boundary that have largely contributed to the historical dipole field drop (Olson and Amit, 2006). Therefore, this is a very rapid change for the geodynamo, which has led to speculations about the possibility of an impending field reversal (Constable and Korte (2006); Olson and Amit (2006)). Because an intensity drop is also associated with excursions this second possibility should also be considered.

The inverse model of Leonhardt and Fabian (2007) may allow to have a deeper insight on this scenario. Indeed, when applied to the study of the Laschamp excursion (Leonhardt et al., 2009) the model indicates that both dipolar and non-dipolar energies at the Earth's surface decrease at approximately the same time with similar rates. The dipolar energy reaches its minimum slightly later than the non-dipolar energy. Then, for a brief interval the non-dipolar energy prevails, contrary to the Brunhes–Matuyama reversal discussed above where the non-dipole energy prevails on the dipole some 5 kyrs before the dipole minimum, after which both the non-dipole and dipolar energies decrease (Figure 3b). The time evolution of the dipole versus non-dipole fields is different for a reversal or an excursion and could, therefore, be an indicator for an incoming reversal or excursion.

At present, however, the non-dipole field energy is about one order of magnitude smaller than the dipolar term. This latter is still larger than the dipole energy calculated before either a reversal or an excursion using the Leonhardt and Fabian approach. So despite the decreasing trend of dipole energy, there is no indication in the present field that would point toward either type of instability. Together with Olson and Amit (2006), Leonhardt et al. (2009) therefore consider that it is not justified to consider the present decrease in dipole energy and an indication of an imminent reversal. We will have to wait some more time to know whether we are heading toward a reversal, an excursion, or whether we are simply undergoing an intensity fluctuation.

Acknowledgments

I thank Alain Mazaud for helpful discussions during the writing of the manuscript.

Bibliography

Barton, C. E., and McFadden, P. L., 1996. Inclination shallowing and preferred transitional VGP paths. *Earth and Planetary Science Letters*, **140**, 147–157.

Berhanu, M., Monchaux, R., Fauve, S., Mordant, N., Pétéelis, F., Chiffaudel, A., Daviaud, F., Dubrulle, B., Marié, L., Ravelet, F., Bourgoin, M., Odier, P. H., Pinton, J.-F., and Volk, R., 2007. Magnetic field reversals in an experimental turbulent dynamo. *Europhysics Letters*, **77**, 59001–59005.

Brunhes, B., 1906. Recherches sur la direction de l'aimantation des roches volcaniques. *Journal de Physique*, **V**, 705–724.

Cande, S. C., and Kent, D. V., 1995. Revised calibration of the geomagnetic polarity timescale for the late cretaceous and cenozoic. *Journal of Geophysical Research*, **100**, 6093–6095.

Channell, J. E. T., and Kleiven, H. F., 2000. Geomagnetic palaeointensities and astrochronological ages for the Matuyama-Brunhes boundary and the boundaries of the Jaramillo Subchron: palaeomagnetic and oxygen isotope records from ODP Site 983. *Philosophical Transactions of the Royal Society A*, **358**, 1027–1047.

Channell, J. E. T., Hodell, D. A., and Lehman, B., 1997. Relative geomagnetic paleointensity and $\delta^{18}O$ at ODP site 983 (Gardar Drift, North Atlantic) since 350 ka. *Earth and Planetary Science Letters*, **153**, 103–118.

Channell, J. E. T., Xuan, C., and Hodell, D. A., 2009. Stacking paleointensity and oxygen isotope data for the last 1.5 Myr (PISO-1500). *Earth and Planetary Science Letters*, **283**, 14–23.

Chauvin, A., Duncan, R. A., Bonhommet, N., and Levi, S., 1989. Paleointensity of the Earth's magnetic field and K-Ar dating of the Louchadière volcanic flow (central France): new evidence for the Laschamp Excursion. *Geophysical Research Letters*, **16**, 1189–1192.

Chauvin, A., Roperch, P., and Duncan, R. A., 1990. Records of geomagnetic reversals from volcanic islands of French polynesia, 2-paleomagnetic study of a flow sequence (1.2 to 0.6 Ma) from the Island of Tahiti and discussion of reversal models. *Journal of Geophysical Research*, **95**, 2727–2752.

Clement, B. M., 1991. Geographical distribution of transitional V. G.P.'s: evidence for non-zonal equatorial symmetry during the Matuyama-Brunhes geomagnetic reversal. *Earth and Planetary Science Letters*, **104**, 48–58.

Clement, B. M., 2004. Dependence of the duration of geomagnetic polarity reversals on site latitude. *Nature*, **428**, 637–640.

Coe, R. S., Hongre, L., and Glatzmaier, G. A., 2000. An examination of simulated geomagnetic reversals from a paleomagnetic perspective. *Philosophical Transactions of the Royal Society A*, **358**, 1141–1170.

Constable, C. G., and Korte, M., 2006. Is the Earth's magnetic field reversing? *Earth and Planetary Science Letters*, **246**, 1–16.

Cox, A., 1969. Geomagnetic reversals. *Science*, **163**, 237–245.

Cox, A., Doell, R. R., and Dalrymple, G. B., 1963. Geomagnetic polarity epochs and Pleistocene geochonometry. *Nature*, **198**, 1049–1051.

Cox, A., Doell, R. R., and Dalrymple, G. B., 1964. Reversals of the Earth's magnetic field. *Science*, **144**, 1537–1543.

Doell, R. R., and Dalrymple, G. B., 1966. Geomagnetic polarity epochs: a new polarity event and the age of the Brunhes-Matuyama boundary. *Science*, **152**, 1060–1061.

Gailitis, A., Lielausis, O., Platacis, E., Dement'ev, S., Cifersons, A., Gerbeth, G., Gundrum, T., Stefani, F., Christen, M., and Will, G., 2001. Magnetic field saturation in the Riga dynamo experiment. *Physical Review Letters*, **86**, 3024–3027.

Glatzmaier, G. A., and Roberts, P. H., 1995. A three-dimensional self-consistent computer simulation of a geomagnetic field reversal. *Nature*, **377**, 203–209.

Glatzmaier, G. A., Coe, R. S., Hongre, L., and Roberts, P. H., 1999. The role of the Earth's mantle in controlling the frequency of geomagnetic reversals. *Nature*, **401**, 885–890.

Gubbins, D., 1999. The distinction between geomagnetic excursions and reversals. *Geophysical Journal International*, **137**, F1–F3.

Harrison, C. G. A., and Somayajulu, B. L. K., 1966. Behaviour of the earth's magnetic field during a reversal. *Nature*, **212**, 1193–1195.

Heirtzler, J. R., Dickson, G. O., Herron, E. M., Pittman, W. C., III, and LePichon, X., 1968. Marine magnetic anomalies, geomagnetic field reversal and motions of the ocean floor and continents. *Journal of Geophysical Research*, **73**, 2119–2136.

Hillhouse, J., and Cox, A., 1976. Brunhes-Matuyama polarity transition. *Earth and Planetary Science Letters*, **29**, 51–64.

Hoffman, K. A., 1979. Behaviour of the geodynamo during a reversal: a phenomenological model. *Earth and Planetary Science Letters*, **44**, 7–17.

Hoffman, K. A., 1982. The testing of geomagnetic reversal models/ récent developments. *Philosophical Transactions of the Royal Society of London. Series A*, **306**, 147–159.

Hoffman, K. A., and Fuller, M., 1978. Polarity transition records and the geomagnetic dynamo. *Nature*, **273**, 715–718.

Hollerbach, R., and Jones, C. A., 1993. Influence of the Earth's inner core on geomagnetic fluctuations and reversals. *Nature*, **365**, 541–543.

Hospers, J., 1951. Remanent magnetization of rocks and the history of the geomagnetic field. *Nature*, **168**, 1111–1112.

Hospers, J., Reversals of the main geomagnetic field I, II, and III. Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen Series C, 56, 467–491; 57, 112–121, 1953–1954.

Kok, Y. S., and Tauxe, L., 1996a. Saw-toothed pattern of relative paleointensity records and cumulative viscous remanence. *Earth and Planetary Science Letters*, **137**, 95–100.

Kok, Y. S., and Tauxe, L., 1996b. Saw-toothed pattern of sedimentary paleointensity records explained by cumulative viscous rémanence. *Earth and Planetary Science Letters*, **144**, 9–14.

Laj, C., Mazaud, A., Fuller, M., and Herrero-Bervera, E., 1991. Geomagnetic reversal paths. *Nature*, **351**, 447.

Laj, C., Kissel, C., Garnier, F., and Herrero-Bervera, E., 1996. Relative geomagnetic field intensity and reversals for the last 1.8 My from a central Pacific core. *Geophysical Research Letters*, **23**, 3393–3396.

Langereis, C. G., Van Hoof, A. A. M., and Rochette, P., 1992. Longitudinal confinement of geomagnetic reversal paths: as a possible sedimentary artefact. *Nature*, **358**, 226–230.

Leonhardt, R., and Fabian, K., 2007. Paleomagnetic recontruction of the global geomagnetic field évolution during the Matuyama/Brunhes transition: Iterative Bayesian inversion and independent vérification. *Earth and Planetary Science Letters*, **253**, 172–195.

Leonhardt, R., Fabian, K., Winklhofer, M., Ferk, A., Laj, C., and Kissel, C., 2009. Geomagnetic field évolution during the Lascamp excursion. *Earth and Planetary Science Letters*, **278**, 87–95.

Love, J. J., 1998. Paleomagnetic volcanic data and geometric regularity of reversals and excursions. *Journal of Geophysical Research*, **103**, 12435–12452.

Love, J. J., 2000. Statistical assessment of preferred transitional VGP longitudes based on paleomagnetic lava data. *Geophysical Journal International*, **140**, 211–221.

Lowrie, W., and Alvarez, W., 1981. One hundred million years of geomagnetic polarity history. *Geology*, **9**, 392–397.
Lowrie, W., and Kent, D. V., 2004. Geomagnetic polarity timescales and reversal frequency regimes. In Channell, J. E. T., et al. (eds.), *Timescales of the Geomagnetic Field*. AGU Geophysical Monograph Series **145**, 117–129.
Mankinen, E. A., Prévot, M., Grommé, C. S., and Coe, R. S., 1985. The Steens Mountain (Oregon) geomagnetic polarity transition. 1. Directional history, duration of the episodes, and rock magnetism. *Journal of Geophysical Research*, **90**, 393–416.
Matuyama, M., 1929. On the direction of magnetization of basalts in Japan, Tyosen and Manchuria. *Proceedings of the Imperial Academy*, **5**, 203–205 (Tokyo).
Mazaud, A., 1996. Sawtooth variation in magnetic intensity profiles and delayed acquisition of magnetization in deep se acores. *Earth and Planetary Science Letters*, **139**, 379–386.
Mazaud, A., and Laj, C., 1991. The 15 m.y. geomagnetic reversal periodicity: a quantitative test. *Earth and Planetary Science Letters*, **107**, 689–696.
Mazaud, A., Laj, C., De Seze, L., and Verosub, K. L., 1983. 15-Myr periodicity in the frequency of geomagnetic reversals since 100 Myr. *Nature*, **304**, 328–330.
McDougall, I., and Tarling, D. H., 1963a. Dating of reversals of the Earth's magnetic field. *Nature*, **198**, 1012–1013.
McDougall, I., and Tarling, D. H., 1963b. Dating of polarity zones in the Hawaiian Islands. *Nature*, **200**, 54–56.
McDougall, I., and Tarling, D. H., 1964. Dating geomagnetic polarity zones. *Nature*, **202**, 171–172.
McFadden, P. L., 1987. Comment on « A periodicity of magnetic reversals?'. *Nature*, **330**, 27.
McFadden, P. L., and Merrill, R. T., 1998. Sawtooth paleointensity and reversals of the geomagnetic field. *Physics of the Earth and Planetary Interiors*, **103**, 247–252.
McFadden, P. L., Barton, C. L. E., and Merrill, R. T., 1993. Do virtual geomagneitc poles follow preferential paths during geomagneitc reversals ? *Nature*, **361**, 342–344.
Merrill, R. T., McElhinny, M., and McFadden, P. L., 1996. *The Magnetic Field of the Earth, International Geophysics Series*. San Diego: Academic, Vol. 63, p. 531.
Meynadier, L., Valet, J.-P., Weeks, R., Shackleton, N. J., and Hagee, V. L., 1992. Relative geomagnetic intensity of the field during the last 140 ka. *Earth and Planetary Science Letters*, **114**, 39–57.
Meynadier, L., Valet, J.-P., and Shackleton, N. J., 1995. Relative geomagnetic intensity during the last 4 MY from the equatorial Pacific. *Proceedings of ODP Scientific Research*, **138**, 779–793.
Niitsuma, N., 1971. Detailed studies of the sédiments recording the Matuyama-Brunhes geomagnetic reversal. *Tohoku Univ.Sci.Rpt 2nd Ser (Geology)*, **43**, 1–39.
Olson, P., and Amit, H., 2006. Changes in the Earth's dipôle. *Die Naturwissenschaften*, **93**, 519–542.
Opdyke, N. D., and Channell, J. E. T., 1996. *Magnetic Stratigraphy.* Academic Press: San Diego, p. 346.
Pitman, W. C., III, and Heirtzler, J. R., 1966. Magnetic anomalies over the Pacific-Antarctic Ridge. *Science*, **154**, 1164–1171.
Prévot, M., and Camps, P., 1993. Absence of preferred longitudinal sectors for poles from volcanic records of geomagneitc reversal. *Nature*, **366**, 53–57.
Prévot, M., Mankinen, E., Coe, R. S., and Grommé, C. S., 1985. The steens mountain (Oregon) geomagnetic polarity transition. 2. Field intensity variations and discussion of reversla models. *Journal of Geophysical Research*, **90**, 10417–10448.
Quidelleur, X., and Valet, J.-P., 1994. Paleomagnetic records of excursions and reversals: Possible biases caused by magnetization artefacts. *Physics of the Earth and Planetary Interiors*, **82**, 27–48.
Roche, A., 1950. Sur les caracteres magnétiques du système éruptif de Gergovie. *Comptes Rendus Academic Science*, **230**, 113–115.
Roche, A., 1951. Sur les inversions de l'aimentation remanente des roches volcaniques dans les monts d'Auvergne. *Comptes Rendus Academic Science*, **223**, 1132–1134. Paris.
Roche, A., 1956. Sur la date de la derniere inversion du champ magnetique terrestre. *Comptes Rendus Academic Science*, **243**, 812–814. Paris.
Roperch, P., and Chauvin, A., 1987. Transitional geomagnetic field behavior: volcanic records from French Polynesia. *Geophysical Research Letters*, **14**, 151–154.
Roperch, P., and Duncan, R. A., 1990. Records of geomagnetic reversals from volcanic islands from French polynesia 1. Paleomagnetic study of a polarity transition on a lava sequance from the Island of Huahine. *Journal of Geophysics Research*, **95** (B3), 2713–2726.
Rutten, M. G., and Wensink, H., 1960. Paleomagnetic dating glaciations and chronology of the Plio-Pleistocene in Iceland. *International Geological Congress*, **IV**, 62–70. Sess 21.
Stieglitz, R., and Müller, R., 2001. Experimental demonstration of a homogeneous two-scale dynamo. *Physics of Fluids*, **13**, 561, doi:10.1063/1.1331315.
Tauxe, L., and Hartl, P., 1997. 11 million years of oligocenegeomagnetic field behavior. *Geophysical Journal International*, **128**, 217–229.
Tauxe, L., and Shackleton, N. J., 1994. Relative paleointensity record from the Ontong-Java plateau. *Geophysical Journal International*, **117**, 769–782.
Valet, J.-P., and Herrero-Bervera, E., 2003. Some characteristics of geomagnetic reversals inferred from detailed volcanic records. *Comptes Rendus Geosciences*, **335**, 79–90.
Valet, J.-P., and Laj, C., 1981. Paleomagnetic record of two successive Miocene geomagnetic reversals in Western Crete. *Earth and Planetary Science Letters*, **54**, 53–63.
Valet, J.-P., and Laj, C., 1984. Invariant and changing transitional field configurations in a sequence of geomagnetic reversals. *Nature*, **311**, 552–555.
Valet, J.-P., and Meynadier, L., 1993. Geomagnetic field intensity and reversals during the last four million years. *Nature*, **366**, 234–238.
Valet, J.-P., Laj, C., and Langereis, C. G., 1983. Two different R-N geomagnetic reversals with identical VGP paths recorded at the same site. *Nature*, **304**, 330–332.
Valet, J.-P., Laj, C., and Tucholka, P., 1986. High resolution sedimentary record of a geomagnetic reversal. *Nature*, **322**, 27–32.
Valet, J.-P., Laj, C., and Langereis, C. G., 1988a. Sequential geomagnetic reversals recorded in upper tortonian marine clays in Western Crete (Greece). *Journal of Geophysical Research*, **93**, 1131–1151.
Valet, J.-P., Tauxe, L., and Clark, D. R., 1988b. The Matuyama-Brunhes transition recorded from Lake Tecopa sédiments 5 (California). *Earth and Planetary Science Letters*, **87**, 463–472.
Valet, J. P., Tucholka, P., Courtillot, V., and Meynadier, L., 1992. Paleomagnetic constraints on the geometry of the geomagnetic field during reversals. *Nature*, **356**, 400–407.
Vine, F. J., 1966. Spreading of the ocean floor: new evidence. *Science*, **154**, 1405–1415.
Vine, F. J., and Matthew, D. H., 1963. Magnetic anomalies over oceanic ridges. *Nature*, **199**, 947–949.
Weeks, R., Fuller, M., Laj, C., Mazaud, A., and Herrero-Bervera, E., 1992. Sedimentary records of reversal transitions – magnetization smoothing artefct or geomagnetic field behaviour? *Geophysical Research Letters*, **19**, 2007–2010.
Wensink, H., 1966. Paleomagnetic stratigraphy of younger basalts and intercalated Plio-Pleistocene tillites in Iceland. *Geol. Rund.*, **54**, 364–384.

GEOMAGNETIC FIELD, SECULAR VARIATION

Monika Korte
Sektion 2.3, Erdmagnetfeld, Helmholtz-Zentrum Potsdam, Deutsches GeoForschungsZentrum – GFZ, Potsdam, Germany

Geomagnetic field: secular variation

The geomagnetic field observed at and above the Earth's surface varies on a broad range of timescales, from less than seconds to millions of years. In general, fast variations with periods less than a year originate from sources external to the Earth. Secular variation is the continuous variation of the geomagnetic core field, associated to changes of fluid flow in the outer core (e.g., Wardinski, 2007). The name originates from the Latin "seculum," referring to a time interval of about a century. Secular variation describes the velocity of field change or the first temporal derivative of a time series of field observations.

Secular variation consists of mostly non-periodic, non-predictable field changes. The high-frequency end of the temporal secular variation spectrum is not well-understood. Very rapid variations that might occur in the core are damped and filtered out when passing through the mantle before being observed. The fastest known secular variation features, namely, geomagnetic jerks and rapid fluctuations, occur on timescales on the order of one year (Mandea and Olsen, 2009). It is difficult to discriminate between internal and external variations in observational data on these timescales. Long-term external field changes, e.g., the approximately 11-year sunspot cycle, are also reflected in the measured time series. Secular variation on several thousands to million year timescales includes excursions and complete polarity reversals of the geomagnetic core field.

Geomagnetic secular variation is not globally uniform. The geomagnetic dipole moment, a measure of the overall global field strength, has been decreasing at the rate of 5% per century since the beginning of systematic magnetic field intensity observations around 1840. The present spatial distribution of intensity secular variation is characterized by a large area of intensity decrease with two foci over North America and the southern Atlantic – south American region, and a smaller focus of intensity increase over the Indian Ocean (e.g., Lühr et al., 2009). Historical and paleomagnetic data indicate that secular variation has always been weak in the Pacific compared to the Atlantic hemisphere, which might be a consequence of lateral variations in lower mantle temperature (Gubbins and Gibbons, 2004).

Bibliography

Gubbins, D., and Gibbons, S., 2004. Low Pacific secular variation. In Channell, J. E. T., Kent, D. V., Lowrie, W., and Meert, J. G. (eds.), *Timescales of the Geomagnetic Field*. American Geophysical Union, Geophysical Monograph Series 145, pp. 279–286.

Lühr, H., Korte, M., and Mandea, M., 2009. The recent geomagnetic field and its variations. In Glaßmeier, K.-H., Soffel, H., and Negendank, J. (eds.), *Geomagnetic Field Variations*. Berlin: Springer, pp. 25–63.

Mandea, M., and Olsen, N., 2009. Geomagnetic and archeomagnetic jerks: where do we stand? *EOS, Transactions, AGU*, **90**, 24.

Wardinski, I., 2007. Geomagnetic secular variation. In Herrero-Bervera, E. (ed.), *Gubbins, D*. Encyclopedia of Geomagnetism and Paleomagnetism: Springer, pp. 346–349.

Cross-references

Core Dynamo
Geomagnetic Excursions
Geomagnetic Field, Polarity Reversals
Geomagnetic Field, Theory

GEOMAGNETIC FIELD, THEORY

Friedrich H. Busse
Institute of Physics, University of Bayreuth, Bayreuth, Germany

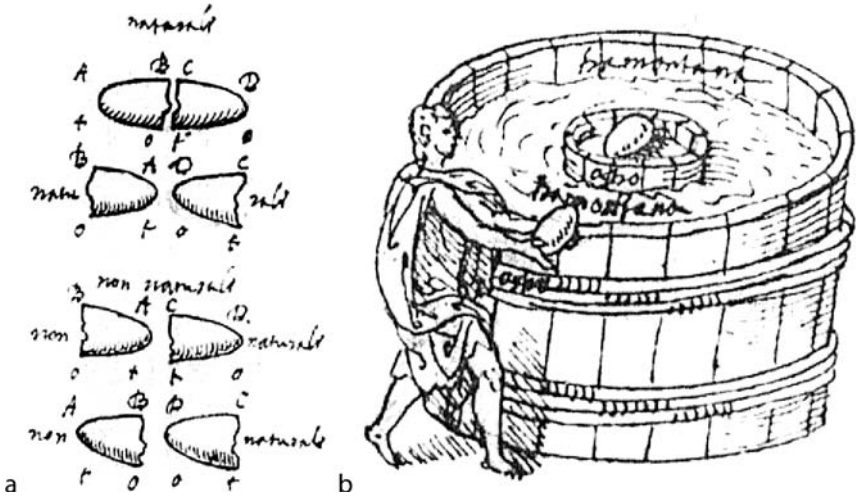
Definition

Geomagnetic field. Nearly dipolar magnetic field emerging from the Earth's interior.
Dynamo process. Conversion of kinetic energy of motions into magnetic energy.
Geodynamo. Dynamo process operating in the Earth's liquid outer core.

Introduction

The forces exerted by magnetized rocks or lodestones have been known to mankind for many centuries before Christ, but the directivity of the lodestone appears to have been first observed by the Chinese in the first century B.C. It is known that at this time Chinese geomancers used a freely turnable spoon-like lodestone on a polished plate as a compass, not for navigational purposes, but probably for aligning tombs and houses. As navigational instrument the compass was introduced to European sailors by arabic merchants in the twelfth century A.C. A first systematic experimental study of the forces exerted by lodestones was published by Petrus Peregrinus in 1269. Some of his results are shown in Figure 1a. More refined experiments were carried out and described in his book *De Magnete* by William Gilbert (1600). He concludes that the Earth itself is a magnet like a lodestone.

In 1839 Carl-Friedrich Gauss introduced the representation in terms of spherical harmonics for the observed geomagnetic field. Through this representation it became possible to separate the magnetic fields of internal and external origin. The latter, being much smaller than the former and much more variable in time, is caused by ionospheric and magnetospheric currents. Only during magnetic storms can the externally generated magnetic field reach a significant fraction of the total strength of the



Geomagnetic Field, Theory, Figure 1 Lodestone experiments of Petrus Peregrinus. (**a**) Experiments showing the attracting and repelling forces of a broken piece of lodestone. Upper configuration: "naturale," lower configuration: "non naturale." (**b**) A simple compass in the form of afloating lodestone. (After Stefani et al., 2008).

Earth's magnetic field (see *Magnetic Storms and Electromagnetic Pulsations*). Together with Alexander von Humboldt Gauss founded in 1836 the Göttinger Magnetischer Verein (Göttingen Magnetic Union) in order to create a worldwide network of observatories for the measurement of the magnetic field and its variations in time. Since that time the knowledge about the geomagnetic field and its time dependence has improved continuously (see *Geomagnetic Field, Global Pattern*).

The idea that the geomagnetic field originates from the ferromagnetism of the Earth's interior prevailed until the end of the nineteenth century even though the phenomenon of secular variation – known since the seventeenth century – was difficult to reconcile with the hypothesis of remanent magnetism. In early years of the twentieth century it became clear that the Curie temperature of ferromagnetic material is exceeded within the entire Earth except for the uppermost few tens of kilometers and that the magnetization of the Earth's crust is far to weak to act as a source of geomagnetism. Numerous other hypotheses have been proposed as origin of geomagnetism and rejected (Rikitake, 1966) except for the dynamo hypothesis. In the meantime magnetic fields had been observed on the sun and in 1919 Larmor proposed the dynamo mechanism for the origin of the magnetic field of sunspots.

According to the dynamo hypothesis motions within the outer liquid-iron core of the Earth can amplify a magnetic field from arbitrarily small amplitudes and sustain it at the observed finite strength against Ohmic dissipation. The feasibility of this mechanism in a simply connected region of nearly uniform electrical conductivity appeared to be doubtful for a long time, and the dynamo hypothesis experienced a setback when Cowling (1934) demonstrated that axisymmetric magnetic fields cannot be generated by the dynamo mechanism. Since the geomagnetic field is axisymmetric in first approximation, Cowling's theorem stimulated the search for alternative mechanisms. It was not until the late 1950s that Backus (1958) and Herzenberg (1958) demonstrated independently that dynamos are indeed possible in spheres of uniform conductivity.

Since that time dynamo theory has developed rapidly and it has become generally accepted that a geodynamo operates in the liquid outer core of the Earth. While in the 1960s and 1970s of the past century analytical models of the geodynamo have prevailed, the strong growth of computer capacities in recent decades has permitted numerical simulations with ever increasing details of convection driven dynamos in Earth's core-like rotating spherical shells. The principles of the dynamo process in such an environment are now reasonably well understood. The remaining uncertainties are caused by an insufficient knowledge of conditions in the Earth's core and lower mantle and the inability to model the dynamo process under the highly turbulent conditions of flow in the outer core.

Before outlining the concepts of dynamo theory we briefly mention the role of the crustal magnetic field which originates from ferromagnetic remanence of material below its Curie temperature in the lithosphere. When the observed magnetic field is represented in terms of spherical harmonics (see *Geomagnetic Field, IGRF*) it is found that crustal contributions dominate in coefficients of the degree 14 and higher. The crustal contributions to coefficients of lower degrees are still substantial, of course, but they do not exceed those originating from the core in general (see *Magnetic, Global Anomaly Map*). Crustal magnetic fields are often correlated with geological features and may indicate iron ore deposits such as the famous Kursk formation in the Ukraine. Other important crustal magnetic features are the magnetic stripes of the ocean floors extending parallel to the mid-oceanic ridges

which represent the most convincing evidence for the concept of plate tectonics (see *Magnetic Anomalies, Interpretation*).

The following sections are devoted to the dynamo theory of geomagnetism and its applications to the Earth's core. The current state of numerical models of the geodynamo will be described under the title Core dynamo.

Kinematic dynamo theory

The simplest description of a dynamo is obtained for a fixed velocity field, i.e., when the feedback of the generated magnetic field on the flow through the action of the Lorentz force is neglected. A single equation of magnetic induction describing the time dependence of the magnetic flux density $\vec{B}$ can be derived from Maxwell's equations in the magnetohydrodynamic approximation, in which the displacement current is neglected, and from Ohm's law which are given by

$$\nabla \cdot \vec{B} = 0, \qquad \frac{\partial \vec{B}}{\partial t} = \nabla \times \vec{E}, \qquad \nabla \times \vec{B} = \mu \vec{j}, \qquad \vec{j} = \sigma(\vec{E} + \vec{v} \times \vec{B}) \tag{1}$$

where $\vec{B}$ is the magnetic flux density, $\vec{E}$ is the electric field and $\vec{j}$ is the current density. σ, μ and $\vec{v}$ denote the electrical conductivity, the magnetic permeability and the velocity vector, respectively. Ohm's law has been written in the form valid for a moving conductor. Accordingly Equation 1 are invariant with respect to a Galilei transformation; in other words, the same Equations 1 hold in a system moving with constant velocity $\vec{V}$ with respect to the original system of reference when $\vec{v}$, $\vec{E}$ and $\vec{B}$ obey the relationships

$$\vec{v}' = \vec{v} - \vec{V}, \qquad \vec{B}' = \vec{B}, \qquad \vec{E}' = \vec{E} + \vec{V} \times \vec{B} \tag{2}$$

where the primed quantities refer to the new system. It is convenient to eliminate $\vec{E}$ and $\vec{j}$ from Equations 1 and to write

$$\frac{\partial \vec{B}}{\partial t} = \nabla \times (\lambda \nabla \times \vec{B}) = \nabla \times (\vec{v} \times \vec{B}) \tag{3}$$

where $\lambda \equiv 1/(\mu\sigma)$ is the magnetic diffusivity. Because of the absence of ferromagnetic effects in the Earth's core, the permeability is usually replaced by the permeability of the vacuum. From the form of Equation 3 it is obvious that $\nabla \cdot \vec{B} = 0$ remains satisfied if this equation is satisfied by the initial conditions. It is remarkable that Equation 3 is not only invariant with respect to Galileo transformations and thus compatible with the Navier–Stokes equations of motion, but is also invariant with respect to a transformation to a rotating system.

Since the Earth's outer core can be regarded as an incompressible fluid in first approximation, Equation 3 can be simplified further by the use of equations $\nabla \cdot \vec{v} = 0, \nabla \cdot \vec{B} = 0$,

$$\left(\frac{\partial}{\partial t} + \vec{v} \cdot \nabla\right)\vec{B} - \lambda \nabla^2 \vec{B} = \vec{B} \cdot \nabla \vec{v}, \tag{4}$$

where for simplicity $\lambda =$ const. has been assumed. This form of the equation of induction is revealing since it resembles the heat equation with the term $\vec{B} \cdot \nabla \vec{v}$ on the right hand side assuming the role of a heat source. This term tends to strengthen the magnetic field when a fluid particle is stretched in the direction of the local magnetic field line. Such an interpretation is especially useful for the dynamo problem. In order that a magnetic field $\vec{B}$ may grow, the term on the right hand side of Equation 4 must overcome the effect of the magnetic diffusion term on the left hand side. Using a typical velocity U and a typical length scale d, the ratio of the two terms can be described by the magnetic Reynolds number Rm,

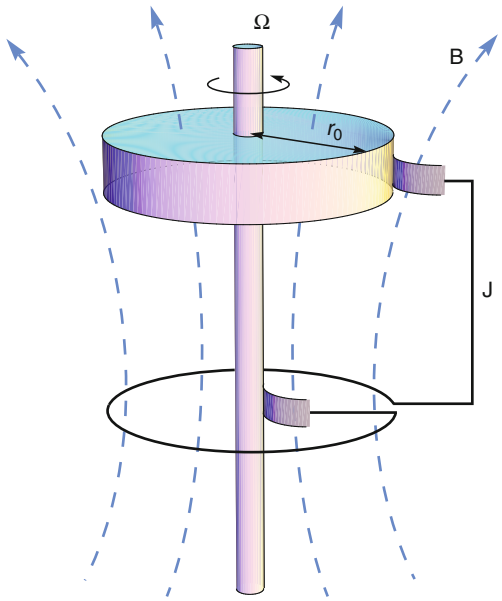
$$Rm \equiv Ud/\lambda. \tag{5}$$

Only when Rm is of the order of 1 or larger may growing magnetic fields become possible. Necessary conditions of this kind for a dynamo have been derived rigorously by Backus and Childress for a fluid sphere of constant magnetic diffusivity,

$$Rm \equiv Ur_o/\lambda > \pi, \tag{6}$$

where r_o is the radius of the sphere. In Childress's formulation U refers to the maximum velocity, whereas in Backus's formulation $U/\pi r_o$ is the maximum rate of strain of the velocity field.

The simplest dynamo is the disk dynamo originally invented by Faraday and shown in Figure 2. A metal disk is rotating about its axis with the angular velocity Ω. When the disk is permeated by an initial magnetic field $\vec{B}_0$



Geomagnetic Field, Theory, Figure 2 The disk dynamo.

parallel to the axis, an electromotive force U_e is generated between axis and rim of the disk,

$$U_e = \Omega \int_{r_1}^{r_0} \vec{B}_0 \, r dr \tag{7}$$

where r_1 and r_0 denote the radii of axis and disk. The electromotive force U_e can be used to drive an electric current J through the circuit indicated in the figure. Denoting by L and R the inductivity and the ohmic resistance of the circuit we obtain for the current J the equation

$$L\frac{dJ}{dt} + RJ = U_e \tag{8}$$

The current J flowing through the circuit is associated with a magnetic field $\vec{B}_1$ which may replace the initial field $\vec{B}_0$. The integral $2\pi \int_{r_1}^{r_0} \vec{B}_1 \, dr/J$ describes the mutual inductivity M between circuit and disk. Equation 8 for the self-excited disk dynamo can thus be written in the form

$$L\frac{dJ}{dt} = (\Omega M/2\pi - R)J \tag{9}$$

which allows for exponentially growing solutions once the dynamo condition

$$\Omega > 2\pi R/M \tag{10}$$

is satisfied. The exponential growth will be modified, of course, as soon as the Lorentz force $\vec{j} \times \vec{B}$ in the disk has grown sufficiently to affect the angular velocity Ω. A larger torque will then be needed to sustain the rotation speed.

The disk dynamo is an inhomogeneous dynamo since it depends on an inhomogeneous, multiply connected distribution of electrical conductivity as given by the wiring of the circuit. The dynamo would not work if the sense of wiring around the axis would be opposite to that shown in the figure or if it would be immersed in a highly conducting fluid. In this respect the disk dynamo differs from the homogeneous dynamo operating in the Earth's core in a singly connected finite domain of essentially homogeneous electrical conductivity. Velocity fields required for driving a homogeneous dynamo are necessarily more complex than the simple rotation velocity of the disk dynamo. It can even be proven that a simple form of motion, such as that of the disk dynamo, cannot drive a dynamo in a sphere if λ is a function of only the radial coordinate. More generally, the toroidal theorem states that growing solutions of Equation 3 do not exist if $\vec{v}$ and λ can be written in the form

$$\vec{v} = \nabla\Psi \times \vec{r}, \qquad \lambda = \lambda(|\vec{r}|). \tag{11}$$

Inspite of their differences homogeneous dynamos and the disk dynamo share a number of properties:

(i) For given external parameters there exists always a solution without magnetic field besides the dynamo solution, just as the disk of Figure 2 can rotate with $\Omega > \Omega_c$ in the absence of any initial field $\vec{B}_0$.

(ii) The existence of a dynamo solution requires that the magnetic Reynolds number Rm exceeds a critical value Rm_c which corresponds to the quantity $\Omega M/R$ in the case of the disk dynamo. The nonmagnetic solution exists, but it is unstable for $Rm > Rm_c$.

(iii) Magnetohydrodynamic dynamos always occur in two forms which are identical except for the sign of the magnetic field $\vec{B}$. This property reflects the fact the Lorentz force is given by an expression that is quadratic in $\vec{B}$.

Property (iii) is the basic reason that the geomagnetic field has often switched its polarity in the geologic past. These "reversals" have occurred nearly randomly about every 200,000 years in the time average. In contrast, the solar magnetic field reverses every 11 years in a surprisingly periodic fashion.

For the description of magnetic fields associated with a spherical system one often uses the general representation for a solenoidal vector field,

$$\vec{B} = \nabla \times (\nabla h \times \vec{r}) + \nabla g \times \vec{r} \tag{12}$$

in terms of a poloidal and a toroidal component each of which is decribed by a scalar function, the poloidal function h and the toroidal function g. Without loss of generality the averages of h and g over surfaces $|\mathrm{r}| = \text{const.}$ can be assumed to vanish. A homogeneous dynamo usually requires the interaction of both components. The functions h and g can be separated into their axisymmetric parts $\bar{h}$ and $\bar{g}$ and nonaxisymmetric parts, $\check{h} = h - \bar{h}$ and $\check{g} = g - \bar{g}$. Because of Cowling's theorem the latter parts play an essential role in the dynamo process. The component $\bar{g}$ can easily be generated in a spherical dynamo through a stretching of the axisymmetric poloidal field by a differential rotation. This process is known as the ω-effect. But the amplification of $\bar{h}$ requires the interaction of the nonaxisymmetric components of magnetic fields and velocity fields. This is often called the α-effect. This latter effect can also be used, of course, for the generation of $\bar{g}$ in the absence of a differential rotation. One thus distinguishes $\alpha\omega$- and α^2-dynamos. These concepts were originally introduced within the framework of mean-field magnetohydrodynamics for which we refer to the book of Krause and Raedler (1980).

Energetics of the geodynamo

For a prescribed stationary velocity field, Equation 3 admits solutions with an exponential time dependence, of which the growing ones are called dynamos. As the magnetic field grows, the action of the Lorentz force on the velocity field can no longer be neglected. The work done by the velocity field against the Lorentz force equals the energy consumed by ohmic heating. Thus, the

sustenance of a finite-amplitude magnetic field against ohmic dissipation requires a source of mechanical energy. A lower bound of about $10^8 W$ on the ohmic heating in the Earth's core can be obtained by a calculation of the minimum dissipation corresponding to the observed geornagnetic field, while an upper bound of about $10^{13} W$ is given by the observed geothermal heat flux. A realistic value is likely to be within the range 10^{11} to $10^{12} W$.

Numerous sources of energy for the geodynamo have been considered in the history of dynamo theory. The major ones that cannot be excluded right away are the precession of the Earth, thermal heat sources, and compositional buoyancy. The latter source first proposed by Braginsky in 1963 is probably the largest contributor, since the turbulence induced by precession appears to be marginal and since the Carnot efficiency of heat sources is relatively low (Gubbins et al., 2004). Compositional buoyancy is generated by the growing solid inner core. Iron and perhaps nickel crystallize at the inner core-outer core boundary while lighter elements stay in solution. The decrease in density caused by the increase in the concentration of light elements provides a source of potential energy that can drive convection in the outer core. Since the outer core contains about 10% light elements by weight, most likely silicon, sulfur and oxygen, the buoyancy provided by this process appears to be more than adequate to drive the geodynamo. The compositional buoyancy mechanism has the advantage that it could explain the apparent absence of a dynamo in the case of Venus (Stevenson et al., 1983; Nimmo, 2002). Since Venus resembles the Earth in most respects, the absence of a measurable magnetic field has posed a puzzle. The somewhat higher pressure at the Earth's center and the more efficient heat transport out of the planet because of plate tectonics seem to have allowed the inner core to form, whereas Venus will reach this stage of its thermal history much later. Estimates for the age of formation of the inner core of the Earth range from 1 to 2 b.y. Since evidence for geomagnetism exist for earlier times in the paleomagnetic record when the inner core had probably not yet nucleated, thermal buoyancy generated by cooling and possibly by radioactive elements is likely to have played a role in driving the geodynamo in the Earth's early history. In this respect the Earth resembles the major planets for which thermal buoyancy appears to be the only viable source of energy to drive the planetary dynamo.

The way in which compositional buoyancy drives convective motion in the outer core is not well understood. It is even possible that the outer part of the outer core is stably stratified. Another peculiar property of the solidification process at the inner core is the so-called mushy zone in which liquid and solid phases coexist (see *Earth's Structure, Core*). But those details have little influence on the theory of the geodynamo in its current state. Since the form of motion in the outer core is primarily determined by Coriolis and Lorentz forces the theory of geomagnetism has proceeded without distinguishing much between thermal and compositional buoyancy.

Geomagnetic Field, Theory, Table 1 Properties of the liquid outer core of the Earth

Property	Symbol	Value	Unit
Angular velocity of rotation	Ω	$7.27 \cdot 10^{-5}$	s^{-1}
Mean outer radius	r_o	3485 ± 3	km
Mean inner radius	r_i	1225 ± 15	km
Density	ρ	$9.9 - 12.2$	$10^3 kg/m^3$
Coefficient of thermal expansion	α	10^{-5}	K^{-1}
Thermal diffusivity	κ	$5 \cdot 10^{-6}$	m^2/s
Magnetic diffusivity	λ	2	m^2/s
Kinematic viscosity	ν	10^{-6}	m^2/s
Permeability of vacuum	μ_0	$4\pi \cdot 10^{-7}$	H/m

Some typical values of parameters used in the theory of the geodynamo are given in Table 1.

Magnetohydrodynamic theory of the geodynamo

The theory of the geodynamo requires the inclusion of the equations of motion for the rotating liquid core for several reasons. First of all, constraints on the velocity fields used in the dynamo equation are needed since rather arbitrary velocity fields are capable of generating fields that resemble the observed geomagnetic field. The velocity fields driven by geophysically realistic forces share special properties, such as the approximate alignment of the vorticity with the axis of rotation, that make them useful in the interpretation of certain features of the observed magnetic field. The second reason for including the equations of motion stems from the necessity of taking into account the Lorentz force without which an equilibrium amplitude of the magnetic field cannot be determined. Thirdly, the time dependence of the geomagnetic field can not be understood without a detailed knowledge of the dynamics of the velocity field in the core. The basic equations of motion for the velocity vector $\vec{v}$, the heat equation for the deviation Θ from the static temperature distribution, and the equation of induction for the magnetic flux density $\vec{B}$ are usually written in the following form:

$$\nabla \cdot \vec{v} = 0, \qquad \nabla \cdot \vec{B} = 0, \tag{13}$$

$$\begin{aligned}(\partial_t + \vec{v} \cdot \nabla)\vec{v} + 2\Omega\vec{k} \times \vec{v} \\ = -\nabla\pi + \alpha\Theta g\vec{r} + \nu\nabla^2\vec{v} + \vec{B} \cdot \nabla\vec{B}/\rho\mu,\end{aligned} \tag{14}$$

$$\partial_t\Theta + \vec{u} \cdot \nabla\Theta = \vec{v} \cdot \nabla T_s + \kappa\nabla^2\Theta, \tag{15}$$

$$\lambda\nabla^2\vec{B} = \partial_t\vec{B} + \vec{v} \cdot \nabla\vec{B} - \vec{B} \cdot \nabla\vec{v}, \tag{16}$$

where ∂_t denotes the partial derivative with respect to time t and where $-g\vec{r}$ denotes the force of gravity. ν is the kinematic viscosity, $\vec{k}$ is the unit vector of the axis of rotation and Ω is the angular velocity of rotation. The potential temperature distribution T_s describes the difference

between the temperature distribution for the static solution of the problem and the adiabatic temperature distribution with the same temperature at the inner core boundary. As mentioned before, Equation 15 remains essentially unchanged if T_s and Θ are interpreted a concentration of light elements. In that case the coefficient of thermal expansion α and the thermal diffusivity κ just need to be reinterpreted. All gradient terms in the equation of motion have been combined into $\nabla\pi$. In writing Equations 13–16 we have assumed the Boussinesq approximation implying constant material properties and a constant density ϱ except in the gravity term.

Using a dimensionless description the parameters of the problem can be reduced to four dimensionless groups, the Rayleigh number Ra, the Coriolis number τ, the Prandtl number Pr, and the magnetic Prandtl number Pm which are defined by

$$Ra = \frac{\alpha g \Delta T d^4}{\nu\kappa}, \quad \tau = \frac{2\Omega d^2}{\nu}, \quad Pr = \frac{\nu}{\kappa}, \quad Pm = \frac{\nu}{\lambda}, \quad (17)$$

where the thickness d of the spherical shell has been used as length scale and ΔT is the difference of the potential temperature between the inner core-outer core boundary and the core-mantle boundary. Instead of the Coriolis number often its inverse, the Ekman number $E = \tau^{-1}$, is used. Since λ/ν is roughly 10^6 in the Earth's core the main sink of mechanical energy is ohmic dissipation and it is generally believed that the term $\nu\nabla^2\vec{v}$ is unimportant except close to the rigid boundaries of the liquid core. Moreover, the magnetostrophic approximation is often applied in which the inertial term $(\partial_t + \vec{v}\cdot\nabla)\vec{v}$ is neglected in comparison to the Coriolis force. Without these terms the important Taylor constraint can be derived: After integrating the azimuthal component of Equation 14 over a coaxial cylindrical surface in the liquid core, one must conclude that an accelerating differential rotation occurs unless the condition

$$\iint [(\nabla \times \vec{B}) \times \vec{B}]_\varphi s d\varphi dz = 0 \quad (18)$$

is satisfied for every distance s from the axis. In deriving Equation 18 we used the fact that the inflow across the cylindrical surface is balanced by the outflow with the result that the contribution from the Coriolis force term vanishes. Different strategies have been designed to accommodate the Taylor constraint (Equation 18). In Braginsky's (1978) model Z the friction of the viscous Ekman layer at the solid surface of the sphere provides a balance on the right side of Equation 18. But this effect is small such that the constraint of Equation 18 becomes nearly satisfied by a poloidal field that is extended parallel to the axis of rotation except for the region close to the boundary.

In modern numerical simulations of convection driven dynamos in rotating spherical shells the Taylor constraint (Equation 18) is usually not approached because relatively high values of the kinematic viscosity must be assumed for numerical reasons and also because the inertial term $(\partial_t + \vec{v}\cdot\nabla)\vec{v}$ plays a larger role than it possibly plays in the geodynamo. Instead of an Earth-like ratio, values of the order unity for λ/ν are used to ensure numerical convergence. The high value of ν is interpreted as an eddy viscosity which is supposed to represent the dissipative action of small scale components of the velocity field which can not be resolved in the numerical simulations.

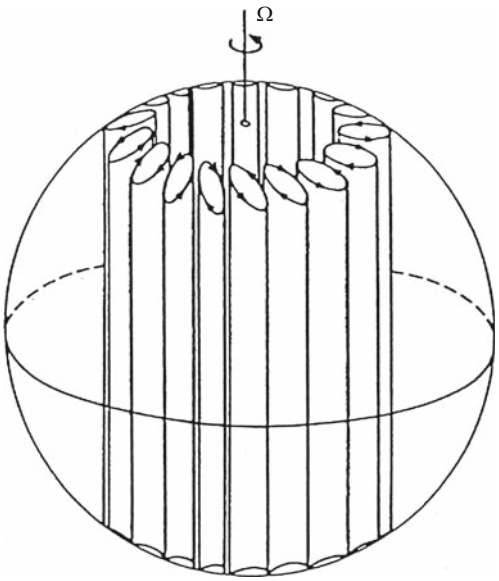
It is instructive to consider Equations 13–16 in the limit of a vanishing magnetic field. In that case the dominant terms are the Coriolis force and the pressure gradient, the balance of which,

$$2\Omega\vec{k} \times \vec{v} = -\nabla\pi \quad (19)$$

is called the geostrophic balance. The curl of Equation 14 together with $\nabla \cdot \vec{v} = 0$ gives rise to the Proudman–Taylor theorem

$$\Omega\vec{k} \cdot \nabla\vec{v} = 0 \quad (20)$$

which states that steady motion of a rotating inviscid fluid must not depend on the coordinate in the direction of the axis of rotation. Convection in a sphere requires, however, a finite radial velocity component and therefore cannot fully satisfy condition (20). As a result nonmagnetic solutions of Eqs. (13–16) exhibit a time dependence in the form of a drift in the prograde azimuthal direction. In addition they are characterized by small length scales in the directions perpendicular to the axis of rotation whereby viscous friction is increased. A sketch of the motion at the onset of convection in a homogeneously heated sphere is shown in Figure 3. For details on the dynamics of these



Geomagnetic Field, Theory, Figure 3 Sketch of convection columns in a rotating, internally heated sphere for Rayleigh numbers close to the critical value.

convection columns and their dynamo action we refer to the review article by Busse (2002).

Because of the small viscosity in the Earth's core, the Lorentz force offers a more effective release of the rotational constraint expressed by condition (20) than does viscous friction. The simplest approach to this problem is to consider the onset of convection in the presence of an azimuthal magnetic field of strength B_0 (Fearn, 1979). The analysis shows that the scale of the preferred mode of convection increases dramatically as B_0 is increased and the critical value of the Rayleigh number and the drift rate decrease. The dimensionless parameter governing this effect is the Elsasser number

$$\Lambda = \frac{B_0^2}{2\Omega\mu_0\rho\lambda} \tag{21}$$

which describes the ratio between Lorentz and Coriolis forces. When the magnetic field becomes very strong such that Λ exceeds unity it exerts an inhibiting influence on convection. It turns out that the minimum Rayleigh number for the onset of convection is achieved for $\Lambda \approx 1$. Eltayeb and Roberts suggested in 1970 that this condition is likely to determine the equilibrium strength of the magnetic field inside the Earth's core. While the condition $\Lambda \approx 1$ appears to fit the amplitude of the geomagnetic field reasonably well, numerical simulations of spherical dynamos do not provide much support for such a relationship. In fact, the dynamical effect of a magnetic field generated by dynamo action is quite different from that of an imposed magnetic field. In present geodynamo models it does not seem to affect the scale of convection significantly nor has it been found that it allows convection to exist at Rayleigh numbers below the critical value for onset of convection in the absence of a magnetic field.

There is still no general agreement about the relationship that governs the average strength of the magnetic field reached in the dynamo process. But it appears likely that the available energy for driving motions determines the average strength of the magnetic field as has been proposed by Christensen and Aubert (2006).

The velocity field displayed in Figure 3 is symmetric with respect to the equatorial plane. Even in its turbulent state the convection flow retains approximately this symmetry outside the tangent cylinder which touches the inner core at its equator. Because of this symmetry the solutions $\vec{B}$ of Equations 13–16 can be separated into two classes: a quadrupolar class that has the same symmetry with respect to the equatorial plane as the velocity field, and a dipolar class that exhibits the opposite symmetry. In both cases the Lorentz force is compatible with the symmetry of the velocity field. The observed geomagnetic field is primarily a dipolar field and so are most of the other planetary magnetic fields, but from a basic point of view quadrupolar dynamos are not less likely. The more complex structure of their poloidal field is balanced by the simpler form of their zonal component. The latter is symmetric with respect to the equatorial plane, where as it is antisymmetric in the dipolar case. The observed geomagnetic field includes a substantial contribution from quadrupolar terms which is not surprising, since the symmetry property exhibited in Figure 3 does not hold inside the tangent cylinder. Moreover, the thermal conditions at the core surface are not likely to be spherically homogeneous because of convection in the Earth's mantle.

The solution of Equations 13–16 becomes much simpler when the geometry of the spherical shell is replaced by that of a cylindrical annulus with a height decreasing with distance from the axis. An early analytical model (Busse, 1975) demonstrates that the dynamo action of convection columns such as those shown in Figure 3 depends on their helicity. The latter is defined as the azimuthal average over the vorticity multiplied by the velocity vector, $\vec{v} \cdot (\nabla \times \vec{v})$. The helicity is generated for example in the columns through the suction exerted by the Ekman layers at the upper and lower boundaries and by the finite change in height of the columns with distance from the axis. This helicity effect appears to dominate also in modern numerical simulations of the geodynamo.

Time dependence of the geodynamo

From observations of the secular variation and from the paleomagnetic evidence for reversals of the direction of the geomagnetic field, it is clear that the mechanism of the geodynamo is not steady. Since a complete model of the geodynamo does not yet exist, the various features that may lead to the observed time dependence cannot be derived from basic mechanisms. Thus, the theories of secular variation and of reversals rely on simplifying assumptions (see *Geomagnetic Field, Secular Variation*; *Geomagnetic Field, Polarity Reversals*). As simulations of the geodynamo become more refined, however, the observed secular variation and the increasing body of data on the process of excursions and reversals will be important for discriminating between different models of the geodynamo.

The time dependence of the nonaxisymmetric part of the magnetic field is a result of the basic dynamics discussed in the preceding section. Even without a magnetic field, the convection motion exhibits a time dependence in the form of a prograde or eastward drift at slightly supercritical Rayleigh numbers, to which a chaotic variability is added as convection becomes more vigorous. In the presence of a magnetic field, this drift is decreased substantially because the Lorentz force balances the ageostrophic part of the Coriolis force. The order of magnitude of the drift corresponds roughly to the observed westward drift of the nondipole part of the geomagnetic field. But since there is likely to be a zonal flow of the liquid core relative to the mantle of a similar order of magnitude, no direct comparison between theoretical and observed drift rates can be made without further knowledge about the velocity field in the core.

The geomagnetic time dependence of internal origin with the shortest duration is known as "geomagnetic jerk."

This term refers to rather abrupt changes in the second derivative with respect to time of the East-West component of the geomagnetic field over the time span of a year or so. Geomagnetic jerks are likely to be the result of magnetohydrodynamic waves in the outer core (Bloxham et al., 2002).

In contrast to the solar magnetic field, the geomagnetic field does not reverse in nearly periodic intervals. The time scale of reversals is an order of magnitude larger than the time scale of about 2×10^4 years for the diffusive decay of Earth's magnetic field. This may suggest the geodynamo is basically a steady dynamo. Recent evidence (Roberts, 2008) suggests, however, that geomagnetic excursions, in which the dipole moment of the Earth decreases to almost vanishing values for a time span of the order of a 1,000 years, occur much more frequently than reversals. No consensus about the origin of excursions and reversals has yet been achieved.

Outlook

In the past 2 decades the dynamo theory of geomagnetism has progressed enormously through increasing availability of computer resources; see *Core Dynamo*. This development will continue in the future with the main goal of reaching higher (and thus more realistic) values of the Coriolis number. Ultimately the lack of data on properties of the Earth's core and lower mantle will inhibit the progress in understanding the geodynamo.

For a more detailed exposition of the foundations dynamo theory and the associated fluid dynamical concepts, the reader is referred to the book edited by Dormy and Soward (2007).

Bibliography

Backus, G. E., 1958. A class of self sustaining dissipative spherical dynamos. *Annalen der Physik*, **4**, 372–447.

Bloxham, J., Zatman, S., and Dumberry, M., 2002. The origin of geomagnetic jerks. *Nature*, **420**, 65–68.

Braginsky, S. I., 1963. Structure of the F-layer and reasons for convection in the Earth's core. *Soviet Physics Doklady*, **149**, 8–10.

Braginsky, S. I., 1978. Nearly axially symmetric model of the hydromagnetic dynamo of the earth. *Geomagnetism and Aeronomy*, **18**, 225–231.

Busse, F. H., 1975. A model of the geodynamo. *Geophysical Journal of the Royal Astronomical Society*, **42**, 437–459.

Busse, F. H., 2002. Convection flows in rapidly rotating spheres and their dynamo action. *Physics of Fluids*, **14**, 1301–1314.

Christensen, U. R., and Aubert, J., 2006. Scaling laws for dynamos in rotating spherical shells. *Geophysical Journal International*, **166**, 97–114.

Christensen, U. R., Aubert, J., Cardin, P., Dormy, E., Gibbons, S., Glatzmaier, G. A., Grote, E., Honkura, Y., Jones, C., Kono, M., Matsushima, M., Sakuraba, A., Takahashi, F., Tilgner, A., Wicht, J., and Zhang, K., 2001. A numerical dynamo benchmark. *Physics of the Earth and Planetary Interiors*, **128**, 25–34.

Cowling, T. G., 1934. The magnetic field of sunspots. *Monthly Notices of the Royal Astronomical Society*, **94**, 39–48.

Dormy, E., and Soward, A. M. (eds.), 2007. *Mathematical Aspects of Natural Dynamos*. Boca Raton: CRC Press.

Fearn, D. R., 1979. Thermal and magnetic instabilities in a rapidly rotating fluid sphere. *Geophysical and Astrophysical Fluid Dynamics*, **14**, 103–126.

Gilbert, W., 1600. *De Magnete*. London: Chiswick Press.

Gubbins, D., Alf, D., Masters, G., Price, G. D., and Gillan, M., 2004. Gross thermodynamics of two-component core convection. *Geophysical Journal International*, **157**, 1407–1414.

Herzenberg, A., 1958. Geomagnetic dynamos. *Philosophical Transactions of the Royal Society of London. Series A*, **250**, 543–585.

Krause, F., and Raedler, K.-H., 1980. *Mean-field Magnetohydrodynamics and Dynamo Theory*. Oxford: Pergamon.

Nimmo, F., 2002. Why does Venus lack a magnetic field? *Geology*, **30**, 987–990.

Rikitake, T., 1966. *Electromagnetism and the Earth's Interior*. Amsterdam: Elsevier.

Roberts, A. P., 2008. Geomagnetic excursions: knowns and unknowns. *Geophysical Research Letters*, **35**, L17307.

Stefani, F., Gailitis, A., and Gerbeth, G., 2008. Magnetohydrodynamic experiments on cosmic magnetic fields. *Zeitschrift fur Angewandte Mathematik und Mechanik*, **88**, 930–954.

Stevenson, D. J., Spohn, T., and Schubert, G., 1983. Magnetism and thermal evolution of the terrestrial planets. *Icarus*, **54**, 466–489.

Cross-references

Core Dynamo
Core-Mantle Coupling
Earth's Structure, Core
Energy Budget of the Earth
Geomagnetic Excursions
Geomagnetic Field, Polarity Reversals
Geomagnetic Field, Secular Variation
Paleomagnetism, Principles

GEOPHYSICAL WELL LOGGING

Miroslav Kobr
Institute of Hydrogeology, Engineering Geology and Applied Geophysics, Charles University in Prague, Prague, Czech Republic

Synonyms

Borehole geophysics

Definition

Borehole geophysics includes all methods that make continuous profiles or point measurements at discrete depth stations in a borehole. These measurements are made by lowering different types of probes into a borehole and by electrically transmitting data in the form of either analog or digital signals to the surface, where they are recorded as a function of depth. The measurements are related to the physical and chemical properties of the rocks surrounding the borehole, the properties of the fluid saturating the pore spaces in the formation, the properties of fluid in the borehole, the construction of the well, or some combination of these factors.

Introduction

The science of well logging was begun by Conrad Schlumberger in 1927 as an application of his work on resistivity measurements of the earth in surface exploration (see *Electrical Resistivity Surveys and Data Interpretation*). From the early measurement of electrical properties by means of normal and lateral resistivity curves, modern logging has progressed through such changes as the introduction of wall-contacting microspacing device, dipmeters, and conductive and inductive techniques. Parameters other than electrical have been utilized, such as radioactive properties of rocks both passive and forced as a result of bombardment by both gamma rays and neutrons; the measurement of acoustic properties, the measurements of other rock properties and auxiliary technique developments such as dielectric logging, temperature log, downhole viewers, remote sensors (dipmeter log, radar system, borehole gravimeter), etc.

The rapid advancement of well-logging techniques is observed in the 1980s and 1990s with the motivation of the petroleum industry to develop ever more cost-effective technology to evaluate thin-bed and nonconventional reservoirs. At the same time, geophysical technology also advanced rapidly with development of new techniques and improved mathematical methodology. The increased use of array tool design (multiple sources and receivers) provides improved vertical resolution, improved and multiple depths of investigations, and enables radial and azimuthal imaging by resistivity tools. Cased-hole resistivity logging may be realized, what could not be possible earlier. Acoustic-array designs, using either dipole or quadrupole sources, can acquire high-quality shear wave data in most types of reservoir rocks. A new generation of nuclear magnetic resonance (NMR) tools using pulse-echo technology provides an improved measurement without the operational restrictions of earlier tools. Nuclear tools have been improved by the use of new high-density gamma-ray detectors that allow higher logging speeds and better repeatability and permit a slimhole, through-tubing tool design. Renewed interest in neutron activation methods has produced new tools based on oxygen and silicon activation, like the pulsed-neutron source sensor. Improvements to existing resistivity and nuclear measurement-while-drilling (MWD) tools plus new caliper and acoustic porosity devices and measurement at the bit, ensure that MWD will continue to replace wireline logging in many situations. High angle, horizontal, and slimhole wells are now common as a result of advances in drilling technology. Wireline and MWD logging have been adapted for use in these wells; and new directionally focused resistivity tools have been developed; logging on coiled tubing is now routine. Improvements of downhole microprocessor-based technology and high-speed digital telemetry make possible the use of high data-rate array and imaging tools as well as increased possibilities of combinations of different tool types. Concurrent improvements in uphole computer processing combined with new visualization techniques made possible by the advent of powerful desktop computers enhance presentation and interpretation of the data acquired with these tools.

A chronological listing of these developmental steps is presented in literature (Johnson, 1962; Segesman, 1980; Prensky, 1994). The applications for geophysical well-logging methods continue to expand away from exploration for natural resources (oil and gas, water, minerals, geothermics) into the field of civil engineering, rock mechanics, disaster prevention, and environmental preservation. Information about recent and emerging developments and trends in well-logging and rock-characterization technology offer a lot of publications from meetings of geophysical societies (proceedings) and geophysical journals and books.

The main objective of borehole geophysics is to obtain more information than can be obtained from conventional drilling, sampling, and testing. Drilling a borehole or well is expensive, but the borehole provides access to the subsurface for geophysical probes. The most obvious log is a description of extracted cores, to provide information of lithology, microfossils, permeability, porosity, and fluids. These logs are used for geological correlation between different wells across an oil field and for the quantitative interpretation of the geophysical logs. The ideal would be to obtain continuous core, but this is very slow and thus expensive, therefore, it is done only for formations of special interest, usually potential oil-producing formations when they can be identified in advance. After drilling, samples may be obtained by sidewall sampling. Samples of fluids from surrounding rocks are also taken with a formation tester. Borehole logs provide continuous records that can be analyzed in real time at the well site. Thus, they can be used to guide completion or testing procedures. Log data are readily reproduced over long periods. Logs also aid the lateral and vertical extrapolation of geologic and fluid sampling data or hydraulic test data obtained from wells. In contrast, samples of rock or fluid from a borehole provide data only from sampled depth intervals and only after laboratory analysis. Data from some geophysical logs, such as acoustic velocity and resistivity, are also useful in interpreting surface geophysical surveys.

The manner in which a well is drilled affects the logging environment, through such factors as the resistivity of the fluid in the well bore, the diameter of the well, the roughness (rugosity) of the borehole wall, and the mud-filtrate invasion of porous formations surrounding the well. The thickness and composition of material between a logging probe and the borehole wall can have an effect on the data set. In general, the larger the borehole diameter, the poorer the quality of the geophysical logs obtained.

Wellbore environment

The physical properties of the geological formations in the vicinity of a borehole are often affected by the drilling process. Further, the temperature, lithostatic stress, and fluid

pressure increase with depth. Logging cable and tools must be constructed in such a way that can withstand the highest temperatures and pressures encountered in the well. Interpretation of well-log data may commonly require temperature correction.

Drilling is usually done using a rotating bit that grinds away the formations to produce cuttings from shales, sand from unconsolidated sandstones, or chips up to a few millimeters in size from solid rocks. The bit is fitted to the bottom of the drill stem, which typically consists of 10-m lengths of steel pipe screwed together. The bit may be turned by rotating the whole drill string of 10 m pipes from its top, but more commonly nowadays the bit is turned up by the mud being pumped through the turbine just behind the bit. Turbine drilling allows much greater control of drilling direction.

Down through the drill stem is pumped drilling fluid, commonly termed mud. It is a mixture of ingredients (liquid and solids) formulated to produce the required rheological properties such as density and viscosity. Drilling mud expedites the drilling operation by cooling, lubricating and cleaning the bit, reducing friction, and transporting cuttings from bit to the surface. Mud also serves to stabilize the wall of the hole, and - by its weight - to prevent any high-pressure gas encountered from blasting out of the hole, and also forms a mud cake, which is a layer made up of the solids in the mud deposited on the sides of hole opposite an intergranular permeable formation ensuring that the hydrostatic pressure of the drilling mud is higher than that of groundwater. The formation is invaded by the liquid of the mud, the mud filtrate, to a distance that depends on the porosity of the formation and properties of the mud. Nearest to the hole is the flush zone, in which the fluids are entirely mud filtrate (apart from some residual fluid that cannot be removed); next is the transition zone, with decreasing replacement until a sufficient distance is the uninvaded zone, quite unaffected by the drilling. In very permeable and porous formations, the mudcake builds up quickly and forms a barrier to further invasion. In less-permeable, low-porosity formations, the buildup is slow and invasion can be very deep. Thus, the resistivity of permeable rocks changes markedly because of the origin of these radial zones (Schlumberger, 1987). In pelitic formations (claystones, siltstones), the changes in elastic and mechanical (strength) properties occur along the borehole walls. Under the influence of drilling mud, the clay minerals swell and the cohesion of the rock is reduced; a low-velocity zone is formed along the borehole wall in the pelitic formations, being the more marked the longer the time interval between drilling and the moment of velocity measurement.

Calibration and standardization of logs

Standardization of logging measurements is the process of checking the response of logging probes in the field, using some type of portable field standard, usually before and after logging a well, to provide the basis for system drift correction in output with time. The Gamma-ray calibration pit and Neutron pit of the American Petroleum Institute (API) in the University of Houston are generally accepted over the world as a basic standard for nuclear logs because of definition of the API Gamma-ray unit and API Neutron unit.

Calibration is the process of converting measuring units of log response to units that represent rock characteristics. Calibration facilities should closely simulate the lithology and geometry to be logged. In many applications, it becomes difficult to utilize such calibration facility for several reasons (namely, the local lithologic environment substantially differs from lithology of calibrations blocks). Then, the calibration is performed by developing a regression between the measured geophysical property (the log) and the property of interest determined from laboratory tests using recovered core. The necessary condition is that the boreholes under study have been carefully cored. In principle, the quality of calibration is indicated by the correlation coefficient given from the regression. This calibration approach is also called "method of key beds."

Benefits and limitations of logging

Well logging differs from most other geophysical methods, not only in being carried out down a borehole, but in relating the physical quantities measured more specifically to geological properties of interest, such as lithology, porosity, and fluid composition. Their main advantage is greater detail, the measurements are made within formations of interest and can be compared with the geological information obtained from the same hole. Borehole logs usually provide continuous data that can be used to guide testing procedure and thus may reduce the number of samples needed and in this way the cost. The data are useful for interpreting surface geophysical surveys. Log data are repeatable over a long period. The repeatability and comparability provide the basis for measuring changes in exploitation wells. Changes in quality of casing, such as grade of corrosion, clogging, or changes in exploitation regime, may also be recorded.

Well logging cannot replace sampling completely. Background information is needed on each new geologic environment to aid log analysis. A log analysis cannot evaluate a set of logs properly without information on the local geology. Logs do not have a unique response, so it is necessary to combine more methods. And the limitation of logging measurements, which should be considered, is the borehole environment (diameter, mud, dimension of a probe, temperature, pressure, velocity of measurement). High technology and sophisticated interpretation procedures can overcome these limitations.

Principles

Well logging uses the principles of almost all methods in geophysical surface surveys: electrical, nuclear, seismic,

geothermal, gravity, magnetic, and electromagnetic and additionally some procedures particular to drilling activity. The latter involve measurement of borehole diameter, bearing and deviation of the borehole from vertical, the strike and dip of beds, and the direction and velocity of fluid moving within a borehole. The instruments and measurements have to be adapted for the special conditions of the subsurface, particularly the confined space of a borehole, where they also have to overcome the alterations produced in the surrounding formation by the drilling. The drilled holes occur from a few meters depth for engineering purposes, to over 10 km for the crust at depth investigation. Most important are those for hydrocarbon exploration and extraction, which commonly reach depths of a few kilometers. The boreholes can be vertical, oblique, even horizontal and geophysical measurement can be also carried out between holes or between hole and the surface: nearwell and crosswell reflection imaging and tomography.

The variation of a property down a borehole is recorded against depth as a log. Most logs are run by lowering the tool (measuring device) on a cable, often called wireline. In some situations, especially if the borehole is deviated far from the vertical, or may be unstable, the tool is placed in a modified section of drill pipe, called a collar, and lowered into a hole at or near the end of the drilling string.

Many geophysical logging techniques were developed in the petroleum industry, which is the most important application in the exploration, evaluation, and production of oil and gas, by providing information on porosity, permeability, fluid content, and saturation of the formations penetrated by a borehole, but logging methods have also been applied in holes drilled for other purposes. Logging for uranium and coal is routinely used because it is relatively easy to identify and evaluate uranium and coal beds for their quality and thickness. Evergrowing applications of logs can observed in water wells for evaluation of porosity and/or fractures, hydraulic parameters and water quality and groundwater dynamics. Shallow holes drilled for the characterization and remediation of contaminants has attracted some new logging techniques. Metal-mining applications continue to rely on the acquisition of core because borehole elemental assay methods have only been developed for a few specific cases. Core is also required for civil engineering (geotechnical) investigations: holes are generally shallow and core makes use of mechanical testing, whereas logging determines rock quality, namely, fractures under in situ conditions. Subsurface temperature measurements serve as critical input to many fields of engineering, exploration, and research such as gas and fluid production engineering, in well completions, in the exploration for hydrocarbons and ore minerals and for high-temperature geothermal resources as well, for testing hypotheses concerning the evolution of the Earth's crust and tectonic processes and for corrections of many temperature-sensitive borehole measurements. Various deep boreholes have been drilled for scientific purposes. Logging technology facilitates data generation for special geological studies concerning sedimentology and stratigraphic correlation, physical properties, mineralogy and geochemistry, fracture and stress, etc.

Instrumentation

Geophysical logging equipment appears complex, and the variety of instruments in use is very large. Any basic measuring system (measuring a natural physical field) consists of a sensor, signal conditioners, and recorders. Logging probes, also called sondes or tools, enclose the sensors, electronics for transmitting and receiving signals, and power supplies. Often there is an isolated artificial signal source in the lower part of the probe, which prevents the direct flow of the signal between source and sensor(s), so that the signal has to go through the rock surrounding the well. The probes (must be watertight and resistant to a high pressure and temperature) are connected to a cable by a cable head screwed onto the top of the probe. Some electric- and acoustic-logging devices are flexible; most are rigid. Probes vary in diameter from 25 mm to 150 mm. The probes are either axially symmetric (centralized) or use bowsprings or springloaded arms to force the probe sensors pad or pads against the borehole wall (side collimation). Some tool arms may be hydraulically driven. All producers publish temperature, pressure, and diameter specifications for their sondes. Because the diameter is limited and much components must be installed in the sonde, in some cases sondes are very long (some meters) even if a large distance between a source and detector is not required. Usually, in deep holes, the probes are stacked to obtain multiple measurements for economical reasons. Measurements are usually made with the probe pulled upward by a winch fitted with a depth counter to insure steady tension on the logging cable when data are recorded. This results in more uniform speed and more accurate depth determination.

The most recent logging systems are fully digital, with control of the logging equipment implemented from a computer keyboard. Logs are displayed by the computer, and a field record is obtained from a printer. In logging of oil or gas wells, the measured data are telemetered to a computation center where they are promptly evaluated automatically.

The well-log equipments are mounted on vehicle chassis, lighter ones (with cable length $\leq$ 500 m) on minibus chassis, while the larger ones (with cable length 1,000–3,000 m) are truck-mounted and those for logging deep wells (to a depth over a 5,000 m) are mounted on heavy-duty trucks. In equipments of medium and large size, the winches are driven by the vehicle motors, in light equipments by separate electric motors or hydraulically. Portable devices are used for logging mine and engineering-geological boreholes of difficult access. They are provided with a logging cable up to 300 m long. The devices for logging in coal mines must be sealed to prevent electrically detonated explosions.

The most widely utilized logging methods are summarized in Table 1. The principles and applications of quoted methods in the Table 1 are explained at length in literature (Serra, 2007; Serra and Serra, 2004; Hearst et al., 2000; Asquith and Gibbon, 1982; Keys, 1997; Chapellier, 1992; Schlumberger, 1987, 1989; Serra, 1986; Rider, 1986; Desbrandes, 1985; Serra, 1984; Pirson, 1977, etc.)

Geophysical Well Logging, Table 1 The most commonly used logs

Type of log	Potential uses	Borehole conditions required	Limitations
Spontaneous potential SP	Lithology, correlation, water quality	Uncased, mud filled only	Needs contrast in solutes in borehole versus formation
Normal and/or lateral resistivity logs (non focused log), Ra	Lithology, correlation, water quality	Uncased, fluid filled only	Influenced by fluid resistivity and bed boundaries
Conductively focused-current logs, LL Micro-focused logs, MLL	Lithology, correlation, water quality, monitoring formation fluid conductivity	Uncased, fluid filled only	Removes some but not all fluid and bed boundary effects
Inductively focused-current logs IL	Lithology, correlation, water quality, monitoring formation fluid conductivity	Uncased or cased with PVC or fiberglass, air or fluid filled	Poor response in massive resistive formations
Gamma-ray log GR Spectral GR	Lithology, correlation	Most conditions	Unusual lithology interpretations in glauconite, arkose and phosphate sands
Gamma–gamma log GGL-D	Bulk density, porosity, moisture, lithology, correlation, well construction	Wide range	Mineralogy may influence calibration through Z effect
X-ray fluorescence log XRF	Elemental composition of ores	Slim uncased holes, constant diameter, pure water or dry	Very short radius of investigation, mineralogy affects the sensibility and accuracy
Neutron log NNL Pulsed neutron porosity logging	Saturated porosity, moisture, lithology, correlation	Wide range, preferably in slim holes, fresh or brackish fluid filled	Ineffective porosity if clay minerals are included in total porosity
Spectroscopy of neutron-induced γ-rays	Mineralogy from elemental analysis	Wide range	Mineralogy may influence calibration
Nuclear magnetic resonance logging NMR	Amount of free fluids in the reservoir rocks	Uncased	Occurrence of magnetic minerals
Elastic wave propagation log (Acoustic log) AL	Saturated porosity, lithology, correlation	Uncased, fluid filled only	Measures primary or matrix porosity rather than total porosity, soft formations
Cement bond log CBL	Fluid content differentiation, bonding of casing	Uncased or cased, fluid filled only	Response gives a circumferential average of bonding, may not detect channels
Acoustic televiewer AT	Character and orientation of fractures, solution openings, and bedding	Uncased, fluid filled, water or mud	Fracture interpretation may be affected by drilling damage to borehole wall
Borehole television BTV	Well construction, secondary porosity, lost objects	Uncased or cased, air or clear fluid	No information in opaque fluid, dark mineral bands mimic fractures
Caliper log CL	Borehole diameter, log correction, well construction, fractures, lithology	Most conditions	Probe performance often affected by borehole deviation
Temperature log TM	Fluid movement in the borehole and/or behind casing, gas inflow, heat flux determination	Water filled	Affected by disturbance to fluid column by probe motion
Fluid conductivity log FRL	Fluid movement in the borehole	Water filled	Affected by disturbance to fluid column by probe motion
Flowmeter FM	Vertical fluid movement	Fluid filled	Spurious results may be cause d by convection, or by movement of fluid in annulus behind screens
Dipmeter and Borehole image technologies	Dip of strata, features of structural geology	Uncased, fluid filled only	
Clinometer	Orientation of the Dip log tool, deviation of the hole from vertical	Non metal casing, wide range	
Well completion methods	Well construction, numerous techniques for checking casing and completion materials, borehole deviation, etc.	All conditions	Labor and equipment intensive; completions may be irreversible and often depend on the specific interpretation

Fundamentals of log analysis

The process of log analysis for a project can be simplified into several basic steps:

- Data processing, including depth matching and meeting all logs and other data from a single well and editing and smoothing the data set
- Correcting borehole effects and other errors
- Converting log measurements to geologic parameters of interest; this is best done with calibration data for each probe
- Combining logs and other data from all wells on a project, so that the data can be extrapolated laterally in cross sections or on maps
- Writing a report describing the analytical procedures and results from the logs

The analysis of log data by computer offers a number of time-saving advantages: a large mass of data can be collated and displayed; logs can be corrected and replotted; scales can be changed, smoothing and filtering operations can be carried out; cross plots can be made between different kinds of logs and core data and calibration curves; correlation functions and rations can be plotted, as well as cross sections and isopach maps. An excellent introduction into this matter is presented by Doveton (1994).

Cross plots

Because valuable quantities can only be deduced from two or more physical properties, a number of logs are run; often several that will function together are combined in a single tool. Frequently used interpretation technique for more reliable estimation of resulting data is called cross plots. Fertl (1981) gives a good overview of the crossplot technique. The basic procedure is to plot log data for a given depth interval in a space defined by two log parameters. That is, one plots all pairs of observed values of neutron versus acoustic or density parameters, or density versus acoustic. Often, at each point in the plane of coordinate system corresponding to a measured pair of values, one plots a digit representing the number of times that pairs of values occur in the interval. Sometimes, the value of a third variable, such as the gamma-ray log, is measured at the same depth interval as the pair of data being crossplotted or is indicated by a color at the point in space; that is called a z-plot. This is particularly convenient in more complex lithologies.

Applications of geophysical well logging

In the oil industry

The *spontaneous potential* (*SP*) *log* represents the measurements of the naturally occurring potential difference between two electrodes: one is movable in the well, the other is fixed (grounded) on the surface. It is measured in millivolts. Electric currents arising primarily from electrochemical factors within the borehole create the SP log response. These electrochemical factors are brought about by differences in salinities between mud filtrate (R_{mf}) and formation water resistivity (R_w) within permeable beds. Because a conductive fluid is needed in the borehole for the SP log to operate, it cannot be used in nonconductive (i.e., oil-based) drilling muds. The concept of static spontaneous potential (SSP) is important because SSP represents the maximum SP that a thick, shale-free, porous, and permeable formation can have for a given ratio between R_{mf}/R_w. SSP is determined by formula $SSP = -K.\log (R_{mf}/R_w)$ [here R_{mf} is derived from the resistivity of mud, which is measured in the laboratory on a sample, while K has a value that depends on the temperature, it is 71 at the typical temperature of 25°C] or chart and is a necessary element for determining accurate values of R_w and volume of shale. The record of SP shows permeable beds and their boundaries and determines formation water resistivity and also volume of shale in permeable beds.

The *resistivity log* is used to determine the fraction of pore space that is occupied by water, called the water saturation, S_w (the fraction that is hydrocarbons $S_{hc} = 1 - S_w$). Because the rock's matrix or grains are commonly nonconductive, the ability of the rock to transmit a current is almost entirely a function of water in pores. Hydrocarbons are also nonconductive; therefore, as the hydrocarbon saturation of the pores increases, the rock's resistivity also increases. The most simple (standard) form of resistivity logging requires three electrodes that are lowered into a borehole, usually two current electrodes (a pair) A, B, and one measuring electrode M. The measuring electrode N is placed on the surface. The electrodes that are lowered into the borehole consist of lead rings fastened to the lower part of the cable or bar. Such an arrangement of electrodes is known as the resistivity logging probe. The contact between electrodes and formation is made through mud, which therefore must be conducting. The resulting curve is called apparent resistivity R_a. The true resistivity R_t would be measured by this simple sonde only in ideal case of homogeneous isotropic rock environment, which case in the borehole does not exists. The shape of resistivity curves obtained by logging is largely controlled by the configuration of electrodes. There are two basic types of simple resistivity probes, that is, *normal and lateral probes*. Normal probes have a distance between unpaired electrodes far smaller than between paired electrodes and represent the classical way of measurement. Measuring a set of R_a curves with different spacing is required to obtain a plot of values of R_a versus probe spacing. Connecting them by a line, we can correlate in this way obtained curves with theoretical curves (master curves) and thus to determine the true resistivity of the layer investigated, the resistivity and the diameter of the invaded zone. Since this process is extremely laborious, a method was sought which would replace this classic resistivity measurement and enable the true resistivity of rocks to be determined in a simpler way.

Research led to the development of a combination tool, known as laterologs, which consist of a central electrode and symmetrical placed other (shielding) electrodes.

They have great advantage as the registered R_a values are closed to the true resistivity (Hearst et al., 2000; Schlumberger, 1989; Serra, 1986). Examples are the microlog, microlaterolog, proximity log, and induction log.

The *radioactivity logs* involve the use of radioactivity to identify lithologies and to help deduce density and porosity, particularly when a hole is cased because they will function with steel case in place. The simple *natural gamma ray-log* responds to total natural gamma ray activity of the surroundings rocks in the borehole, which is due mainly to potassium (K), thorium (Th), and uranium (U, resp. Ra). Of lithologies commonly present, usually shales have the highest activity; sands are intermediate; limestones, dolomite, and quartz are low; while anhydrite, salt, coal, and chert are least active. Occasionally, potash beds, and accumulations of radioactive ores or minerals with higher activity than shales may be encountered. Although the gamma ray activity of shales varies widely on worldwide basis, it tends to be constant in a particular field. It is therefore used to estimate shale content, assuming the maximum value on the log means 100% shale and the minimum value means no shale. The *spectral gamma ray log* records separately the activities of K, Th, and U, so allowing more precise identification of shales, for these differ from other rocks in their relative proportions of the three elements. The *gamma–gamma or formation density log* (GG-D) has a strong source of gamma rays (mostly 137 Cs) at one end of the probe, at the other is a detector of the gamma rays that is shielded so that it cannot receive rays directly from the source, so only gamma rays that results from scattering from the surrounding rock are counted. The *neutron–neutron log* (NNL) is a record of the neutron flux after its passing from a source in the probe through a formation continuous with the borehole to the detector(s). The tool is equipped with a radioactive source that emits fast neutrons. Neutrons travelling through the formation only slow down significantly when they collide with atoms of a similar mass, that is, hydrogen atoms. That is why the measured value of NNL corresponds to the content of hydrogen (not only as its part of water or oil, but independently in the form of hydrogen occurrence in some minerals), referred to as hydrogen index. Other varieties of neutron–neutron log can be used, interrelated with application of different i) sources such as plutonium/beryllium or americium/beryllium or californium sources using neutron accelerators working in pulse regime and ii) detectors besides neutron tubes for epithermal and thermal neutrons using spectral detection of gamma rays produced as a result of few usual interactions of neutrons with formation (nonelastic collisions, neutron absorption, activation).

The *sonic log* measures a record of the elastic waves velocity versus depth in a borehole (see *Seismic Properties of Rocks*). The most simple tool design is single-transmitter tool, which commonly has a radial piezometric, electrostrictive, or magnetostrictive transducer placed along the axis of an acoustic tool. Receivers are mounted in much the same way. The borehole fluid must be a liquid for most tools to function. Care must be taken to acoustically isolate the transmitter from the receiver so signals are not transmitted through the tool (compare with *Seismic instrumentation*). This measuring system fails if the tool is not parallel to the borehole wall or if there are increases in hole diameter. To solve those problems, the *compensated acoustic log* was developed. It has two transmitters at each end, and for each tramsmitter there are two receivers positioned far enough along the tool that the first arrivals at both are rays refracted from the wall of the borehole. The choice of tool frequency is critically important for the generation of useful acoustic logs. High frequency and short wavelength are characterized by detailed spatial resolution and strong attenuation. Low frequency and long wavelength are characterized by poor spatial resolution but low attenuation. Careful study of the effects of frequency on acoustic logging indicates that the most important requirement for logging tool performance is excitation of a single mode. Hearst et al. (2000) find that the most commonly used acoustic-logging frequencies (10–20 kHz) correspond with the source band most likely to excite the single fundamental pair of P and S peaks under typical logging conditions for 15–25 cm diameter boreholes. There are many variations of the continuous velocity log, but all rely more or less on the same principle, though the number and type of sources and receivers are variable. The advent of compact downhole electronics controlled by powerful computers in the logging truck has offered the collection of digital waveforms from logging tools consisting of many receivers. Thus, more information than just the first arrival can be presented, and several methods of presenting the information have been developed: the full-wave display, variable-intensity display. It is now common practice to record the entire waveform digitally for later processing (full-wave acoustic logging). Tool using this mode of acoustic propagation measure both the velocity and the amplitude of the various components of the full acoustic waveform, that is, compressional, shear, and tube (Stoneley) waves (Paillet et al., 1992; Chen, 1989). Advances in characterization of full-waveform data facilitate fracture detection and evaluation, permeability measurement, and well completion. The measured data by continuous velocity log can help identify lithologies although the main application is to estimate the porosity. Other developments in acoustic logging are depicted in publication of Prensky (1994): *reflexion (Pulse-echo) tool* which employs ultrasonic pulse-echo transducers (borehole-televiewer type) that measure signal velocity and amplitude and have application in both open and cased hole. Further advances in acoustic logging are improvements in techniques such as *Downhole seismic, Wellbore resonance, Behind-pipe flow*, etc.

The *temperature log* provides the values of temperature of fluids in a borehole (which is not necessarily in thermal equilibrium with the formation temperature) with depth (in °C) or the temperature increment over a constant distance, that is, the temperature gradient (in $^{\circ}C.m^{-1}$).

The instruments used are thermometers and differential thermometers. As a measuring element, temperature-dependent resistors are used. They are located in the probe so as to be in contact with the fluid with their entire surface, if possible, and to be simultaneously thermally insulated from the body of the measuring probe. The time constant of the device is thus reduced to a minimum. There are two types of temperature-dependent resistors: *thermally positive* (metal wires, e.g., platinum, copper, etc.) and *thermally negative* semiconductor materials (thermistors). These thermally negative materials exhibit substantially higher sensitivity (by as much as 10 times) than the thermally positive materials, but their drawback is that their resistance varies with temperature nonlinearly (in contrast to metals). The temperature sensor independent of its material is usually inserted in a Wheatstone bridge so that temperature-dependent resistor is located in one branch of the bridge, and remaining three branches consist of resistors independent of temperature. For calibration, the logging thermometers are put into a water bath whose temperature is changed gradually and measured as accurately as possible. Desired accuracy is of the order of 0.01°C, the effective accuracy of commercial temperature log is 0.5°C. Very often, the temperature is recorded during the lowering of the probe so as to meet a thermal field with the minimum possible disturbance. The application of borehole-temperature data can be grouped in two broad categories: engineering uses and scientific uses (Prensky, 1992). Engineering uses incorporate: (1) correction of resistivity logs because temperature normally increases with depth and resistivity is a function of temperature (and salinity); (2) reservoir engineering: temperature is required for calculations of hydrocarbon recovery factors, including pressure–volume–temperature relationships and gas–oil ratios; (3) well completion and production logging: this issue includes the production of hydrocarbons and geothermal energy, location of top of casing cement (owing an exothermic reaction of setting cement producing an obvious temperature anomaly), designing cement slurries, detecting channels in the cement, Fluid flow – temperature anomalies are used for identifying the depth of fluid and gas entry or exit, detection of casing leaks and intervals of lost circulation, fluid injection – the contrast in temperature between injected fluid and formation fluids can identify zones of fractures.

Scientific uses are based on heat-flow measurements in boreholes, which are made by combining sets of temperature measurements under a steady temperature regime and thermal conductivity data, following the formula $Q = K\,(dT/dz)$, where Q is heat flow, K is thermal conductivity (known from literature for a given lithology or measured in laboratory), and T is temperature at depth z (when the temperature of mud has equalized with the temperature of the formation). Temperature in a borehole also generally changes with time as it returns toward equilibrium after the disturbance caused by drilling, which can take long time and may range from a few days for a shallow (100–150 m) air-drilled hole to several months for deep mud-drilled oilwells. The heat-flow knowledge contributes to solving a lot of scientific work, such as study of the Earth's evolution, stratigraphic correlation (changes in thermal gradient may correspond to lithology variations), basin analysis and modeling, detection of zones of overpressuring, detecting of regional aquifer flow, exploration for geothermal reservoir, etc.

Borehole imaging technologies are based on three types of sensors: electrical, ultrasonic, and optical. The *dipmeter* can also be considered a form of imaging. The dipmeter is used to determine the dip angle and dip direction of planar features intercepted by the borehole.

Indirect-imaging technologies: *Acoustic and electrical*

The electrical scanning tools rely on a resistivity contrast in the formation to obtain an image and thus limited to holes with conductive mud. Acoustic-imaging tools use a rotating, focused, ultrasonic transducer to measure reflectivity times and amplitude of the pulse-echo from the borehole wall or from the casing. Acoustic amplitude is sensitive to borehole irregularities, thus enabling determination of borehole shape, that is, borehole caliper. Prensky (1994) gives an excellent summary of the literature as regards the possibilities of these borehole imaging tools for identification and analysis of natural and induced fractures; borehole breakouts; thin beds; geologic structure, such as strike, dip, and faults; identification of sedimentary features; interpretation of depositional environment; and evaluation of casing corrosion.

Direct Imaging Technology: *Borehole video* is a wireline tool containing a miniature television camera and light source to obtain high-resolution, real-time visual images of the borehole or tubulars, in both air- or fluid-filled boreholes. The recent introduction of fiber-optic-based systems offers reduction in tool diameter and operation in greater depths. Applications include inspection of casing, tubing and perforations, detection of produced fluid entry, fracture deification, etc.

The *caliper logs* facilitate measurement of the true size and shape of a drilled hole. Most caliper tools are mechanical. The mechanical tool consists of several arms, which are pushed out against the wall of the hole by springs or hydraulic pressure; their extent is converted by an electromechanical device into an electrical signal suitable for recording (proportional to real diameter). Most independent calipers have three or four arms; some have as many as six. The diameter of the borehole is necessary to be known to introduce corrections in quantitative interpretation in almost all methods measured in the boreholes; to calculate the amount of cement needed to fix the casing; to give lithologic information, because washouts are indicative of formation properties – weak formations will tend to wash out, strong ones will not. Caliper logs can also give evidence of the presence of fractures.

The simplest relation between electrical conductivity and formation properties is given by an empirical relation

between formation factor F, defined as the ratio of the resistivity of a rock fully saturated by water R_0 to the resistivity of the water R_w with which the rock medium is saturated, and porosity Φ of clean sands (known as Archie's law): $F = R_0/R_w = a/\Phi^m$, where the coefficient a 'and the exponent m' (cementation factor) can be determined independently from laboratory measurements on core samples and are usually consistent within a lithologic unit. The Archie's formula has some limitations dealing with conductivity of pore space, which is frequently discussed in literature, for example in Hearst et al., 2000. *Dielectric log* is an electromagnetic method using high frequencies, when dielectric properties begin to dominate the propagation characteristics of electromagnetic waves. The presence of water continues to govern the tool response in this high-frequency regime (100 kHz and greater) because the dielectric permittivity of water is significantly greater than that of most earth materials. *Pulsed neutron logging* measures the rate of decay of thermal neutrons in the interval between two pulses of a neutron generator. *Nuclear magnetic resonance logging* (NMR) has been available for many years, since 1960s, but the literature on the applications of borehole NMR has grown very rapidly in the 1990s, and is constantly increasing. Originally, the method is based on the measurement of nuclear spin and magnetic moment of a hydrogen nucleus providing data of the free fluid index – how large a volume of fluid can be obtained from the rock by pumping, and the character of the fluid at the selected depths. The total signal is proportional to the total number of protons present in both the free and bound fluid. NMR measures proton spin–lattice (longitudinal, T1) and spin–spin (transverse, T2) relaxation times; these relaxation times are proportional to pore size and to bound (clays) and free (pore) fluids. In petrophysical terms, these parameters provide information for quantitative determination of permeability, pore-size distribution lithology in reservoir rocks. Kenyon (1997) is an excellent summary of the application of NMR to petroleum industry.

Porosity derived from logs shows summary in the Table 2.

Hydrogeologic applications

Geophysical well logs in water wells can be used to determine lithologic parameters with the aim to evaluate quality and geometry of aquifer system, both with intergranular porosity and fracture and estimate the quality and amount of contained water in the surroundings rocks. In addition to the lithologic parameters, a second important category of information that can be derived from well logs includes data on the character and movement of water in the borehole and the formation. Characteristics of fluids in the well bore can be measured directly, but the data on formation fluids must be inferred from logs. *Logs of the conductivity and temperature* of fluid in the borehole can be related to formation-fluid characteristics, provided that the well has had time to attain chemical and thermal equilibrium with the groundwater reservoir, that an adequate hydraulic connection exists between the hole and the rocks penetrated, and that the inhole flow does not disturb these conditions. *Flowmeters* (Schlumberger, 1993) are probes designed to measure groundwater flow rates along boreholes under ambient and forced gradient (during pump-test) conditions. Impellor flowmeters (Hill, 1990; Keys and Sullivan, 1979) require a significant velocity to rotate impellor blades, so that this probe cannot measure velocities less than about 10^{-1} m/s. Heat pulse (Hess, 1986) and electromagnetic (Molz et al., 1994) flowmeters can detect velocities as low as 10^{-3} m/s.

Techniques based on *time sequence measurements of flow logs* are applied only in clean wells, where mud has been replaced by water and the filter cake does not affect aquifer permeability in the vicinity of the borehole. Under these conditions, changes in the physical properties of the borehole fluids under ambient or stressed conditions reflect the properties of the formation and the water contained in the formation. Various properties of water

Geophysical Well Logging, Table 2 Summary of porosity derived from logs

Log	Property measured	Response to the type of porosity	Extraneous effects	Borehole requirements
Gamma–gamma	Electron density	Total porosity, best for high porosity	Matrix, salinity, water saturation, well construction	Wet or dry, cased or open
Neutron	Hydrogen content	Total porosity, best for low porosity	Bound water, matrix chemistry, salinity, water saturation	Wet or dry, cased or open
Acoustic velocity	Average P-wave velocity	Total porosity, only primary and intergranular	Cycle skipping, poor in high-porosity sediments	Fluid filled and open, high pressure
Resistivity	Pore fluid and interconnected porosity	Effective porosity both primary and secondary	Pore shape, fluid chemistry, water saturation	Open or with perforated PVC casing
Gravity logging	Bulk density	Total porosity	Stationary measurement	Deep reading, >50 ft (18 m)
Nuclear magnetic resonance	Relaxation time of protons	Amount of fluid in pores larger than some minimum	Bound water, very small pores, clay and magnetic minerals	Open hole

in boreholes, such as the conductivity, temperature, transparency, and/or radioactivity (using tracers) as a function of depth, may be used in this fluid column logging method (Chapellier, 1992; Kobr et al., 2005). Many of these properties can be measured using conventional borehole logging equipment. Log-derived hydraulic parameters, such as hydraulic conductivities or transmissivities can be obtained from both flowmeter analysis and time sequence of fluid logs. The last named method is also used to study water flow in the rock medium. Movement of the groundwater flow in sandy aquifers is oriented in the direction of the hydraulic velocity, by the hydraulic conductivity and the azimuth (direction) of the flow. Properly interpreted borehole fluid log can be used to locate zones where water enters and/or leaves the well bore, to estimate the specific yields of these zones and all parameters cited above. Also, this methodology is advantageous in contamination studies, to check the quality of well completion and in shallow boreholes for engineering geology (Hess, 1986; Tsung et al., 1990; Kelly and Mareš, 1993; Mareš et al., 1994; Molz et al., 1994; Paillet and Crowder, 1996; Keys, 1997; Paillet, 2000).

Mineral exploration

Although boreholes are routinely cored to detect and quantify the presence of economic minerals, often the information is incomplete because of missing core, or misleading because of rapid lateral variation of the mineralization (hydrothermal veins). The depth of investigation may be increased by lowering down the hole closer to the ore bodies. The logging system is adapted to the conditions of a mineral exploration program (smaller holes as diameters – from 49–122 mm; as depths – often less than 100 m) using portable logging system with appropriate methodology. The physical properties of interest are density, average mass number, electrical conductivity, and magnetic susceptibility. For disseminated types of ores induced polarization method is useful. Sometimes, self potential method is also used. In this case, the spontaneous potential is of interest in connection with minerals that are electronic conductors, such as massive sulfides, magnetite, and graphite. If two boreholes are available at reasonable mutual distance a cross-hole tomographic survey can be carried out either with seismic or electromagnetic method. This later can map the conductivity distribution between boreholes in the exploration for minerals as well as in environmental and crustal studies (see *Instrumentation, EM*). The modern radioactive methods (X-ray fluorescence analysis, spectroscopy of neutron-induced γ-rays) enable implementation of elemental analysis of the rocks in situ conditions. Due to heavy expenses involved, these methods are not usually considered as the routine ones in mineral exploration. But in some special conditions they may be economic (for example, neutron activation in the borehole on fluorine can be used for calculation of reserves).

Civil engineering, rock mechanics, and environmental preservation

Other well-logging applications for various geological studies can be found in literature (Deltombe and Schepers, 2004; Hurst et al., 1992; Hurst et al., 1990). Almost all methods discussed earlier can be applied, some modifications have been introduced. The most important methods for this reason are the acoustic methods: *sonic log, vertical seismic profiling, seismic tomography.* The main goal of this application is direct (in situ) determination of dynamic elastic properties because sonic logging and waveform analysis provide the means for obtaining continuous measurements of compression and shear velocities. More information can be found in Takahashi et al. (2006).

Summary

Geophysical logs are made by instruments in sondes, or tools, suspended mostly on a wire, with readings usually taken as they are pulled to the surface. Sondes often contain several instruments that can operate without mutual interference. The records of the variation of various quantities down boreholes are called logs, and are of both geophysical and nongeophysical quantities. Geophysical logs are often more valuable when combined than when used singly. Geophysical logging is not only cheaper than continuous coring but it provides information that cannot be obtained on cores, partly because of the alterations produced by the dribbling, and reduces the need for sampling.

Measurements of geophysical parameters in boreholes present the follows characteristics:

- Instruments have to be adapted for the dimensions of a borehole and usually to operate submerged in drilling mud.
- Measurements have to take into account the conditions (pressure, temperature, and technical conditions of the borehole) resulting from the drilling.

The most important application of well logging in the exploration, evaluation, and production of oil and gas, as well as other uses in different geological and engineering branches has been discussed.

Bibliography

Asquith, G., and Gibbon, C., 1982. *Basic Well Log Analysis for Geologists*. Tulsa: Association of Petroleum Geologists.

Chapellier, D., 1992. *Well Logging in Hydrogeology*. Rotterdam: A. A. Balkena.

Chen, S. T., 1989. Shear wave logging with multi-pole sources. *Geophysics*, **54**, 509–517.

Deltombe, J. L., and Schepers, R., 2004. New developments in real time processing of full waveform acoustic televiewer data. *Journal of Applied Geophysics*, **55**, 161–168.

Desbrandes, R., 1985. *Encyclopedia of Well Logging*. Paris: Editions Technip.

Doveton, J. H., 1994. *Geologic Log Analysis Using Computer Methods*. Tulsa: AAPG.

Fertl, W. H., 1981. Open hole crossplot – a powerful technique in well log analysis. *Journal of Petroleum Technology*, **33**(3), 535–549.

Hearst, J. R., Nelson, P. H., and Paillet, F. L., 2000. *Well Logging for Physical Properties*. Chichester: John Wiley.
Hess, A. E., 1986. Identifying hydraulically conductive fracture with a slow velocity borehole flowmeter. *Canadian Journal of Earth Sciences*, **23**, 69–78.
Hill, A. D.,1990. *Production logging-Theoretical and interpretive elements*. Richardson, Texas: Society of Petroleum Engineers, Monograph 14.
Hurst, M. A., Lovell, M. A., and Morton, A. C. (eds.), 1990. *Geological Application of Wireline Logs*. London: The Geological Society.
Hurst, M. A., Griffiths, C. M., and Worthington, P. F. (eds.), 1992. *Geological Application of Wireline Logs II*. London: The Geological Society.
Johnson, H. M., 1962. A history of well logging. *Geophysics*, **27**(4), 507–527.
Kelly, W. E., and Mareš, S., 1993. *Applied Geophysics in Hydrogeological and Engineering Practice*. Amsterdam: Elsevier.
Kenyon, W. E., 1997. Petrophysical principles of applications of NMR logging. *The Log Analyst*, **38**(2), 21.
Keys, W. S., 1997. *A Practical Guide to Borehole Geophysics in Environmental Investigations*. Boca Raton, FL: CRC Press, Lewis Publisher.
Keys, W. S., and Sullivan, J. K., 1979. Role of borehole geophysics in defining the physical characteristics of the raft river geothermal reservoir. *Geophysics*, **44**, 1116–1141.
Kobr, M., Mareš, S., and Paillet, F., 2005. Borehole geophysics for hydrogeological studies: principles and applications. In Rubin, Y., and Hubard, S. (eds.), *Hydrogeophysics*. Dordrecht: Springer-Verlag, pp. 291–331.
Mareš, S., Zbořil, A., and Kelly, W. E., 1994. Logging for determination of aquifer hydraulic properties. *The Log Analyst*, **35**(6), 28–36.
Molz, F. J., Bowman, G. K., Zouny, S. C., and Waldrop, W. R., 1994. Borehole flowmeters – field applications and data analysis. *Journal of Hydrology*, **163**, 347–371.
Paillet, F. L., 2000. A field technique for estimating aquifer parameters using flow log data. *Ground Water*, **38**(4), 510–521.
Paillet, F. L., and Crowder, R. E., 1996. A generalized approach for the interpretation of geophysical well logs in groundwater studies. Theory and application. *Ground Water*, **34**(5), 883–898.
Paillet, F. L., Cheng, C. H., and Pennington, W. D., 1992. Acoustic-waveform logging-advances in theory and application. *The Log Analyst*, **33**(3), 239–258.
Pirson, S. J., 1977. *Geologic Well Log Analysis*. Houston: Gulf Publishing.
Prensky, S., 1992. Temperature measurements in boreholes: an overview of engineering and scientific applications. *The Log Analyst*, **33**(3), 313–333.
Prensky, S. E., 1994. A survey of recent developments and emerging technology in well logging and rock characterization. *The Log Analyst*, **35**(2), 15–45. No. 5, pp.78–84.
Rider, M. H., 1986. *The Geological Interpretation of Well Logs*. Blackie, Glasgow & London, New York: Halsted Press.
Schlumberger, 1987, 1989. *Log Interpretation: Principles/Applications*. SMP/7017. Houston: Schlumberger Educational Services.
Schlumberger, 1993. *Production Log Interpretation*. Houston: Schlumberger Educational Services.
Segesman, F. F., 1980. Well-logging method. *Geophysics*, **45**(11), 1667–1684.
Serra, O., 1984. *Fundamentals of Well-Log Interpretation. Vol. 1: The Acquisition of Logging Data*. Developments in Petroleum Science, 15A. Amsterdam: Elsevier.
Serra, O., 1986. *Fundamentals of Well-Log Interpretation. Vol. 2: The Interpretation of Logging Data*. Developments in Petroleum Science, 15B. Amsterdam: Elsevier.
Serra, O., 2007. *Well Logging and Reservoir Evaluation*. Paris: Editions Technip.
Serra, O., and Serra, L., 2004. *Well Logging: Data Acquisition and Interpretation*. France: Serralog.
Takahashi, T., Takeuchi, T., and Sassa, K., 2006. ISRM suggested methods for borehole geophysics in rock engineering. *International Journal of Rock Mechanics and Mining Sciences*, **43**(3), 337–368.
Tsung, C. F., Hufschmied, P., and Hale, F. V., 1990. Determination of fracture inflow parameters with a borehole fluid conductivity logging method. *Water Resources Research*, **26**, 561–578.

Cross-references

Electrical Properties of Rocks
Electrical Resistivity Surveys and Data Interpretation
Gravity Method, Principles
Heat Flow Measurements, Continental
Instrumentation, EM
Magnetic Methods, Principles
Magnetometers
Propagation of Elastic Waves: Fundamentals
Sedimentary Basins
Seismic Instrumentation
Seismic Properties of Rocks
Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat
Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity

GEOTHERMAL HEAT PUMPS

Ladislaus Rybach
Geowatt AG Zurich, Zurich, Switzerland

Definition

A geothermal heat pump (GHP) is a heating and/or cooling system that moves heat to or from the ground. It uses the earth as a heat source (in the winter) or a heat sink (in the summer). Geothermal heat pumps are also known by a variety of other names, including ground-source, geoexchange, earth-coupled, earth energy, or water-source heat pumps. They can be designed and installed in sizes from a few thermal kW to several MW capacity (the latter in modular assemblage). Currently GHPs represent the largest share in geothermal direct use and account mainly also for the global growth.

The resource

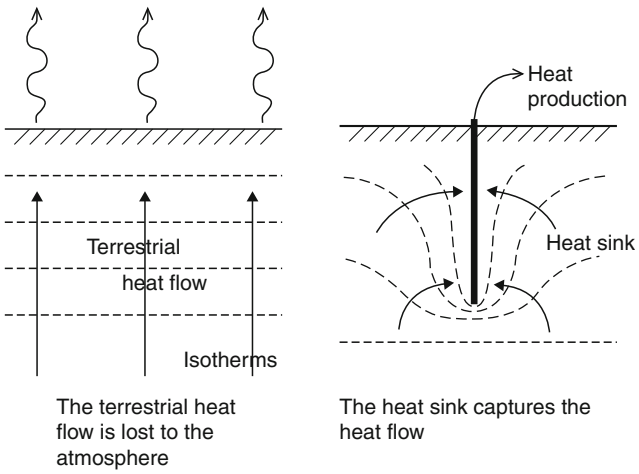
Shallow geothermal resources (the heat content of rocks in the top few meters of the continental crust) represent a major and ubiquitous energy source. The earth as planet can afford to give off heat by a thermal power of 40 million MW, without cooling down. Without utilization, the terrestrial heat flow is lost to the atmosphere. In this case, the isotherms run parallel to the earth's surface (i.e., horizontal in flat terrain) and the heat flow lines are perpendicular to them. If, instead, the heat flow can be captured, e.g., by a heat extraction device like

a borehole heat exchanger (BHE) (see later), the isotherms are deformed and the heat flow lines can be diverted toward heat sinks (Figure 1).

Geothermal heat pump (GHP) technology

A new chapter in the history of geothermal applications in buildings opened with the advent of geothermal heat pumps. This technology enables space heating, cooling, and domestic warm water production with the same installation. The GHP application is now in the focus of private, public, and municipal interest (Banks, 2008).

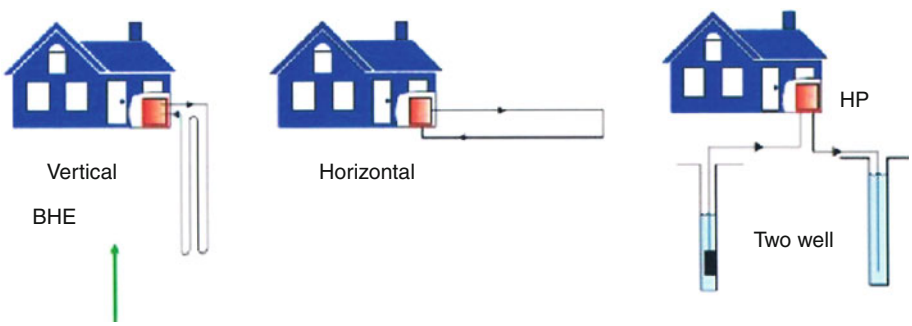
Geothermal heat pumps (GHPs) are one of the fastest growing applications of renewable energy in the world and definitely the fastest growing segment in geothermal technology, in an increasing number of countries (Rybach, 2005). GHP represent a rather new but already well-established technology, utilizing the immense renewable storage capacity of the ground. GHPs use the relatively constant temperature of the ground to provide space heating, cooling, and domestic hot water for homes, schools, factories, public buildings, and commercial buildings. The applicational size can vary from single-family homes with 1–2 borehole heat exchangers (BHEs) to large-scale complexes with hundreds of BHEs. The decentralized systems can be tailor-made, taking into account the local conditions. It is essential to employ proper installation design that takes into account meteorologic, ground property, and technical supply conditions. By these means, reliable long-term operation can be secured. Of the local conditions, the thermal conductivity of ground materials and the groundwater properties are of key importance.



Geothermal Heat Pumps, Figure 1 Principle of geothermal heat extraction and production.

There exist mainly two types of geothermal heat pumps (Figure 2). In ground-coupled systems, a *closed loop* of plastic pipe is placed in the ground, either horizontally at 1–2 m depth or vertically in a borehole down to 50–300 m depth. A water-antifreeze solution is circulated through the pipe. Thus, heat is collected from the ground in the winter and optionally heat is rejected to the ground in the summer. An *open loop* system uses groundwater or lake water directly as a heat source in a heat exchanger and then discharges it into another well, a stream, or lake, or even into the ground. The installation of horizontal coils needs relatively large surface area and extensive earthworks (digging the ground down to the level of coil layout); the prerequisite for extracting the heat of groundwater is the presence of a shallow water table. For these reasons, the most widespread technology of shallow heat extraction is by BHEs. Heat extraction is established by closed-circuit fluid circulation (a few m^3/h of pure water or with an antifreeze additive) through the BHE and the evaporator side of the heat pump.

Three basic components make up a GHP system: (1) the heat extraction/storage part in the ground, (2) the central heat pump, and (3) the heat distributor/collector in the building (e.g., floor panel). The key component is the heat pump. In essence, heat pumps are nothing more than refrigeration units that can be reversed. In the heating mode, the efficiency is described by the coefficient of

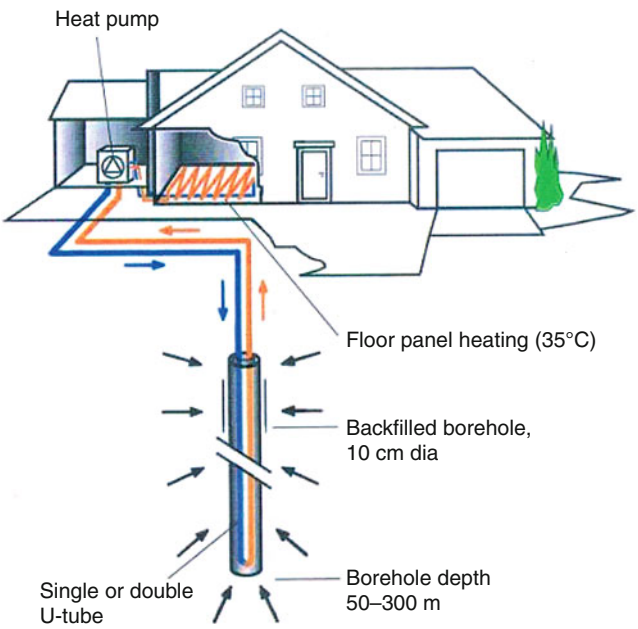


Geothermal Heat Pumps, Figure 2 Closed loop and open loop (groundwater) heat pump systems. The arrow indicates the most common system, with borehole heat exchangers (BHE). The heat pump (HP) is the red box in the house. (Modified from Lund et al., 2004).

performance (COP) which is the heat output divided by the electrical energy input for the heat pump. Typically, this value lies between 3 and 4 (Rybach, 2005). Except for larger, singular applications where gas-driven heat pumps are used, most heat pumps use electricity for their compressors. Therefore GHPs are electricity consumers. The source of electricity varies from country to country; it would be elegant if the electricity to drive the GHP heat pumps would originate from renewable sources like solar, wind, or even geothermal! The principal components of a BHE-based GHP system are depicted in Figure 3.

Heating and cooling with GHPs

As mentioned above, GHP systems can provide space cooling also. In moderate climate, the ground below about 15 m depth is significantly colder than outside air in summer. Thus, a large geothermal store with favorable heat capacity is available where the heat can be exchanged (extracted from the building and deposited in summer, extracted from the ground store and supplied to the building in winter). The thermal capacity of the system depends, besides the volume, on the thermal and hydrogeologic characteristics of the installation site; these must be carefully considered in system dimensioning. In summer, most of the time the heat pump can be bypassed and the heat carrier fluid circulated through the ground by the BHEs and through the heating/cooling distribution (e.g., floor panels). By these means, the heat is collected from the building and deposited in the ground for extraction in the next winter (Lund et al., 2003).



Geothermal Heat Pumps, Figure 3 Sketch of a GHP system with a single BHE. Horizontal/vertical arrows indicate circulation in the U-tube heat exchanger, oblique arrows indicate heat extraction from the ground (heating mode in winter). In summer the oblique arrows are reversed; heat is extracted from the building and stored in the ground.

Design, costs

BHEs can be installed in nearly all kinds of geologic material. Design and installation must consider numerous influence factors. The sizing must consider the demand characteristics of the object to be supplied (size, extension; heating alone, or heating + domestic water, combined heating/cooling) as well as the local site conditions (climate, ground properties). The key ground property is the thermal conductivity of the ground surrounding the BHE. The higher the rock thermal conductivity λ (W/m,K), the higher the specific heat extraction rate (W/m) and the energy yield (kWh/m,a) per unit BHE length (see Table 1).

The installation and operation of a BHE/HP system needs installation and running costs. BHE drilling and installation costs amount to about 50 US$ per meter in the USA and about 70 € per meter length in Europe. The operational cost is significantly less than with fossil-fueled systems. Currently, the price of heat and cold provided by GHP systems is comparable with conventional systems; GHP installation cost payback (the period of years before the savings in annual operating costs exceed the initial premium of an installation) ranges from 4 to 10 years, depending on the size and complexity of the system.

GHPs represent by far the fastest growing market in geothermal development. In the USA, over 100,000 systems are installed annually (IGSHPA, 2009a). Annual GHP sales reached 35,000 units in Germany in 2008, and 40,000 units in Sweden in 2007 (EGEC, 2009). In Switzerland, over 2,000 km borehole length was drilled in 2008 to install BHEs (Rybach and Signorelli, 2010).

Licensing, environmental benefits

In general, GHP installation needs permits. In most countries, the water protection agencies (local, regional, national) are providing permits. Regulation varies from country to country, in some cases even within the same

Geothermal Heat Pumps, Table 1 Borehole heat exchanger performance (single BHE, depth ~ 150 m) in different rock types. From Rybach (2001)

Rock type	Thermal conductivity ($W\,m^{-1}\,K^{-1}$)	Specific extraction rate (W per m)	Energy yield ($kWh\,m^{-1}\,a^{-1}$)
Hard rock	3.0	max. 70	100–120
Unconsolidated rock, saturated	2.0	45–50	90
Unconsolidated rock, dry	1.5	max. 25	50

country. Usually, some standard forms need to be filled and submitted. It is common practice that in groundwater protection areas the installation of GHP systems is limited or forbidden.

GHPs operate with little or no greenhouse gas (GHG) emissions since no burning processes are involved. GHPs driven by fossil-fueled electricity reduce the CO_2 emission by at least 50% compared with fossil-fuel-fired boilers. If the electricity that drives the geothermal heat pump is produced from a renewable energy source like hydropower or geothermal energy, the emission savings are up to 100%.

Reducing current CO_2 emissions is the central option in all the efforts to mitigate global warming. Here, a clear distinction must be made between actual emission reduction and merely avoidance of additional emission: by new GHPs only additional CO_2 emission can be avoided ("saving"), not a reduction of actual emissions. When GHPs are installed in refurbishment (to replace fossil-fueled systems), actual emission reduction can be achieved. Emission reduction is also evident when electric heater/cooler installations, driven by fossil-based electricity, are replaced by GHP systems.

Applications, development trends

There is a growing number of successful examples of GHP applications, for all kinds of buildings of highly varying sizes. The largest installation currently under construction is the geothermal store at the Bell State University (Indiana, USA) campus, with 4,100 BHEs, each 120 m deep (IGSHPA, 2009b). Several international organizations cover and/or promote the applications of GHPs. Here, only a few links are provided.

- International Ground Source Heat Pump Association (IGSPHA) http://www.igshpa.okstate.edu/
- Geothermal Heat Pump Consortium (GHPC) http://www.geothermalheatpumpconsortium.org/
- International Energy Agency Heat Pump Centre http://www.heatpumpcentre.org/
- European Heat Pump Association (EHPA) http://www.ehpa.org/

Besides, there are national associations active in many countries.

Most GHP systems operate with borehole heat exchangers (BHEs). In the early days of BHE applications, the boreholes have been drilled outside the objects to be served by the GHP installations. Nowadays, also dictated by the high prices of land, the boreholes are more and more located beneath of the buildings to be erected.

Energy piles are now becoming increasing popular. At many locations, the ground properties are such that load-bearing rock formations occur only at some depth. Therefore, foundation piles need to be placed to secure firm standing of buildings. These piles can be equipped with heat exchanger pipes; the assemblage is then called "energy pile." Special care must be taken to secure that the temperature changes caused by heat extraction and storage does not affect the pile/ground contact. A special handbook has been published in Switzerland that describes design, installation, and operation of energy pile systems (SIA, 2005).

Summary

Currently, geothermal heat pumps (GHPs) represent the fastest growing branch in the utilization of geothermal energy. There are various factors that can lead to the successful dissemination of the GHP technology. Promoting factors are manifold: technical, economic, environmental awareness, governmental support, etc. The most important driving force is simply knowledge and know-how. Besides, various applications and wide-scale realizations, high-level quality assurance and successful demonstration facilities are needed.

The use of GHP systems represents an important contribution of renewable energies to the mitigation of climatic warming: the increasing use of the environmentally friendly geothermal technology enables the saving of great amounts of fossil fuels, avoid corresponding CO_2 emissions and, in the case of the application of GHP systems in building renovation, it results in a significant emission reduction. It can be expected that GHP systems, along with other geothermal technologies (especially for power generation), will further increase the benefits to future generations.

Bibliography

Banks, D., 2008. *An Introduction to Thermogeology – Ground Source Heating and Cooling*. Oxford: Blackwell, p. 349.

EGEC, 2009. *Geothermal Geat Pumps – Ground Source Heat Pumps*. Brussels: European Geothermal Energy Council, p. 16.

IGSHPA, 2009a. *GeoOutlook Magazine*,Vol. 6 no. 2.

IGSHPA, 2009b. *GeoOutlook Magazine*, Vol. 6 no. 4.

Lund, J., Sanner, B., Rybach, L., Curtis, R., and Hellström, G., 2003. Ground source heat pumps – A world review. *Renewable Energy World*, July-August 2003, pp. 218–227.

Rybach, L., 2001. Design and performance of borehole heat exchanger/heat pump systems. Oradea/Romania. *Proc. European Summer School on Geothermal Energy Applications* (CD-ROM).

Rybach, L., 2005. The advance of geothermal heat pumps worldwide. *IEA Heat Pump Center Newsletter*, Vol. 23 No. 4, pp. 13–18.

Rybach, L., and Signorelli, S., 2010. Country update of Switzerland. *Proceedings, World Geothermal Congress 2010, Nusa Dua, Bali/Indonesia*, 13 p. (CD-ROM).

SIA, 2005. *Nutzung der Erdwärme mit Fundationspfählen und anderer erdberührten Betonebauteilen – Leitfaden zu Planung, Bau und Betrieb*. SIA D 0190, Zurich: Schweizerischer Ingenieur- und Architektenverein, 101 p., ISBN 3-908483-58-1.

Cross-references

Heat Flow, Continental
Lithosphere, Continental: Thermal Structure
Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat
Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity

GEOTHERMAL RECORD OF CLIMATE CHANGE

Michael G. Davis[1], David S. Chapman[1], Robert N. Harris[2]
[1]Department of Geology and Geophysics, The University of Utah, Salt Lake City, UT, USA
[2]College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, OR, USA

Introduction

Because changes in temperature at the Earth's surface are transferred to the subsurface through the process of thermal diffusion, past changes in decadal and centennial surface temperature trends are now evident in subsurface temperatures. Thus, temperature-depth measurements in boreholes register not only the heat flowing out of the Earth but also transient departures from the background attributable to past surface temperature changes (e.g., Lachenbruch and Marshall, 1986; Pollack and Chapman, 1993; Pollack and Huang, 2000; Beltrami, 2002a; Harris and Chapman, 2005). Identifying and interpreting these transient departures in terms of the past few hundred years of surface temperature change is now known as the geothermal method of climate reconstruction. Because thermal diffusion behaves as a low-pass filter, such that high-frequency changes in surface temperature are attenuated more quickly and at shallower depths than longer-term changes, this method is most useful in identifying centennial scale climatic trends. The geothermal method of climate reconstruction has several benefits including (1) bypassing the empirical temperature calibration inherent with proxy records of climate change; (2) extending coverage to remote and widely dispersed continental sites where meteorological instrumental records are not available; and (3) establishing a surface baseline temperature prior to the industrial revolution when anthropogenic greenhouse gases started their unprecedented growth phase.

The geothermal method of temperature reconstruction

The geothermal method of temperature reconstruction is most easily understood in terms of the one-dimensional heat diffusion equation,

$$\frac{\partial T(z,t)}{\partial t} = \alpha \frac{\partial^2 T(z,t)}{\partial z^2}, \tag{1}$$

where T is temperature, z is depth, t is time, and α is thermal diffusivity. Time and depth are scaled nonlinearly through the thermal diffusivity such that a surface temperature change at time t prior to the present is largely captured within a depth, l, called the thermal length:

$$l = \sqrt{4\alpha t}. \tag{2}$$

Because the thermal diffusivity of rocks is about 1×10^{-6} m^2 s^{-1}, the majority of the past 100 years of surface temperature change is stored within the upper 150 m of the Earth; the majority of the last 1,000 years of surface temperature change is captured in the uppermost 500 m. Careful analysis of curvature in the upper few hundred meters of a temperature-depth profile, therefore, can be used to reconstruct surface temperature change over the past millennium (e.g., Huang et al., 2000; Harris and Chapman, 2001; Beltrami, 2002b; Pollack and Smerdon, 2004).

Useful analytical solutions to Equation 1 for a temperature-depth profile resulting from simple functional changes of temperature at the Earth's surface include the following (Carslaw and Jaeger, 1959; Lachenbruch and Marshall, 1986; Chisholm and Chapman, 1992):

(a) Single step ΔT change in surface temperature at time τ before present,

$$\Delta T(z) = \Delta T erfc\left(\frac{z}{\sqrt{4\alpha\tau}}\right), \tag{3}$$

(b) Multiple step ΔT_i changes in surface temperature at respective times τ_i before present,

$$\Delta T(z) = \sum_{i=1}^{n} \Delta T_i erfc\left(\frac{z}{\sqrt{4\alpha\tau_i}}\right), \tag{4}$$

and

(c) A single ramp change in surface temperature starting at time τ before present and with amplitude ΔT when the temperature log is obtained,

$$\Delta T(z) = 4\Delta T i^2 erfc\left(\frac{z}{\sqrt{4\alpha\tau}}\right), \tag{5}$$

where the complementary error function is denoted by *erfc* and $i^2 erfc$ indicates the second integral of the complementary error function.

Interpretations of surface temperature change from the geothermal method of climate reconstruction are most confident when curvature in the temperature-depth profile can be shown to be transient and fully attributable to a changing surface temperature in the past. Because the Earth is not homogeneous, phenomena and processes other than a changing surface temperature can also cause curvature in temperature-depth profiles. Steady-state source of curvature includes variations in subsurface thermal conductivity, radioactive heat production, heat refraction due to topography, and differential solar insolation due to variations in slope and azimuth. Transient sources of curvature include: changes in ground surface temperature (GST) in response to changes in surface air temperature (SAT), spatial changes in surface temperature around the borehole caused by temporal variations in albedo, non-isothermal groundwater flow, precipitation, and other microclimate effects (e.g., Beck, 1982; Chisholm and Chapman, 1992; Lewis and Wang, 1992; Harris and Chapman, 1995; Bartlett et al., 2004). Chisholm and Chapman (1992) and Harris and Chapman

(1995) explore the magnitude of many of these effects quantitatively. While most studies of climate change inferred from borehole temperature profiles attempt to select sites that minimize these non-climatic sources of borehole temperature profile curvature, it is often difficult to partition curvature between steady-state and transient sources of curvature. Repeated temperature measurements in the same borehole can help improve the confidence in separating and interpreting steady-state and transient sources of curvature (Davis et al., 2010).

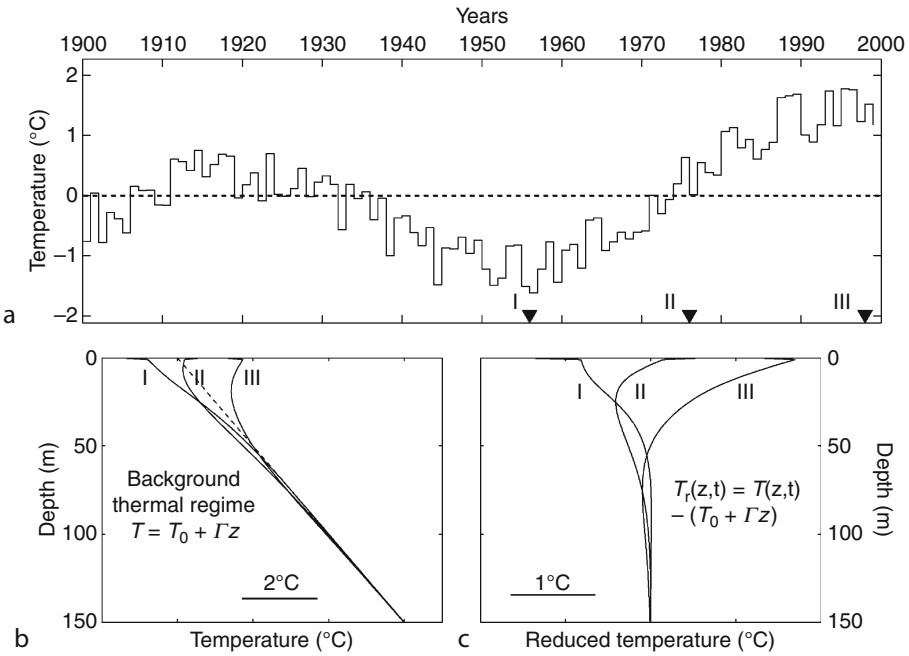
Tutorial

Important elements of the geothermal method of climate reconstruction are illustrated in Figure 1. First, consider a ground surface temperature (GST) that is essentially constant in time (dashed line; Figure 1a). In this case, the measured borehole profile has no transient curvature and the surface temperature intercept is equal to the GST (Figure 1b). However, if the GST is not constant, but fluctuating in some way such that there is a deviation from the mean GST, it will impart curvature to the subsurface according to Equation 1 (Figure 1b). For interpretations of climate change based on temperature profiles, one is interested in the departures from the background thermal regime. To highlight these departures, we compute reduced temperatures expressed as

$$T_r(z) = T_{obs}(z) - (T_o + \Gamma z), \tag{6}$$

where T_o is the reducing intercept and Γ is the reducing gradient. In Figure 1c, the background gradient has been removed, making the anomalous temperature strikingly apparent. The transient nature of the anomaly is illustrated by measuring the borehole temperatures at several times. The three temperature-depth profiles marked I, II, and III in Figures 1b and 1c were measured at three different times shown by the solid triangles in Figure 1a. Cooling from 1915 to 1955 creates a negative transient in the 1955 log I. Surface warming since 1955 results in warming features in the temperature-depth logs II and III that propagate deeper with time. This example also illustrates the low-pass smoothing of thermal diffusion. While trends in subsurface curvature can be related to trends in GST, individual annual means are not apparent. Further, the magnitude of GST change is much better resolved than the timing of the change, consistent with the physics of diffusion (Chisholm and Chapman, 1992).

Because SATs represent the gold standard of temperature change, and because climatic proxies such as tree rings are empirically calibrated to SATs, a final step is often taken whereby derived changes in GST are related to SATs. Many borehole temperature-depth profiles are located in areas where SAT records exist and the combination of datasets offer additional information about surface

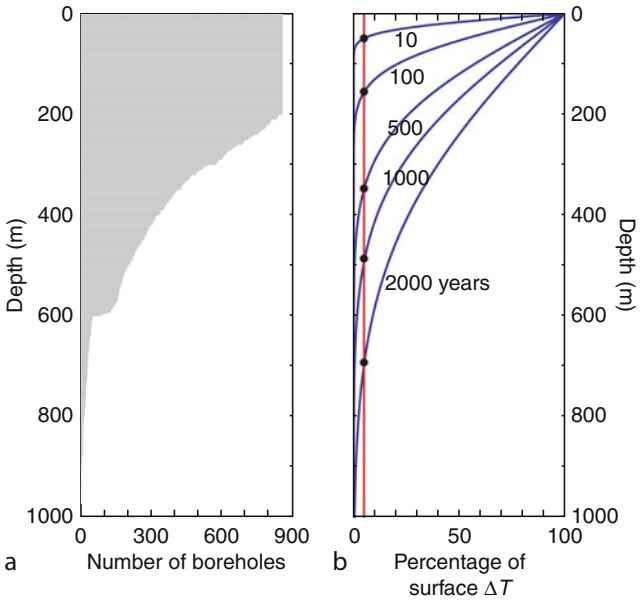


Geothermal Record of Climate Change, Figure 1 Basic aspects of the geothermal method of climate reconstruction. (**a**) An arbitrary time series of surface temperatures for the last century. Dashed line represents a climatic history with no temperature change. Triangles indicate times of borehole temperature logs illustrated in (**b**) and (**c**); log I in 1955, log II in 1975, and log III in 1998. The method asserts that changes in ground surface temperature (GST) are manifested as transient departures from the background, steady-state thermal regime. (**b**) Temperature-depth profiles for times I, II, and III. The linear profile results from the zero change scenario in (**a**) – dashed line. (**c**) Reduced temperature-depth profiles with background removed at times I, II, and III.

temperature change. Techniques to compare or use borehole temperatures with surface records include overlaying a GST reconstruction on an appropriate SAT curve (e.g., Huang et al., 2000) or using an appropriate SAT curve as a forcing function and comparing the resulting transient temperature-depth curve to a reduced temperature profile (e.g., Harris and Chapman, 2001). In this second technique, the best-fitting initial temperature, sometimes called the pre-observational mean (POM; Chisholm and Chapman, 1992; Harris, 2007), is found by minimizing the fit between the reduced temperatures and the synthetic transient calculated from the SAT time series. An obvious uncertainty in this step is the nature of coupling between air and ground temperatures. Regional comparisons at the centennial timescale generally show good agreement between changes in air and ground temperature (Huang et al., 2000; Harris and Chapman, 2001; Beltrami, 2002b; Pollack and Smerdon, 2004). Modeling studies using General Circulation Models allow comparisons between changes in air and ground temperatures at longer timescales and also suggest good agreement between these measures (Gonzalez-Rouco et al., 2003, 2006). Questions regarding changes in the relationship between air and ground temperatures have prompted detailed investigations often in combination with other meteorological parameters (e.g., Smerdon et al., 2003, 2004, 2006; Bartlett et al., 2006; Chudninova et al., 2006; Stieglitz and Smerdon, 2007). These studies have found that variations in air and ground temperatures generally track each other over the time period of study.

A global geothermal climate change database

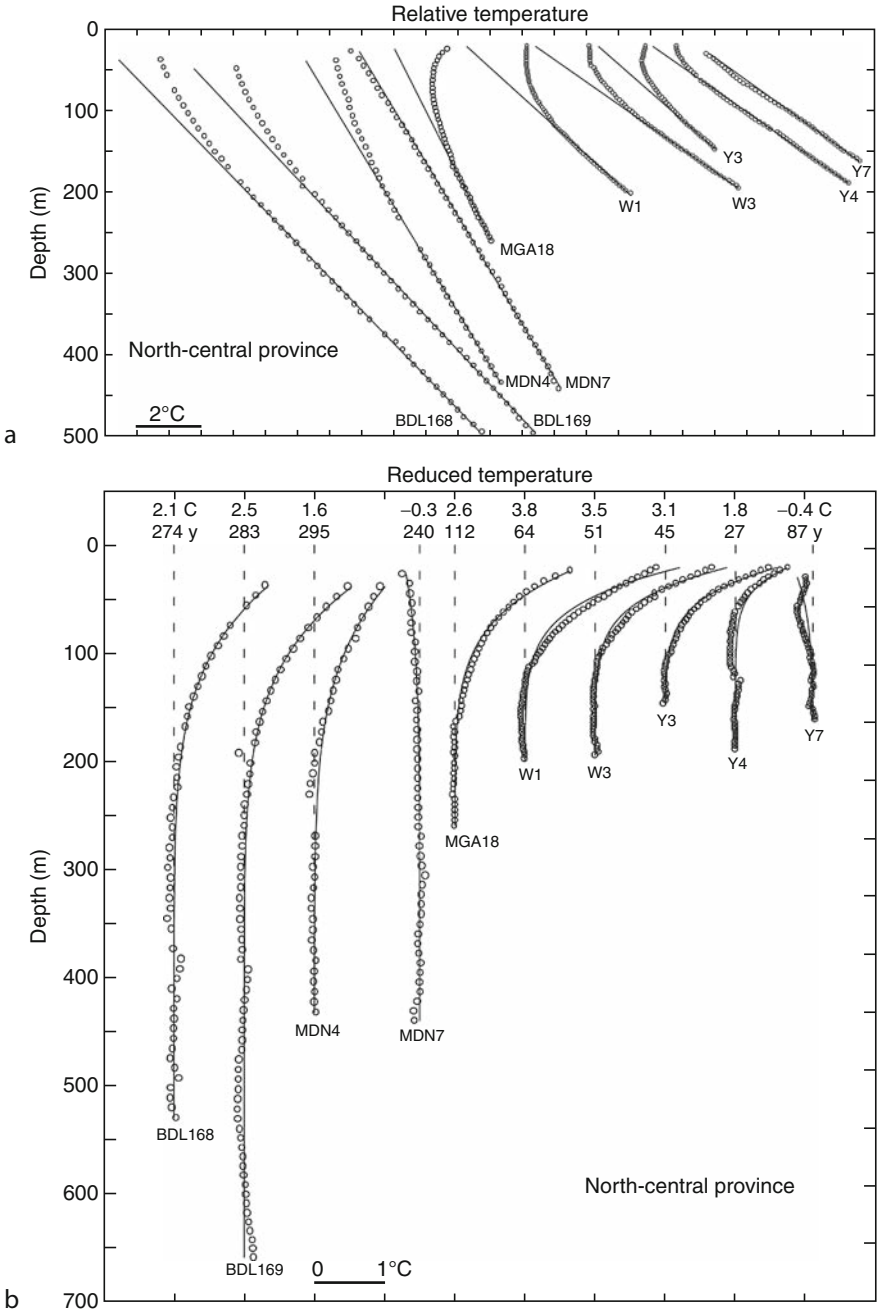
An archive of borehole temperature profiles at 862 sites (Huang et al., 2000) suitable for studying climate change has been compiled at the University of Michigan. Individually, the boreholes exhibit many characteristics of the meteorological record of climate change: hemispheric scale variation, regional variation, local effects, and noise. Collectively, however, borehole temperatures show clear evidence of significant twentieth century warming and a potential for inferring longer-term climate change (Huang et al., 2000). Figure 2a shows a histogram of the number of boreholes in the global climate borehole database with temperature observations to a given depth. Boreholes at all 862 sites extend to at least 200 m depth, one criteria for inclusion in the archive. Half of the boreholes are deeper than 350 m but fewer than 10% extend more than 600 m into the subsurface. The depth distribution of boreholes is important because the depth of borehole temperature anomalies is related to the timing of surface temperature changes (Equation 2), as shown in Figure 2b. The resolution of borehole temperature measurements can generally resolve a surface temperature signal that is greater than 5% of the surface temperature change. Figure 2b shows that to reconstruct past surface temperature adequately for the entirety of the last millennium, temperature-depth measurements must be collected over at least 500 m. In practice, these records must be deeper than 500 m so that the background thermal gradient and surface temperature intercept can be adequately estimated (Equation 6).



Geothermal Record of Climate Change, Figure 2 (**a**) The number of boreholes in the global climate borehole database with temperature observations to a given depth (*shaded*). (**b**) Temperature versus depth at times following a step change ΔT in surface temperature. Solid dots mark depths where the signal is 5% of the surface change.

Temperature-depth profiles

An ensemble of climatically perturbed temperature-depth profiles from India (Roy et al., 2002) is shown in Figure 3a. The boreholes are from the north-central climatic province of India. The first-order effect shown in these profiles is the general increase in temperature with depth, consistent with an upward flux of heat toward the Earth's surface. The deepest part of the borehole temperature profiles has a constant background thermal gradient. Figure 3b shows the reduced temperature profiles where the background gradient and surface temperature intercept have been removed (Equation 6), illustrating the transient temperature departures in the uppermost 200 m of the boreholes. The transient departure from a steady-state background signal is most consistent with a ramp change of amplitude −0.3–3.8°C starting 27–295 years ago. This example is notable because it shows the predominantly warming signal shown in ground temperatures consistent with surface temperature warming over the past two centuries, as well as the site-specific climatic conditions recorded in subsurface temperatures.



Geothermal Record of Climate Change, Figure 3 (**a**) Plots of temperature-depth data (*open circles*) obtained at 10 selected sites from the north-central climatic province of the Indian peninsula (Modified from Roy et al., 2002). The background temperature profiles (*solid lines*) are obtained by fitting a line just below the curved upper portion in each temperature-depth profile. Temperatures below 500 m are not displayed. (**b**) Reduced temperature profiles (*open circles*) for the 10 boreholes are shown in (**a**). Each profile is offset to avoid overlap and plotted relative to zero reduced temperature (*dashed line*). Solid lines show the best-fitting ramp models for the reduced temperature profiles. The magnitude and onset time of the ramp model are shown at the top of the profiles. See Roy et al. (2002) for more details.

Repeat temperature logs

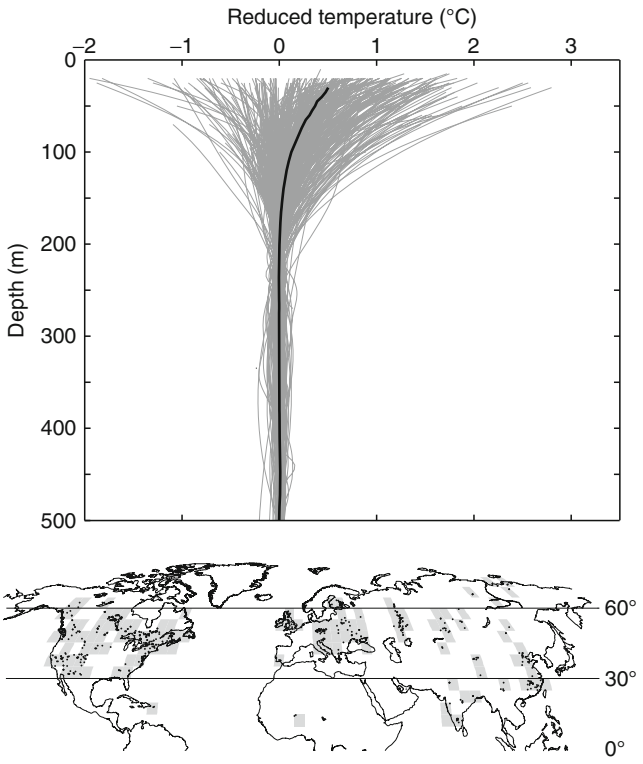
The regional analysis of Davis et al. (2010) employed multiple repeat temperature-depth profiles from three individual boreholes in northwestern Utah over nearly 30 years to provide observational constraints for understanding the relationship between ground and air temperatures. A key part of the study was to calculate and compare synthetic temperature profiles from nearby SAT records to the

observed temperature profiles. The synthetic profiles closely fit observed temperature-depth profiles and match both the amplitude and pattern of subsurface transient changes. Further, forward modeled ground temperatures and differences between temperature logs isolate transient variations and indicate that the observed profiles can be ascribed to changes in GST (Davis et al., 2010). Such observations of transient temperatures in boreholes offer strong support for using GST histories to complement SAT data and multi-proxy reconstructions in climate change studies.

Temperature reconstruction analysis

Several groups have used boreholes to analyze past temperatures on both regional (e.g., Lachenbruch and Marshall, 1986; Beltrami and Mareschal, 1991; Chisholm and Chapman, 1992; Harris and Chapman, 1995; Roy et al., 2002; Davis et al., 2010) and global (e.g., Huang et al., 2000; Harris and Chapman, 2001; Beltrami, 2002b) scales. The temperature reconstruction of Huang et al. (2000) utilized the global database and a Bayesian estimation technique, similar to the functional space inversion of Shen and Beck (1991). The method parameterized the reconstruction into century-long rates over the past 500 years, resulting in a 1.1°C increase over the entire period (Huang et al., 2000). Further analysis of this dataset partitioned the borehole into both hemisphere and continental reconstructions. Huang et al. (2000) found that warming was greatest in the Northern Hemisphere, and greatest in the twentieth century.

Another analysis used a Northern Hemisphere subset of the global database (Harris and Chapman, 2001) to obtain an average reduced temperature profile (Figure 4). The surface temperature record was then used as the forcing function, calculating the best-fitting synthetic profile. This process requires determination of a baseline or POM temperature prior to the beginning of the surface temperature record (Chisholm and Chapman, 1992; Harris, 2007). The result, which combined the POM and the instrumental surface temperature record, indicated that there was approximately 1°C of ground warming over the past 100–200 years (Harris and Chapman, 2001), consistent with proxy methods of paleoclimate reconstruction (e.g., National Research Council, 2006, and references therein) and comparable to other borehole analyses (Huang et al., 2000; Beltrami, 2002b).



Geothermal Record of Climate Change, Figure 4 (*Top*) Northern Hemisphere reduced temperature profiles from Harris and Chapman (2001) and Roy et al. (2002). Lines represent 509 profiles whose locations are shown on the map (*bottom*). Bold line shows average reduced temperature profile. Figure modified from Harris and Chapman (2001).

Summary

Temperature-depth profiles measured in boreholes can be an important record of surface temperature change and can be summarized as follows:

1. Thermal diffusion results in low-pass filtering of changes in surface temperature trends to be related to subsurface curvature. Unfortunately, annual means are not apparent. Further, the depth distribution of borehole temperature anomalies is related to the timing of surface temperature changes.
2. Both globally and regionally, ground temperatures show a predominantly warming signal consistent with surface temperature warming over the past two centuries.
3. Repeated temperature measurements in the same borehole can help improve the confidence in separating and interpreting steady-state and transient sources of curvature.
4. Geothermal climate change reconstruction techniques show approximately 1°C of ground warming over the past one to two centuries, consistent with proxy methods of paleoclimate reconstruction.

Bibliography

Bartlett, M. G., Chapman, D. S., and Harris, R. N., 2004. Snow and the ground temperature record of climate change. *Journal of Geophysical Research*, **109**, F04008, doi:10.1029/2004JF000224.

Bartlett, M. G., Chapman, D. S., and Harris, R. N., 2006. A decade of ground-air temperature tracking at emigrant pass observatory, Utah. *Journal of Climate*, **19**, 3,722–3,731.

Beck, A. E., 1982. Precision logging of temperature gradients and the extraction of past climate. *Tectonophysics*, **83**, 1–11.

Beltrami, H., 2002a. Earth's long-term memory. *Science*, **297**, 206–207.

Beltrami, H., 2002b. Climate from borehole data: Energy fluxes and temperatures since 1500. *Geophysical Research Letters*, **29**(23), 2111, doi:10.1029/2002GL015702.
Beltrami, H., and Mareschal, J. C., 1991. Recent warming in eastern Canada inferred from geothermal measurements. *Geophysical Research Letters*, **18**, 605–60.
Carslaw, H. S., and Jaeger, J. C., 1959. *Conduction of Heat in Solids*, 2nd edn. New-York: Oxford University Press.
Chisholm, T. J., and Chapman, D. S., 1992. Climate change inferred from analysis of borehole temperatures; an example from western Utah. *Journal of Geophysical Research*, **97**, 14155–14175.
Chudninova, S. M., Frauenfeld, O. W., Barry, R. G., Zhang, T., and Sorokovikov, V. A., 2006. Relationship between air and soil temperature trends and periodicities in the permafrost regions of Russia. *Journal of Geophysical Research*, **111**, F02008, doi:10.1029/2005JF000342.
Davis, M. G., Harris, R. N., and Chapman, D. S., 2010. Repeat temperature measurements in boreholes from northwestern Utah link ground and air temperature changes at the decadal time scale. *Journal of Geophysical Research,* **115**, pp.12, doi:10.1029/2009JB006875.
Gonzalez-Rouco, J. F., von Storch, H., and Zorita, E., 2003. Deep soil temperature as proxy for surface air-temperature in a coupled model simulation of the last thousand years. *Geophysical Research Letters*, **30**, 2116–2119.
Gonzalez-Rouco, J. F., Beltrami, H., Zorita, E., and von Storch, H., 2006. Simulation and inversion of borehole temperature profiles in surrogate climates: Spatial distribution and surface coupling. *Geophysical Research Letters*, **33**, L01703, doi:10.1029/2005GL024693.
Harris, R. N., 2007. Variations in air and ground temperature and the POM-SAT model: results from the northern hemisphere. *Climate of the Past*, **3**, 611–621.
Harris, R. N., and Chapman, D. S., 1995. Climate change on the Colorado Plateau of eastern Utah inferred from borehole temperatures. *Journal of Geophysical Research*, **100**, 6367–6381.
Harris, R. N., and Chapman, D. S., 2001. Mid-latitude (30°-60°N) climatic warming inferred by combining borehole temperatures with surface air temperatures. *Geophysical Research Letters*, **28**, 747–750.
Harris, R. N., and Chapman, D. S., 2005. Borehole temperatures and tree-rings: Seasonality and estimates of extratropical northern hemispheric warming. *Journal of Geophysical Research*, **110**, F04003, doi:10.1029/2005JF000303.
Huang, S. P., Pollack, H. N., and Shen, P. Y., 2000. Temperature trends over the past five centuries reconstructed from borehole temperatures. *Nature*, **403**, 756–758.
Lachenbruch, A. H., and Marshall, B. V., 1986. Climate change: Geothermal evidence from permafrost in the Alaskan Arctic. *Science*, **234**, 689–696.
Lewis, T. J., and Wang, K., 1992. Influence of terrain on bedrock temperatures. *Global Planetary Change*, **6**, 87–100.
National Research Council, 2006. *Surface Temperature Reconstructions for the Last 2,000 Years*. Washington: The National Academies Press, p. 145.
Pollack, H. N., and Chapman, D. S., 1993. Underground records of changing climate. *Scientific American*, **268**(6), 44–50.
Pollack, H. N., and Huang, S., 2000. Climate reconstruction from subsurface temperatures. *Annual Review of Earth and Planetary Sciences*, **28**, 339–365.
Pollack, H. N., and Smerdon, J. E., 2004. Borehole climate reconstructions: Spatial structure and hemispheric averages. *Journal of Geophysical Research*, **109**, D11106, doi:10.1029/2003JD004163.
Roy, S., Harris, R. N., Rao, R. U. M., and Chapman, D. S., 2002. Climate change in India inferred from geothermal observations. *Journal of Geophysical Research*, **107**(B7), 2138, doi:10.1029/2001JB000536.
Shen, P.-Y., and Beck, A. E., 1991. Least squares inversion of borehole temperature measurements in functional space. *Journal of Geophysical Research*, **96**, 19965–19979.
Smerdon, J. E., Pollack, H. N., Enz, J. W., and Lewis, M. J., 2003. Conduction-dominated heat transport of the annual temperature signal in soil. *Journal of Geophysical Research*, **108**(B9), 2431, doi:10.1029/2002JB002351.
Smerdon, J. E., Pollack, H. N., Cermak, V., Enz, J. W., Kresl, M., Safanda, J., and Wehmiller, J. F., 2004. Air-ground temperature coupling and subsurface propagation of annual temperature signals. *Journal of Geophysical Research*, **109**, D21107, doi:10.1029/2004JD005056.
Smerdon, J. E., Pollack, H. N., Cermak, V., Enz, J. W., Kresl, M., Safanda, J., and Wehmiller, J. F., 2006. Daily, seasonal and annual relationships between air and subsurface temperatures. *Journal of Geophysical Research*, **111**, D07101, doi:10.1029/2004JD005578.
Stieglitz, M., and Smerdon, J. E., 2007. Characterizing land-atmosphere coupling and the implications for subsurface thermodynamics. *Journal of Climate*, doi:10.1175/JCLI3982.1, 21–37.

Cross-references

Heat Flow Measurements, Continental
Lithosphere, Continental: Thermal Structure

GPS, DATA ACQUISITION AND ANALYSIS

Carine Bruyninx, Wim Aerts, Juliette Legrand
Department of Reference Systems and Planetology, Royal Observatory of Belgium, Brussels, Belgium

Definition

GPS, Global Positioning System, is a US satellite-based navigation system which can accurately determine position, velocity, and time in a common reference system, anywhere on or near the Earth on a continuous basis.

Satellite positioning

Point positioning

Satellite-based point positioning consists in determining the 3-dimensional position (X, Y, Z) of a receiver, using the radio signals emitted by a number of satellites with known position. The receiver, mostly located on the Earth, measures the travel time τ of the radio signal. Then it reconstructs the distance D between the receiver and the satellite by multiplying the travel time τ by its velocity. In the ideal case, neglecting the influence of the Earth's atmosphere, radio signals travel with the speed of light c (299,792,458 m/s) and the distance reads $D = \tau . c$.

Because the receiver and satellite clock are not perfectly synchronized, the measured travel time is corrupted with synchronization errors causing an error on the measured distance with respect to the true satellite-receiver distance. The unknown synchronization error dT between the satellite clock and the receiver clock becomes then the fourth unknown the receiver has to determine, next to the three-dimensional (X, Y, Z) receiver position.

Consequently, to solve this problem with four unknowns, at least four satellites have to be tracked simultaneously.

Relative positioning

Differential or relative satellite positioning is a technique developed to improve the accuracy of satellite point positioning by combining the measured travel time at different receivers. The basic principle consists in reducing the biases at a "remote receiver" using measured biases at a fixed "reference receiver" with known coordinates. This allows removing the errors common to the travel times measured at both stations. The position of the remote receiver (or the baseline vector) is then determined relative to the reference receiver. The corrections may be used in real time or later, using post-processing techniques. The remote receiver can be stationary or moving.

System description

Introduction

Development work on the NAVSTAR/GPS (Navigation by Timing And Ranging/Global Positioning System) started in 1973, when the US Department of Defense (DoD) applied its technical resources to develop a highly accurate, space-based navigation system to replace the TRANSIT system. For background to the development of the GPS system, the reader is referred to Parkinson (1994).

The first GPS satellite was launched on February 22, 1978, and GPS Full Operational Capability (FOC) was declared in July 1995 (US Coast Guard, 1995). FOC status means that the system meets all the requirements specified in a variety of formal performance and requirements documents. The system consists of three major segments: a space segment, satellites that transmit radio signals; a ground segment, ground-based equipment to monitor the satellites and update their signals; and a user segment, equipment which passively receives the satellite signals and converts satellites signals into position and navigation information.

Satellites, signals, and system time

The GPS is designed as a constellation of 24 satellites distributed among 6 orbital planes of which the inclination with respect to the equator is 55°. The orbital planes have a radius of 26,600 km. Each satellite makes two complete orbits each sidereal day. With this constellation, one can always see minimum four satellites from any point on the Earth. In the original GPS design, the satellite signal consists of

- The L1 and L2 carrier waves, with a respective wavelength of 19 and 24 cm
- Information, encoded in the form of binary bits on the carrier signals by a process known as phase modulation:
  - The Coarse/Acquisition code or C/A, which is freely available to the public, is a binary code sequence which repeats every 1 ms. It is transmitted on the L1 carrier at 1.023 Mbps (million bits per second) using a Binary-Phase Shift Keying (BPSK) modulation technique. The "chip length," or physical distance between binary transitions (between digits +1 and −1), is 293 m. Each satellite has a different C/A code, so that it can be uniquely identified.
  - The Precision code or P-code is transmitted on both the L1 and L2 frequencies at 10.23 Mbps with a chip length of 29.3 m and using the same BPSK modulation technique. The P-code repeats every 7 days, and like for the C/A code, each satellite has a unique P-code. Whereas C/A is a courser code appropriate for initially locking onto the signal, the P-code is designed for more precise positioning. The P-code is also not classified, but since January 31, 1994, a process called "Anti-Spoofing" encrypts the P-code in order to prevent an enemy to imitate the GPS signal. To that end, the P-code was modulated with the W-code, a special encryption sequence, to generate the Y-code.
  - The navigation message is modulated at 50 bit/s on the L1 wave on top of both ranging codes. It is sent in blocks of 1,500 bits (taking 30 s to transmit). One message block is subdivided in three parts or five subframes. The first part, or first subframe, contains GPS date and time, plus the satellite's status and an indication of its health. The second part, consisting of the second and third subframe, contains orbit parameters (often called the "broadcast ephemeris") and allows the receiver to calculate the position of the satellite in the earth-fixed WGS 84 system (ICD-GPS-200C, 2000). The last part, consisting of the fourth and fifth subframe, are partly for military use, but also contain information on the ionosphere (8 coefficients of the Klobuchar model, see Klobuchar (1986)), UTC data (Universal Time Coordinated), and the almanac data part. The almanac data contain information and status concerning all the satellites, their locations, and PRN numbers. The first and second parts are identical in all transmitted 30-s blocks. The third part differs as it contains only 1/25 of the entire almanac. Hence, to receive the almanac of the entire constellation, a total of 25 blocks must be received, which takes 12.5 min.

The GPS signal is transmitted with enough power to ensure a minimum signal power level of −160 dBW at the Earth's surface. The orientation of the satellites is always changing, such that the solar panels face the sun, and the antennas face the center of the Earth. More details on the GPS signals can be found in GPS Signal Specifications from the US Coast Guard (GPS-SPS, 1995). For a full description on the generation of the navigation message, we refer to the GPS Interface Specification (IS-GPS-200D, 2004).

New demands on the existing GPS system pushed for a gradual modernization. Within that frame, different

generations of GPS satellites have been developed, where the newer generations contain new satellite signals not present in the original design:

- Block I (no more operational): satellites launched between 1978 and 1985, with an average lifetime of 8–9 years.
- Block II and IIA: launched after 1985, these satellites are able to voluntary degrade the signals and have a typical lifetime of 7.3 years. They contain two cesium and two rubidium atomic clocks.
- Block IIR: launched after 1996, they have a lifetime of 10 years. They contain three rubidium atomic clocks.
- Block IIR-M: launched since 2005, these satellites transmit a second civil signal L2C on the L2 frequency and the military M signal on the L1 and L2 frequencies. The L2C signal aims at improving navigation accuracy, providing an easy to track signal, and acting as a redundant signal in case of localized interference. The satellites contain three rubidium atomic clocks.
- Block IIF: first launch scheduled for May 2010, the satellites will emit a new L5 civil signal providing a second safety-of-life signal that meets the technical requirements for enabling aircrafts to make precision landings in high multipath environments. Future launches of satellites with L5 capability will continue throughout the coming years until a full constellation of 24 satellites emitting L5 will be available by 2018. The satellites will carry four atomic clocks (cesium and rubidium).
- Block III: first launch scheduled for 2014, will carry a new L1C civil code which will be transmitted on L1 in addition to current signals. This seeks to maximize interoperability with Open Service signal of the European Galileo satellite navigation system. Further, Japan's Quazi-Zenith Satellite System (QZSS) will transmit a signal with virtually the same design as L1C. More background on the L1C signal can be found in Betz et al. (2007).

GPS system time is a paper clock based on the atomic clocks in the GPS ground control stations and the GPS satellites themselves. The GPS Control Segment maintains GPS system time to within 1 μs of UTC(USNO) (ICD-GPS-200C, 2000), which is the realization of the UTC kept by the USNO (United States Naval Observatory, Washington, USA). However, GPS system time is continuous and does not follow the UTC leap-seconds. UTC (USNO) itself is kept very close to the international benchmark UTC as maintained by the Bureau International des Poids et Mesures, in Paris.

GPS ground instrumentation

High-end GPS ground instrumentation generally consists of an antenna with hemispherical radiation pattern, connected to a receiver through a coaxial cable. GPS devices for the consumer market generally use small integrated antennas, resulting in a single portable device. For fixed GPS stations, portability is not an issue, but performance in terms of accuracy is. Consequently for this application, better, but larger, antennas and more complex, but more power consuming, receivers are used. Only the high-end instrumentation relevant for scientific use is discussed in this entry.

Antenna

Most GPS antennas in consumer electronics solely receive the primary L1 frequency. As a second frequency, however, allows for, e.g., correction of ionospheric delays which was already intended in the original system design (Spilker, 1978), antennas for high-precision positioning will receive on both L1 and L2. Typically, these high-precision antennas will be mounted on locations with clear sky view in order to be able to track as many satellites as possible. This implies that the cable connecting the antenna to the receiver, somewhere inside a building, will have a length of several meters. To compensate for the signal loss in this cable, most high-precision antennas have a built-in Low Noise Amplifier (LNA). In most cases, the LNA is fed by a dc voltage put on the antenna cable by the receiver.

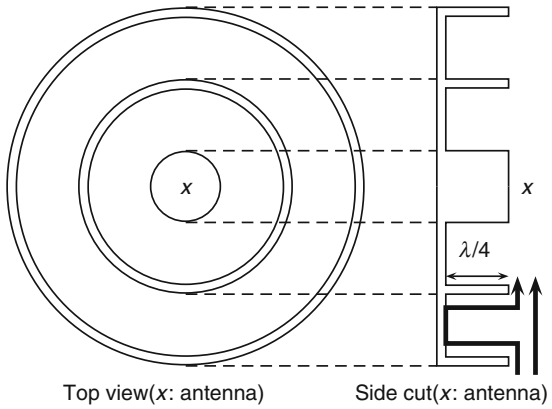
As the fields emitted by the satellites are Right Hand Circularly Polarized (RHCP) (IS-GPS-200D, 2004), any antenna capable of receiving this polarization is able to receive the GPS signals. This can be wire antennas, such as helix antennas or crossed dipoles (e.g., the Dorne Margolin antennas included by many manufacturers), or circular patches.

In many cases, choke rings are added to the antenna in order to suppress the reflections (also known as multipath). These choke rings are cavities with a depth of approximately a quarter of a wavelength. As a consequence, multipath waves, generally coming in from the back side of the antenna, will partly reflect in the cavities. These reflected waves will cancel out with the multipath waves which did not enter the cavity, thanks to a difference in path length of half a wavelength. This working principle is illustrated in Figure 1. The concept of choke rings was first introduced by Jet Propulsion Laboratory (Tranquilla et al., 1989).

As the choke ring is typically a narrow band design, its performance can never be perfect on both L1 and L2 frequencies. Therefore, many design improvements such as dual depth choke rings (Zhodziskhky et al., 1998) were proposed. Another problem linked with the antenna being multiband is the considerable variation of the phase center, which is the physical point where the signal is received. Fortunately, this can be measured and compensated for through an appropriate calibration procedure (Cohen and Parkinson, 1991).

Receiver

A generic GPS receiver consists of a Radio Frequency (RF) part that demodulates the messages transmitted by the GPS satellites, and a positioning part that, based



Gps, Data Acquisition and Analysis, Figure 1 Working principle of the choke ring ground plane.

on the information received from the RF part, determines the position of the user.

Most GPS receivers allow storage of some variables (such as position and speed), and status information (such as system errors), and observables (such as code and phase measurements) in log files for further processing, mostly on removable memory cards. In combination with battery feeding, this should allow for autonomous operation for several days.

The RF block in the receiver demodulates the radio signals of all satellites in view. These signals are received by a single antenna, meaning that the receiver essentially gets a signal from the antenna that is the sum of all satellite signals. The receiver will hence use Code Division Multiplexing (CDM) techniques to extract the messages from the different satellites out of the single received signal (Braasch and Van Dierendonck, 1999). To do so, the receiver duplicates the antenna signal into several identical copies. Each of the identical copies will be sent into a different channel of the receiver. In every single channel, a signature from a different GPS satellite will be sought for, and the navigation message and the propagation delay of this satellite will be extracted. For cost reasons, this duplication is performed after down conversion of the original signal to a lower Intermediate Frequency (IF) and digitization with an Analog to Digital Converter (ADC). Indeed, a digital signal can easily be copied to feed the distinct channels. Contemporary receivers have about 100 such channels.

Using a trial and error procedure by varying Doppler shift, carrier phase, and PRN time delays, the following steps are performed in each of the channels. The first step consists in removing the carrier by a numerical (digital) mixing of the IF signal with an internally generated carrier. The remaining signal only contains the PRN code and data bits. This signal is correlated with the satellite-specific PRN code generated by the receiver for varying time delays. In Figure 2 the correlation with the received signal as a function of the time shift of the internally generated PRN is shown. Once the correlation of the internally generated replica of the PRN and the received signal succeeds, the Doppler shift, carrier phase, and code time delay of a certain satellite signal is determined. The channel is then said to be locked on to that specific satellite. Else, another PRN is tried.

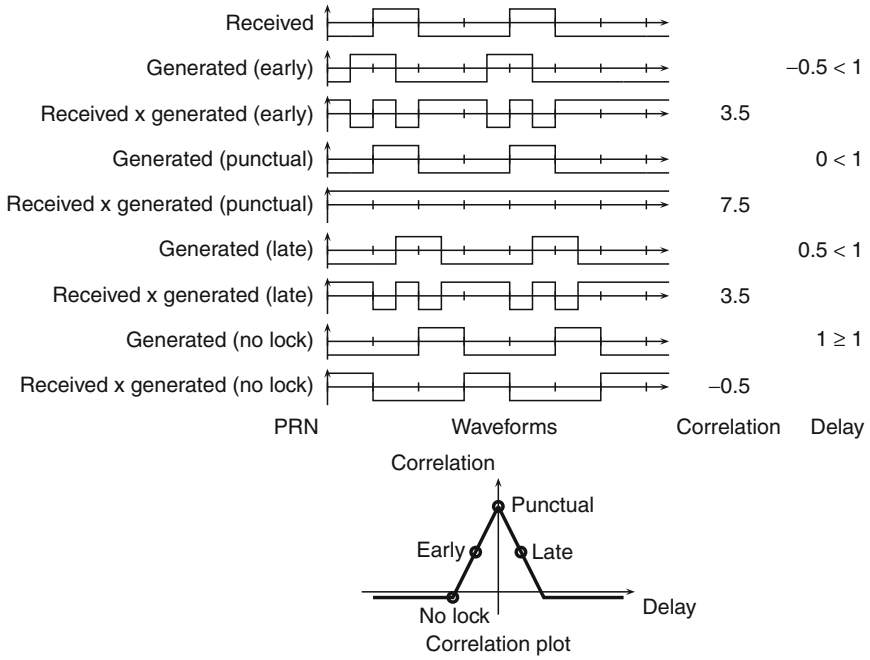
Obviously, the more stable the receiver internal clock (mostly a quartz crystal) is, the more accurately the Doppler shift, carrier phase, and code delay can be determined. However, due to systematic errors, the timescale of the receiver's clock deviates from the GPS timescale. If the difference between both approaches 1 ms, the receiver's time is resynchronized to the GPS timescale, resulting in a time jump, but assuring time correspondence of measurements from different receivers. In addition, signal noise and noise in the channel control loops will cause minor errors in the determination of these variables. These errors, referred to as measurement noise will limit the precision of the position determination.

Only the PRNs of the C/A code, sent on L1, are publicly known. The encrypted Y-code's PRNs cannot be generated by the civil GPS receivers. Consequently, the civil receivers cannot receive the messages that were encoded with the military CDM. However, as the Y-code is sent on both L1 and L2, techniques of cross-correlating the signals on both frequencies are used to still allow determination of difference in arrival time of the P-code ay both frequencies (Woo, 1999).

If the channel is locked to a satellite, the correlation between the received satellite and the internally generated PRN code sequence will be strong. Then, in order to decode the navigation message, the sign of this correlation is investigated. If the correlation is strongly positive, a 1 was sent by the satellite; if it is negative, a 0 was sent. This way, the bits sent by the satellite can be demodulated. The demodulated navigation message, together with the Doppler shift, carrier phase, and pseudo range of the channel are then passed to the positioning block that will combine all information from all channels to determine the receiver's position. Besides using the information from the RF block, the positioning block of most contemporary receivers can also take into account signals from correction streams from neighboring GPS receivers (see Sect. Positioning Modes) or inertia sensors.

Measurements

The basic principle of the GPS measurement is that the GPS receiver determines for each satellite the time shift between the signal emission time and the signal reception time (see the "Receiver" part for more details). By multiplying this time difference with the signal travel speed, the distance between the GPS receiver and emitting antenna of each locked satellite can be reconstructed. As will be shown later, this measured distance does not correspond with the geometric distance separating the receiver and satellite, and is therefore called "pseudo range." As already outlined in the Sect. Receiver, two main types of measurements can be done by the RF block of the GPS receiver:



Gps, Data Acquisition and Analysis, Figure 2 Correlation with the received PRN as a function of time shift of the internally generated PRN.

- Range observations, based on the time shift of the internally generated PRN codes, also called "pseudo-range" measurements.
- Carrier phase observations, based on the phase shift in the numerical IF mixing step in each receiving channel. They are more precise than the pseudo-range type measurements, but have a much higher degree of "ambiguity" than the code ranges.

Code pseudo-range model

The GPS code pseudo-range P_i^k measurement at receiver i to satellite k can be expressed according to the following simplified equation (Wells et al., 1987):

$$\rho_i^k = \rho_i^k + c.\left(dt_i - dt_i^k\right) + I_i^k + T_l^k + \varepsilon_{Pi}^k \tag{1}$$

where ρ_i^k is the geometric range from receiver i to satellite k, dt_i and dt_i^k are, respectively, the receiver and satellite clock offsets from GPS system time, while I_i^k and T_i^k take respectively into account the effect of the ionosphere and troposphere on the signal. Finally, ε_{Pi}^k stands for the measurement noise components, including multipath.

Using Pythagoras Theorem, we know that

$$\rho_i^k = \sqrt{\left(x^k - x_i\right)^2 + \left(y^k - y_i\right)^2 + \left(z^k - z_i\right)^2} \tag{2}$$

with $\overrightarrow{x_i} = (x_i, y_i, z_i)$ and $\overrightarrow{x^k} = \left(x^k, y^k, z^k\right)$ resp. the receiver and satellite positions (at signal transmission time).

For positioning or navigation purposes, the user wishes to determine the receiver position $\overrightarrow{x_i}$. As Equation 2 is a nonlinear equation, it is linearized around the a priori receiver position $\overrightarrow{x_{i0}} = (x_{i0}, y_{i0}, z_{i0})$:

$$\begin{aligned} P_i^k(t_i) = {} & \frac{\left(x_{i0} - x^k\right)}{\rho_{i0}^k} dx_i + \frac{\left(y_{i0} - y^k\right)}{\rho_{i0}^k} dy_i \\ & + \frac{\left(z_{i0} - x^k\right)}{\rho_{i0}^k} dz_i + c.\left(dt_i - dt_i^k\right) \\ & + I_i^k + T_i^k + \varepsilon_{\mathrm{Pi}}^k \end{aligned} \tag{3}$$

where (dx_i, dy_i, dz_i) denotes the deviation of the unknown receiver position with respect to its a priori value and ρ_{i0}^k denotes the geometric range from receiver i to satellite k computed using the a priori receiver position.

For most applications, the satellite position $\overrightarrow{x^k}$ and clock synchronization error dt_i^k can be considered as known. This leaves the following unknowns (dx_i, dy_i, dz_i, dt_i) in Equation 3. As previously mentioned, a minimum of four GPS satellites has to be observed simultaneously to solve for these four unknowns.

Carrier phase model

The GPS carrier phase φ_i^k measured at receiver i to satellite k can be modelled according the following simplified equation (expressed in [m]):

$$\phi_i^k = \rho_i^k + c.\left(dt_i - dt_i^k\right) + \lambda N_i^k - I_i^k + T_i^k + \varepsilon_{\phi i}^k \tag{4}$$

This equation differs from the pseudo-range (Equation 1) as follows:

N_i^k, the initial cycle ambiguity: an integer associated with the first measurement (fractional phase) of a satellite track (or after a tracking interruption). After the initial measurement (and as long as there is no tracking interruption), the receiver keeps track of the cumulative phase differences reflecting the changes in distance to the satellite.

λ, the wavelength of the carrier wave (L1, L2 or any of their linear combinations, see Dach et al., 2007).

$\varepsilon_{\phi i}^k$, the measurement noise of the phase observable, including phase multipath.

But this equation can be linearized in exactly the same way as the code pseudo-range equation (Equation 1).

Remarks

The following comments can be made with regard to Equations 1 and 4:

- These equations are generic and several of the terms in the equations differ depending on the observed code pseudo-range (C/A, P1, or P2) or carrier phase (L1 or L2) signal.
- High-end GPS receivers are programmed to measure GPS satellites simultaneously at the epoch $\mathbf{t_i}$ (e.g., at the full second). Even if the reception times are identical, respective signals leave the satellites at slightly different times because the satellite-receiver distances differ. The signal transmission time t_i^k corresponding to the reception time becomes thus receiver-dependent (hence the use of the double index).
- The receiver clock error is identical in all simultaneous measurements at receiver i from all satellites.
- Using simultaneous code pseudo-range measurements to at least four satellites, the absolute position of the receiver can be computed rather straightforward. This is not possible when using carrier phases as the initial cycle ambiguity is an additional satellite-receiver-dependent unknown which has to be dealt with.
- Note the different sign for the ionospheric effect in Equations 1 and 4. The ionosphere, being a dispersive medium, acts differently on the group velocity (code) than the phase velocity and is consequently delaying the code measurements while it advances the phases (Leick, 2004).

Single and double differences

For relative positioning between two stations (i, j), the observations made simultaneously to the same satellite k are differenced:

$$\begin{aligned}\Delta P_{ij}^k &\cong P_i^k - P_j^k \\ &= \Delta\rho_{ij}^k + c.\Delta dt_{ij} + \Delta I_{ij}^k + \Delta T_{ij}^k + \Delta\varepsilon_{\mathrm{p}ij}^k\end{aligned} \tag{5}$$

$$\begin{aligned}\Delta\varphi_{ij}^k &\cong \varphi_i^k - \varphi_j^k \\ &= \Delta\rho_{ij}^k + c.\Delta dt_{ij} - \Delta I_{ij}^k + \Delta T_{ij}^k + \lambda\Delta N_{ij}^k + \Delta\varepsilon_{\phi ij}^k\end{aligned} \tag{6}$$

$\Delta(.)_{ij}^k$ denotes the single difference. Note that the satellite clock difference errors are eliminated from Equations 5 and 6. If in addition, the difference between the single differences to two satellites k and l is taken, then the so-called double differences are constructed (indicated with $\Delta(.)_{ij}^{kl}$):

$$\Delta P_{ij}^{kl} \cong \Delta P_{ij}^k - \Delta P_{ij}^l = \Delta\rho_{ij}^{kl} + \Delta I_{ij}^{kl} + \Delta T_{ij}^{kl} + \Delta\varepsilon_{\mathrm{p}ij}^{kl} \tag{7}$$

$$\begin{aligned}\Delta\varphi_{ij}^k &\cong \Delta\varphi_{ij}^k - \Delta\varphi_{ij}^l \\ &= \Delta\rho_{ij}^{kl} - \Delta I_{ij}^{kl} + \Delta T_{ij}^{kl} + \lambda\Delta N_{ij}^{kl} + \Delta\varepsilon_{\varphi ij}^{kl}\end{aligned} \tag{8}$$

where the receiver clock difference errors, which are the same for both single differences, are also eliminated.

Although the use of single and double differences has many advantages (see Sect. Errors and Biases), the differencing process increases the measurement noise level by $\sqrt{2}$ for each difference and introduces correlations between the differenced equations. These correlations can, however, be taken into account using the procedure developed in Beutler et al. (1986).

The double difference integer ambiguity ΔN_{ij}^{kl} in Equation 8 plays an important role in accurate positioning, using carrier phase double differences. When the ambiguity is estimated together with the other parameters as a real number, a so-called "float solution" is obtained. If the estimated ambiguities can, however, be successfully constrained to their integer value, then an "ambiguity-fixed solution" is obtained. This process of estimating ambiguity parameters, and then selecting the likeliest integer values, is known as "ambiguity resolution." Once the L1 and L2 ambiguities are resolved, and can thus be removed from the Equation 8, the double differenced phase observations become precise pseudo-range observations.

Errors and biases

Introduction

GPS measurements are corrupted by biases (measurement errors that cause true ranges to be different from measured ranges by a "systematic amount") and errors (internal instrument noise or residual biases) which limit the positioning accuracy. Three groups of GPS errors and biases are generally distinguished: receiver-dependent, satellite-dependent, and signal propagation errors.

Receiver-dependent errors

Measurement noise

The measurement noise is included in the terms $\varepsilon_{\mathrm{P}i}^k$ and $\varepsilon_{\phi i}^k$ within the code pseudo-range (Equation 1) and carrier phase

(Equation 4) measurements. Without biases contaminating the measurements, low measurement noise is expected to result in high accuracy positioning. A "rule-of-thumb" (Wells et al., 1987) states that the measurement noise is at the level of 1%, or better of the wavelength (or chip length) of the signal. In the case of code pseudo ranges, this results in a measurement noise of 3 m for the C/A code and 0.3 m for the P-code. With their shorter wavelength, carrier phase measurements are less noisy implying millimeter noise for carrier phase measurements.

Receiver clock errors

GPS receivers are equipped with relatively inexpensive quartz crystal oscillators. The receiver clock error dt_i contaminates all satellite-receiver ranges. As mentioned in "The code pseudo-range model", the solution to this problem is to treat the clock error as an additional parameter in the code pseudo-range estimation procedure. This allows estimating the receiver clock error with a precision of 10^{-8} to 10^{-7} s which is in addition satisfactory to access the exact measurement epoch of both code and phase measurements. An alternative strategy uses single or double differences see (Equations 5–8) to eliminate the receiver clock errors dt_i.

Antenna-specific errors

The precise point whose position is measured by a GPS receiver is the electrical phase center of the GPS receiver's antenna which is not the physical center of the antenna. The phase center will change with the changing direction of the signal from a satellite. The a priori receiver position $\overrightarrow{x_{i0}}$ has to be corrected for this effect. This behavior is different for each antenna, and modelling it requires precise antenna calibrations (Mader, 2001; Görres et al., 2006; Schmid et al., 2007). GPS antenna calibrations consist of two parts: (1) an average phase center offset with respect to a physical feature of the antenna and (2) the phase center variation (PCV) with respect to the satellite elevation angle and azimuth. Ignoring these phase center variations can lead to serious (up to 10 cm) vertical positioning errors. Even if the antenna calibrations have submillimeter accuracy, the access to absolute high-precision positioning is still limited by site-specific near-field multipath errors (Elósegui et al., 1995) which can introduce systematic biases up to 1–2 cm, mostly in the estimated height component.

Multipath effects, included in terms $\varepsilon^k_{\mathrm{Pi}}$ and $\varepsilon^k_{\varphi i}$ in respectively Equations 1 and 4, are propagation errors arising from the interference of the direct signal by reflected signals from water or metallic surfaces or nearby buildings. The combined direct and reflected signals will give rise to incorrect pseudo-range or phase measurements. Low-elevation signals are more susceptible to multipath effects than those at high elevations. The size of pseudo-range multipath can reach 10–20 m on the C/A code (Evans, 1986), while for the carrier phases, it can have a maximum value of a quarter of a cycle (4.8 cm for L1) (El-Rabbany, 2002).

Using specific antenna design (see Sect. Ground Instrumentation), careful antenna mounting, and advanced signal processing to filter out multipath integrated in the receiver, can reduce the multipath errors. Using differenced code or phase observables will not remove the effect of multipath as the effect is receiver and satellite-dependent. However, when using stationary receivers, an appropriate multipath mitigation can be implemented by taking advantage of the multipath time correlation due to the repeating satellite ground tracks (Wu and Hsieh, 2010).

Cycle slips

If the receiver loses lock on a satellite signal, due to, e.g., obstructions or interference, then the carrier phase observables might suffer from cycle slips. These cycle slips are jumps in the carrier phase measurements caused by the fact that when signal lock is reacquired, the integer initial cycle ambiguity N^k_i of the phase observable will be re-initialized. Cycle slips can occur independently on L1 and L2.

The detection and repair of cycle slips (Kim and Langley, 2002) is an important carrier phase data preprocessing step. Each time the software will detect a cycle slip, the cycle slip will be repaired (to restore the continuity of carrier cycle count) or a new integer initial ambiguity N^k_i will be setup (Jonkman and De Jong, 2000 and Dach et al., 2007). If this is properly done, the cycle slips are not contributing to the error budget anymore.

Satellite-dependent biases

Orbit and satellite clock errors

Computing the a priori geometric range, ρ^k_{oi} (Equation 3) implies known satellite position. The navigation message broadcast by the GPS satellites provides the user in real-time information on the satellite position and clock error. Following the International GNSS Service (http://igscb.jpl.nasa.gov/), the accuracy of the broadcast orbits is at the 1-m level while the satellite clock error is given at the 5 ns level. The effect of the satellite orbit and clock errors (dt^k_i in Equations 4 and 6) propagates directly into the point position results and amplifies the positioning error.

In the case of relative positioning, the orbit errors dr (in m) propagate into errors db (in m) in the coordinates of a baseline of length b following the rule of thumb of Bauersima (1983): $\frac{|db|}{b} = \frac{|dr|}{r}$ where r is the receiver–satellite distance (in km), typically 20,000 km. So, for a receiver inter-distance (or baseline length) of 100 km and an orbit error of 1 m, the error on the baseline components will be of the order of 0.5 cm. Following Beutler (1996), the Bauersima rule is a bit too pessimistic and actually gives mainly an approximation for the propagation of the orbit errors on the estimated height component. For the horizontal component, the error can be divided by a factor of three.

Precise GPS positioning (see *GPS, Tectonic Geodesy*) requires orbits with a higher precision than the ones provided in the broadcast navigation message. For that purpose, the International GNSS Service (IGS, see *Geodesy, Networks and Reference Systems*) uses the tracking data

from a worldwide GPS network to generate the most precise satellite orbits and clocks related to the International Terrestrial Reference system (ITRS). Their accuracy depends on their latency (Dow et al., 2009) but typically varies between 2.5 and 5 cm for the orbits and ±75 to 150 ps for the satellite clocks.

Antenna-specific errors

Each GPS satellite has a transmitting antenna which, like the receiver antenna, also has a phase center which changes with elevation angle and is influenced by the local environment around each satellite. The scientific community, mainly driven by the International GNSS Service, has made considerable efforts to accurately determine the satellites antenna phase center offsets and variations (Schmid et al., 2007). The correct location of this phase center with respect to the satellite's center of mass is critical for accurate orbit determination.

Signal propagation biases

When travelling from the GPS satellites to the receiver on Earth, the radio signals cross and interact with the Earth's atmosphere. The two atmospheric layers that influence the most the propagation of the GPS signals are the troposphere and the ionosphere. The atmospheric refraction has consequently an ionospheric I_i^k and tropospheric T_i^k component, each with their own spatial and temporal characteristics.

Ionospheric refraction

The ionosphere is stretching from a height of about 50 km to more than 1,000 km over the surface of the Earth; it is named so because it is ionized by the Sun's ultraviolet light. When the GPS signals propagate through this medium dispersion occurs, bending the signal and changing its velocity (Klobuchar, 1991). The ionosphere speeds up the propagation of the carrier phase and slows down the PRN code (as can also be seen in Equations 1 and 4). The signal bending can be neglected for satellite elevations larger than 5°.

The ionospheric delay I_i^k is proportional to the number of free electrons along the GPS signal path, called the total electron content (TEC) which varies as a function of the latitude of the receiver, the season, the time of day, and solar activity (approximately 11-year cycle). At first order, the ionospheric delay on the carrier phases reads: $I_i^k = -\frac{1}{\cos z}\frac{40.3}{f^2}TEC$, with z the satellite's zenith distance. It ranges from about 50 m for signals at the zenith to as much as 150 m for low-elevation observations.

To reduce the ionospheric effect, a correction (using the Klobuchar model, Klobuchar, (1986)) is transmitted within the navigation message. However, it is generally conceded that the broadcast correction model corrects only 50–60% of the ionospheric delays, which is considered as the performance level of the model. For single frequency receivers, the use of the correction model parameters is often the only option for point positioning. Another source of information on the behavior of the ionosphere is provided by the IGS from their Global Ionospheric Maps (GIM) which consist of a global TEC grid (2,5° lon × 2.5° lat) available at 2-hourly intervals. The IGS GIM maps have a typical accuracy of 2–8 TEC units (Dow et al., 2009).

In relative positioning, it has been shown by Georgiadou and Kleusberg (1988) that the ionospheric error will induce a baseline shortening dependent of the baseline length. Following Santerre (1991), the estimated receiver positions will suffer from scale and orientation biases. For example, at a midlatitude site using an elevation cutoff angle of 20°, a horizontal scale bias of −0.63 ppm is incurred for each 1×10^{-7} m^{-2} of TEC not accounted for.

As the ionosphere acts as a dispersive medium, the ionospheric delay is frequency-dependent. Hence, if dual-frequency receivers are available a special combination of the L1 and L2 signals, called the ionosphere-free combination L3, can eliminate the first order of the ionospheric effect (Beutler et al., 1988a). The ionosphere-free carrier phase measurement is:

$$\phi_{L3} = \frac{f_1^2}{f_1^2 - f_2^2}\phi_{L1} - \frac{f_2^2}{f_1^2 - f_2^2}\phi_{L2} \tag{9}$$

with f_1 and f_2, respectively, the frequencies of the L1 and L2 signals, and ϕ_{L1} and ϕ_{L2} the carrier phase observables (converted in m) on L1 and L2. A similar equation can be derived to construct the ionosphere-free code measurements. Using the ionosphere-free combination removes 99% of the ionospheric error. For applications requiring the highest positioning precision, the usage of ionosphere-free differenced observables is the most effective way of mitigating the ionospheric effect.

Tropospheric refraction

The troposphere is the lower part of the Earth's atmosphere, and it is the seat of all meteorological phenomenas (clouds, rain, hydrometeors...). It contains approximately 75% of the atmosphere's mass and and 99% of its water vapor and aerosols. The troposphere extends from the surface of the Earth to less than 9 km over the poles to 16 km over the equator (Lutgens and Tarbuck, 1989). For radio frequencies below 15 GHz, it is a nondispersive medium meaning that it delays all GPS codes and carriers by the same amount so that the effect cannot be removed by combining observations made on two frequencies.

The tropospheric refraction delay T_i^k depends on atmospheric pressure, temperature, and water vapor pressure. In addition, it is the shortest in the zenith direction (2 m) and increases when going down to the horizon (about 20–28 m for signals at an 5° elevation angle) as the air mass traversed by the signal increases (Brunner and Welsch, 1993).

The tropospheric delay consists of a dry and wet component. The dry component, which depends on surface pressure and represents about 90% of the delay, can be

easily modelled. The most commonly used model is the Saastamoinen model (1972), which is also recommended by the IERS conventions (2003). Today, the necessary pressure values input to the Saastamoinen model can be derived from numerical weather model data from the European Centre for Medium-range Weather Forecasts (ECMWF), see Boehm et al. (2007). The wet component of the tropospheric delay accounts for the remaining 10% of the delay and depends on the amount of water vapor along the signal path; it is much more difficult to accurately model.

Two kinds of troposphere biases can affect relative positioning (Beutler et al., 1988b): relative and absolute tropospheric biases. Relative troposphere biases are caused by errors of (mismodelled) tropospheric refraction at one endpoint of a baseline relative to the other endpoint, while absolute ones are caused by errors common to both endpoints of a baseline. Relative troposphere biases primarily bias station heights. For example, a relative troposphere bias of only 1 cm leads to an error of approximately 3 cm in the estimated relative station height for an elevation cutoff angle of 20°. This error increases to 19 cm for an elevation cutoff angle of 3°. Absolute troposphere biases produce scale biases of the estimated baseline lengths. For example, an absolute troposphere bias of 10 cm induces a scale bias of 0.05 ppm for an elevation cutoff angle of 20° and of 0.3 ppm for a cutoff angle of 3°.

For high-precision applications where mm accuracy is desired, the tropospheric bias at zenith is parameterized in the final position solution (see IERS Conventions (2003)). For that purpose, different mapping functions (Niell, 2000; Boehm and Schuh, 2004; Boehm et al., 2006) are used to express the dependency between the tropospheric delay at zenith and the one experienced at satellite's zenith distance.

In addition, as with the ionospheric bias, the tropospheric bias is spatially correlated over distances up to several tens of kilometers meaning that differential positioning is an effective strategy for mitigating the effect of the tropospheric bias on positioning results.

Summary

As we have seen, all biases exceeding the measurement noise level have a high probability to degrade the accuracy of the positioning and should therefore be accounted for somehow. Various strategies have been developed to account for them:

- They can be estimated as explicit (additional) parameters, e.g., the receiver clock error and the tropospheric refraction.
- Common-mode biases can be eliminated or mitigated through combination of observables or datasets, e.g., by using of the ionosphere-free observable or relative positioning to mitigate the ionospheric and tropospheric refraction as well as satellite clock and orbit errors.
- The biases can be directly measured, e.g., by using antenna calibrations.
- The biases can be considered known or adequately modelled, e.g., by using precise satellite orbits and clocks from the IGS.
- They can be ignored, e.g., the phase multipath error at moving receivers.

Positioning modes

Two main methods of positioning with GPS exist: point positioning and relative positioning. In addition, the precision of the obtained positions depends on the observables used: code and/or carrier phase data, the interval over which the data are accumulated before being analyzed, and the sophistication of the error modelling.

Point positioning

GPS point positioning employs the data from a single GPS receiver.

Using code pseudo-ranges from at least four satellites, Equation 1 can be used to compute the instantaneous receiver position and receiver clock error. With the broadcast ephemeris and ionospheric correction, a typical position accuracy of 22 m (2D rms) can be achieved using C/A codes. Typically, code pseudo-range positioning is used for low-accuracy real-time land, sea, and air navigation. The distinguishing characteristic of navigation is the *urgency* with which positioning information is required.

A more sophisticated method of performing point positioning uses the ionosphere-free combination of dual-frequency code and carrier phase observables. It is known as Precise Point Positioning (PPP) and allows reaching centimeter-level precision when using long observation series. Kouba and Héroux (2001) show that to reach this precision, it is necessary to use the most precise satellite orbit and clock information (from the IGS) as well as a sophisticated error modelling. For that reason, PPP is at the moment mostly implemented in scientific GPS analysis software such as Bernese (Dach et al., 2007), and GISPY/OASIS (Zumberge et al., 1997). As PPP eliminates the need to acquire simultaneous tracking data from a reference station, it has given rise to centralized geodetic positioning services that require from the user a simple submission of a valid GPS observation file (see, e.g., Ghoddousi-Fard and Dare, 2005). As shown by Bisnath and Gao (2008), PPP has the potential to become a positioning technique that could replace the RTK positioning mode (see next section).

Relative positioning

Relative positioning employs two GPS receivers, a remote and a reference receiver, simultaneously tracking the same satellites (minimally four) to determine their relative coordinates (see Sect. Satellite Positioning). It is based on the ability to eliminate, or significantly reduce, the common measurement biases across observations made simultaneously by GPS receivers see (Equations 5–8 and Sect.

Errors and Biases). Depending on the accuracy requirements, code pseudo-range as well as carrier phase measurements can be used in relative GPS positioning. In addition, relative positioning can be implemented with different latencies varying from real-time to post-processing:

– Carrier-phase-based relative positioning

The basis of high-precision relative positioning is carrier phase measurements because of their low measurement noise. The problem of processing the "ambiguous" carrier phase data can be overcome by collecting sufficient epochs of data (up to 1 h or more) to ensure ambiguity resolution and strengthen the baseline solution. The reliability of the ambiguity resolution depends on:

- Baseline length
- The number of satellites (the more, the better)
- The satellite-receiver geometry
- Whether observations are made on both frequencies (it is much easier to resolve ambiguities when dual-frequency observations are available)
- The length of the observation session (the longer, the better)

Throughout time, GPS ambiguity resolution methods have become more and more sophisticated. The LAMBDA (Least-squares AMBiguity Decorrelation Adjustment method by Teunissen (1995)) allows now even on-the-fly ambiguity resolution using few observation epochs. This is typically necessary when the receiver is in motion.

At one extreme, static GPS surveying is the most precise positioning technique in use today. Observation times vary between 20 min to a day, depending on the expected precision, the distance between the two receivers, the number of visible satellites and the satellite geometry. Measurements are typically taken at intervals between 10 and 30 s. After completing the field measurements, the data are processed on a PC using different processing options depending on, e.g., observation time and baseline length. For long baselines (>20 km), typically ionosphere-free carrier phases will be processed. For very long baselines (>1,000 km), the user will process the data using high-precision scientific GPS analysis software, e.g., GAMIT (Herring et al., 2007) and Bernese (Dach et al., 2007). In this case, mm to cm precision can be reached over distances exceeding 1,000 km. Typical applications are the surveying of crustal or intraplate deformations (see *GPS, Tectonic Geodesy*).

At the other extreme, real-time-kinematic (RTK) GPS is the less precise relative GPS positioning method based on carrier phases. The main differences to static GPS surveying is the drastic reduction of the observation time and the fact that the position of the remote receiver is often computed in real-time. RTK can be carried out in real-time if an appropriate communications link (UHF radio modem, satellite link, cell phone, or other transmission media, cf. Langley, 1993; Weber et al., 2005) is provided over which the carrier phase data collected at the reference receiver can be made available to the remote receiver's onboard computer. In this way, the remote receiver can generate the double differences, resolve the initial ambiguities, and perform the position calculations. Typically, measurements are taken at a 1 Hz rate and initial ambiguities are determined on-the-fly. In order to mitigate as much as possible the error sources, baselines should not exceed 10–15 km to reach the expected positioning precision of the order of 2–5 cm. Cadastre surveying is one of the applications that uses RTK. More information on RTK can be found in Langley (1998).

Other relative positioning modes based on GPS carrier phases are rapid static positioning and stop-and-go surveying (cf. El-Rabbany, 2002). They are lying between the above-mentioned extremes.

– Code-based relative positioning is mostly known as real-time Differential GPS (DGPS). There are two ways to implement DGPS: (1) the fixed reference station broadcasts the differences between the measured pseudo-ranges and actual (internally computed) pseudo range so that it can correct its pseudo-ranges by the same amount before determining its position (method of range corrections); (2) the fixed reference station compares its known position with its computed position and sends the difference to the remote receiver which uses this difference to correct its computed position. The correction signal is broadcast using similar means as uses RTK. Range corrections are the preferred mode for real-time DGPS navigation. DGPS can deliver meter-level positioning that typically ranges between 0.5 and 5 m.

The standard format used to exchange GPS data in real-time is RTCM 104 (cf. RTCM104, 2009). For the exchange of GPS data targeting post-processing, the RINEX format (Gurtner and Estey, 2007) is used.

Conclusion and outlook

GPS positioning is based on measuring the travel times of radio signals emitted by satellites to a receiver mostly located on the ground. These travel times can be obtained by measuring on the precise, but ambiguous, GPS carrier phases or on the, less precise, GPS PRN codes. By multiplying the travel time by the speed of light, the distances receiver–satellites are reconstructed. Using minimally four satellites, the receiver can compute its position. Errors that influence GPS positioning are receiver-related (measurement noise, multipath, receiver clock errors, antenna phase center variation), satellite-related (satellite orbit and clock error, antenna phase center variation) or related to the signal propagation (ionospheric and tropospheric effect). As a lot of these errors are similar for GPS receivers distant of a few kilometers, relative GPS positioning was developed to improve the point positioning accuracy.

GPS is active since the mid-1980s. It has grown from a purely military navigation system to a scientific tool allowing today to measure deformations of the Earth's crust with millimeter-level precision. Within the next years, existing Global Navigation Satellite Systems (GNSS) like GPS and the Russian GLONASS will launch new types of satellites and provide additional signals to deliver better accuracy, reliability, and availability of positioning. Moreover, Europe and China are also developing their own GNSS, known as Galileo and Compass, respectively (Gibbons, 2009). All these signals have the potential to allow even more precise positioning.

Bibliography

Bauersima, I., 1983. Navstar/GLOBAL Positioning System (GPS) (II). *Mitteilung der Satellitenbeobachtungsstations Zimmerwald*, No. 10, Druckerei der Universität Bern.

Betz, J. W., Bianco, M. A., Cahn, C. R., Dafesh, P. A., Hegarty, C. J., Hudnut, K. W., Kasemsri, V., Keegan, R., Kovach, K., Lenahan, L. S., Ma, H. M., Rushanan, J. J., Sklar, D., Stansell, T. A., Wang, C. C., and Yi, S. K., 2007. Enhancing the future of civil GPS – Overview of the L1C Signal. *Inside GNSS*, Spring 2007, pp. 42–49.

Beutler, G., 1996. GPS Satellite Orbits. In Kleusberg, A., and Teunissen, J. G. (eds.), *GPS for Geodesy*. Berlin, Heidelberg, New York: Springer. Lecture Notes in Earth Sciences, Vol. 60, pp. 37–101.

Beutler, G., Bauersima, I., Gurtner, W., Rothacher, M., and Schildknecht, T., 1988a. Static positioning with the Global Positioning System (GPS): state of the art. In Groten, E., and Strauss, R. (eds.), *GPS-techniques applied to Geodesy and Surveying*. Berlin Heidelberg New York: Springer. Lecture notes in Earth Sciences, Vol. 19, pp. 363–380.

Beutler, G., Bauersima, I., Gurtner, W., Rothacher, M., Schildknecht, T., and Geiger, A., 1988b. Atmospheric refraction and other important biases in GPS carrier phase observations. In Brunner, F. (ed.), *Atmospheric Effects on Geodetic Space Measurements*, Monograph 12, School of Surveying, Kensington: University of New South Wales, pp. 15–43.

Beutler, G., Gurtner, W., Bauersima, I., and Rothacher, M., 1986. Efficient computation of the inverse of the covariance matrix of simultaneous GPS carrier phase difference observables. *Manuscripta Geodetica*, **11**, 249–255.

Bisnath, S., and Gao, Y., 2008. Current state of Precise Point Positioning and future prospects and limitations. In Sideris, M. (ed.), *International Association of Geodesy Symposia "Observing our changing Earth"*, Vol. 133, pp. 615–623.

Boehm, J., Heinkelmann, R., and Schuh, H., 2007. Short Note: a global model of pressure and temperature for geodetic applications. *Journal of Geodesy*, **81**(10), 679–683, doi:10.1007/s00190-007-0135-3.

Boehm, J., Niell, A., Tregoning, P., and Schuh, H., 2006. Global Mapping Function (GMF): a new empirical mapping function based on numerical weather model data. *Geophysical Research Letters*, **33**(7), L07304, doi:10.1029/2005GL025546.

Boehm, J., and Schuh, H., 2004. Vienna mapping functions in VLBI analyses. *Geophysical Research Letters*, **32**(1), L01603, doi:10.1029/2003GL018984.

Braasch, M. S., and Van Dierendonck, A. J., 1999. GPS receiver architectures and measurements. *Proceedings of the IEEE*, **87**(1), 48–64, doi:10.1109/5.736341.

Brunner, F. K., and Welsch, W. M., 1993. Effect of the troposphere on GPS measurements. *GPS World*, **4**(1), 42–51.

Cohen, C. E., and Parkinson, B. W., 1991. Mitigating multipath error in GPS Based Asttitude Determination. In *Proceedings of the Annual Rocky Mountain Guidance and Control Conference*, Keystone, CO, February 2–6, 1991 (A93-15576 03–12), pp. 53–68.

Dach, R., Hugentobler, U., Fridez, P., and Meindl, M. (eds.), 2007. *Bernese GPS Software Version 5.0*. University of Bern, Switzerland: Astronomical Institute.

Dow, J. M., Neilan, R. E., and Rizos, C., 2009. The international GNSS Service in a changing landscape of global navigation satellite systems. *Journal of Geodesy*, **83**, 191–198, doi:10.1007/s00190-008-0300-3.

El-Rabbany, A., 2002. *Introduction to GPS*. Norwood: Artech House.

Elósegui, P., Davis, J. L., Jaldehag, R. T. K., Johansson, J. M., Niell, A. E., and Shapiro, I. I., 1995. Geodesy using the global positioning system: the effects of signal scattering on estimates of site positions. *Journal of Geophysical Research*, **1000**, 9921–9934.

Evans, A.G., 1986. Comparison of GPS pseudorange and biased Doppler range measurements to demonstrate signal multipath effects. In *Proceedings of the Fourth International Geodetic Symposium on Satellite Positioning*, Austin, TX, Vol. 1, pp. 761–587.

Galileo, 2009. Report from the Commission to the European Parliament and the Council on the implementation of the GNSS programmes and on future challenges, http://ec.europa.eu/enterprise/policies/space/documents/galileo/index_en.htm

Georgiadou, Y., and Kleusberg, A., 1988. On the effect of ionospheric delay on geodetic relative positioning. *Manuscripta Geodetica*, **13**, 1–8.

Ghoddousi-Fard, R., and Dare, P., 2005. Online GPS processing services: an initial study. *GPS Solutions*, **10**(1), 12–20, doi:10.1007/s10291-005-0147-5.

Gibbons, G., 2009. What race? What competition. *Inside GNSS*, March/April 2009, pp. 16–19.

GPS-SPS, 1995. Global Positioning System Standard Positioning Service Signal Specification 2nd edition. http://www.navcen.uscg.gov/gps/geninfo/2001SPSPerformanceStandardFINAL.pdf.

Görres, B., Campbell, J., Becker, M., and Siemes, M., 2006. Absolute calibration of GPS antennas: laboratory results and comparison with field and robot techniques. *GPS Solutions*, **10**(2), 136–145, doi:10.1007/s10291-005-0015-3.

Gurtner, W., and Estey, L., 2007. The Receiver Independent Exchange Format version 3.01, http://igscb.jpl.nasa.gov/igscb/data/format/rinex301.pdf.

Herring, T. A., King, R. W., and McClusky, S. C., 2007. Introduction to GAMIT/GLOBK Release 10.3. Cambridge: Massachusetts Institute of Technology.

ICD-GPS-200C, 2000. GPS Interface Control Document. http://www.navcen.uscg.gov/pubs/gps/icd200/icd200cw1234.pdf.

IERS Conventions, 2003. http://www.iers.org/IERS/EN/Publications/TechnicalNotes/tn32.html.

IS-GPS-200D, 2004. GPS Interface Specification Revision D. ARINC Engineering Services, LLC, http://www.navcen.uscg.gov/gps/geninfo/IS-GPS-200D.pdf.

Jonkman, N. F., and De Jong, K., 2000. Integrity monitoring of IGEX-98 data, part II: cycle slip and outlier detection. *GPS Solutions*, **3**(4), 24–34, doi:10.1007/PL00012812.

Kim, D., and Langley, R. B., 2002. Instantaneous real-time cycle-slip correction for quality control of GPS carrier-phase measurements. *Navigation*, **49**(4), 205–222.

Klobuchar, J. A., 1986. Design and characteristics of the GPS ionospheric time delay algorithm for single frequency users. In *Proceedings IEEE Plans 1986 Position, Location and Navigation Symposium*, Las Vegas, pp. 280–286.
Klobuchar, J. A., 1991. Ionospheric effect on GPS. *GPS World*, **2**(4), 48–51.
Kouba, J., and Héroux, P., 2001. Precise point positioning using IGS Orbit and Clock products. *GPS Solutions*, **5**, 12–28.
Langley, R. B., 1993. Communications links for DGPS. *GPS World*, **4**(4), 47–51.
Langley, R. B., 1998. RTK GPS. *GPS World*, **4**(4), 70–76.
Leick, A., 2004. *GPS Satellite Surveying*, 4th edn. New York: Wiley, p. 435.
Lutgens, F. K., and Tarbuck, E. J., 1989. *The Atmosphere: An introduction to Meteorology*, 4th edn. Englewood: Prentice Hall.
Mader, G. L., 2001. A comparison of absolute and relative GPS antenna calibrations. *GPS Solutions*, **4**(4), 37–40, doi:10.1007/PL00012864.
Niell, A. E., 2000. Improved atmospheric mapping functions for VLBI and GPS, *Earth, Planets, and Space*, **52**, 699–702.
Parkinson, B. W., 1994. GPS eyewitness: the early years. *GPS World*, **5**(9), 32–45.
RTCM104, 2009. Documents Developed by RTCM Special Committee 104 on Global Navigation Satellite Systems (GNSS) Service, http://www.rtcm.org/.
Saastamoinen, J., 1972. Atmospheric correction for the troposphere and stratosphere in radio ranging of satellites. In Henriksen, S. W., Mancini, A., and Chovitz, B. H. (eds.), *The use of artificial satellites for geodesy*. Washington, DC: AGU. Geophysical Monograph Series 15, pp. 247–251.
Santerre, R., 1991. Impact of GPS satellite sky distribution. *Manuscripta Geodetica*, **16**, 28–53.
Schmid, R., Steigenberger, P., Gendt, G., Ge, M., and Rothacher, M., 2007. Generation of a consistent absolute phase-center correction model for GPS receiver and satellite antennas. *Journal of Geodesy*, **81**(2), 781–798, doi:10.1007/s00190-007-0148-y.
Spilker, J., 1978. GPS signal structure and performance characteristics. *Journal of the Institute of Navigation*, **25**, 121–146.
Teunissen, P. J. G., 1995. The least-squares ambiguity decorrelation adjustment: a method for fast GPS integer ambiguity estimation. *Journal of Geodesy*, **70**(1–2), 65–82.
Tranquilla, J. M., Colpitts, B. G., and Cam J. P., 1989. Measurement of low- multipath antennas for Topex. In *Proceedings 5th International Geodetic Symposium on Satellite Positioning*, Las Cruces, pp. 356–361.
U.S. Coast Guard, 1995. News Release: Global Positioning System Fully Operational. http://www.navcen.uscg.gov/gps/geninfo/global.htm.
Weber, G., Dettmering, D., and Gebhard, H., 2005. Networked Transport of RTCM via Internet Protocol (NTRIP). In Sanso, F. (ed.), *International Association of Geodesy Symposia 'A Window on the Future of Geodesy'*, pp. 60–64, doi:10.1007/3-540-27432-4_11.
Wells, D., Beck, N., Delikaraoglou, D., Kleusberg, A., Kralaiwsky, E., Lachapelle, G., Langley, R., Nakiboglu, M., Schwarz, H.-P., Tranquilla, J., and Vanicek, P., 1987. *Guide to GPS Positioning*. University of New Brunswick: Canadian GPS associates.
Woo, K. T., 1999. Optimum semi-codeless carrier phase tracking of L2. In *Proceedings 12th International Technical Meeting of the Satellite Division of the Institute of Navigation*, Nashville, TN, September 14–17, pp. 289–305.
Wu, J., and Hsieh, C. H., 2010. Statistical modeling for the mitigation of GPS multipath delays from day-to-day range measurements. 2010. *Journal of Geodesy*, **84**(4), 223–232, doi:10.1007/s00190-009-0358-6.
Zhodzishsky, M., Vorobiev, M., Khvalkov, A., and Ashjaee, J., 1998. The First Dual-Depth Dual-Frequency Choke Ring. In *Proceedings ION GPS-98*, pp. 1035–1040.
Zumberge, J., Heflin, M., Jefferson, D., Watkins, M., and Webb, F., 1997. Precise point positioning for the efficient and robust analysis of GPS data from large networks. *Journal of Geophysical Research*, **102**, 5005–5501.

Cross-references

Geodesy, Networks and Reference Systems
GPS, Tectonic Geodesy

GPS, TECTONIC GEODESY

Jeffrey T. Freymueller
Geophysical Institute, University of Alaska Fairbanks, Fairbanks, AK, USA

Synonyms

Crustal deformation

Definition

Earthquake cycle. A conceptual model that explains the buildup of stress that causes earthquakes through a period of steady tectonic loading that culminates in sudden slip in an earthquake.
Slip deficit. The difference between the slip that has occurred over a period of time on a fault and the long-term fault slip rate; the slip deficit is a measure of the accumulated stress available to cause an earthquake on the fault.
Coseismic displacement. The final, static displacement of the surface caused by slip in an earthquake.
Postseismic displacement. Displacements caused by postseismic processes, which are transient deformation processes that follow large earthquakes.
Volcanic inflation. Dilatational deformation caused by pressurization of a volcanic system.
Glacial isostatic adjustment (GIA). The deformation, mainly uplift, that results from the removal of glacial surface loads.

Tectonic geodesy

Introduction

Tectonic geodesy uses measurements of changes in the position of a network of points on earth's surface to study motion and deformation of the earth. These motions mainly result from plate motions, slip on active faults (including earthquakes), and the pressurization or depressurization of magma systems beneath volcanoes, but significant deformation can also result from changes in surface loads such as glacial isostatic adjustment. It is one of the most important tools for the study of plate boundary zones, the regions of distributed deformation

that are commonly found at the boundaries of the major tectonic plates.

Measurements of change depend on a precise time series of positions in a stable terrestrial reference system (see *GPS, Data Acquisition and Analysis* and *Geodesy, Networks and Reference Systems*), and the principal measurement system today is the Global Positioning System (GPS). GPS is most commonly used because of the relatively low cost of the equipment and its ease of use in the field, but the same principles apply to positions measured by any space geodetic technique. Usually, GPS time series is dominated by steady, linear motion over time, so site velocities are frequently used to describe site motions. For sudden or large deformations, especially due to earthquake and volcanic activity, measurements of motion from Interferometric Synthetic Aperture Radar (InSAR) have also become very important and are complementary to GPS (see *SAR Interferometry*).

Although tectonic geodesy has become prominent over the last couple of decades, the field dates back more than a century and many of the important conceptual and mathematical models relating deformation measured at the surface to sources within the earth pre-date the availability of precise, modern geodetic data. During the last three decades, the focus of tectonic geodesy has changed from measuring motions, assumed to be steady in time, to a more general representation of time-dependent deformation. This evolution in theory and practice has resulted from the steady improvement in precision and accuracy of the measurements along with the realization that many processes are not steady in time.

History of tectonic geodesy

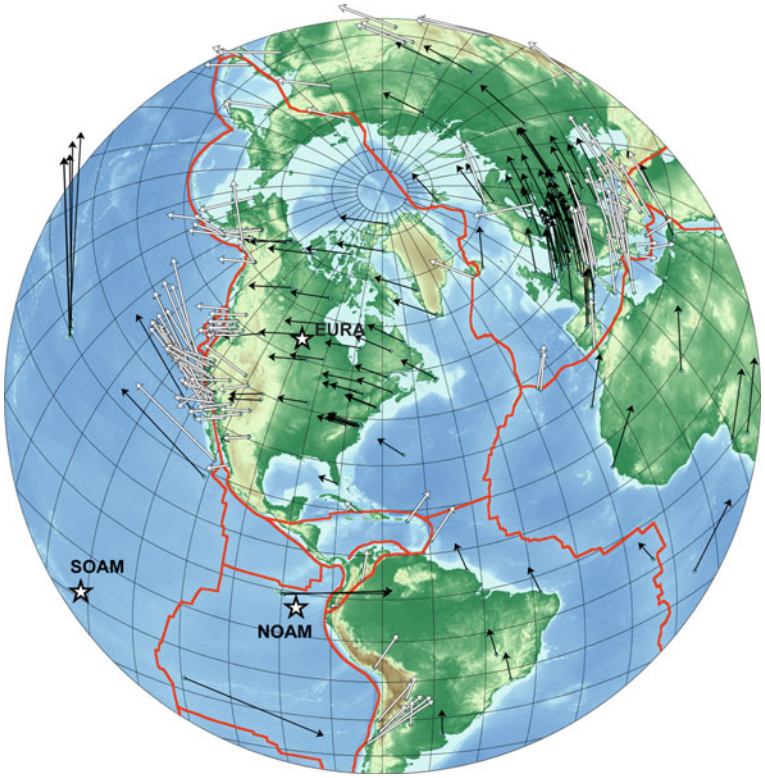
Early work in tectonic geodesy dates back more than a century. The Dutch surveyor J. J. A. Müller reported ground movements associated with the 1892 Tapanuli earthquake in Sumatra, which caused changes in observed angles (Bonafede et al., 1992). H. F. Reid used displacements computed from geodetic surveys conducted before and after the 1906 San Francisco earthquake as the basis for the elastic rebound hypothesis (Reid, 1910, 1911). These observations supported the hypothesis proposed by G. K. Gilbert a few decades earlier that earthquakes were the result of a sudden release of strain that had built up over a long period of time (Gilbert, 1884). Numerous studies over the succeeding decades used all kinds of terrestrial geodetic measurements to document both sudden displacements due to earthquakes and the slow, steady deformation between earthquakes. By the 1970s, the US Geological Survey had begun to make extensive and systematic terrestrial geodetic measurements for the purpose of measuring crustal deformation, and similar developments occurred in Japan and other countries that experienced active tectonics. Because of the limited spatial scale of the terrestrial measurements, these efforts generally focused on the deformation associated with individual faults or fault systems. Although restricted in spatial scale and coverage, these terrestrial data prompted the development of most of the conceptual and numerical models used in tectonic geodesy today.

Measurement of tectonic motions over large spatial scales began in the 1980s with NASA's Crustal Dynamics Project and related international efforts (e.g., the WEGENER project in Europe). This work used a network of fixed (and a few mobile) Very Long Baseline Interferometry (VLBI) and Satellite Laser Positioning, Ranging (SLR) stations (see *Very Long Baseline Interferometry* and *Satellite Laser Ranging*), and achieved its goals of comparing plate motions over a decadal timescale to geologic plate motions (e.g., Smith et al., 1990; Stein, 1993; Robbins et al., 1993; Robaudo and Harrison, 1993) and measuring deformation across broad plate boundary zones in North America and the Mediterranean (e.g., Ma et al., 1990; Gordon et al., 1993; Wilson and Reinhart, 1993; Gendt et al., 1993). Since the early 1990s, the tremendous expansion of GPS networks has resulted in a vast data set available to study plate motions (e.g., Sella et al., 2002) and active deformation of plate boundary zones (e.g., McCaffrey et al., 2007; Meade, 2007).

Today, data from thousands of GPS stations are available and used by researchers around the world. A growing number of these stations also record data from other Global Navigation Satellite Systems (GNSS), such as the Russian GLONASS or European Union's Galileo. These data come from global and regional networks of fixed instruments, as well as from survey markers measured episodically. High-accuracy GPS satellite orbits are available from the International GNSS Service (IGS), and the stations that contribute to the IGS tracking network form a reference network that any researcher can use to determine station positions in the International Terrestrial Reference Frame (ITRF) (Altamimi et al., 2007). Modern studies in tectonic geodesy are based on motions derived from time series of site positions expressed in the ITRF, on regional to global scales.

Plate motions

Plate tectonic theory holds that the surface of the earth can be divided into a set of rigid plates moving relative to each other (Figure 1), with tectonic activity concentrated on the boundaries of the plates and caused by the relative plate motions. The motion of a rigid plate on the surface of a sphere is a rotation about an axis passing through the geocenter, described by a single angular velocity vector. The surface projection of this axis of rotation is called the pole of rotation, or the Euler pole. Models of plate motions thus consist of angular velocities of plates relative to some reference frame, or angular velocities of relative plate motions. Magnetic anomalies at spreading centers and transform fault azimuths record the rates and directions of past relative plate motions, and estimates of current plate motions estimated from these data (averaged over periods of up to a few million years) can be compared



GPS, Tectonic Geodesy, Figure 1 Velocities of geodetic sites illustrate both rigid plate motion and plate boundary deformation. Velocities are from Argus et al. (2010), in the ITRF2005 reference frame. Black vectors are sites on plate interiors, and illustrate rigid rotational motion of the plates. White vectors are sites within plate boundary zones, and generally show motions intermediate between the rigid plates that bound the plate boundary zone. The stars show the poles of rotation (in ITRF) for the North American (NOAM), South American (SOAM), and Eurasian (EURA) plates, also derived from Argus et al. (2010). For all three of these plates, the plate rotation is counterclockwise about the pole. The main plate boundaries are shown in red, based on DeMets et al. (1990).

with geodetic estimates averaged over the last 2–3 decades (DeMets et al., 2010; Argus et al., 2010). The latest work shows that the approximation of rigid plates moving at a steady rate over millions of years is quite good, but there is evidence for small changes in plate motions over the last few million years, and evidence for small but significant internal deformation of the plates.

In the early history of tectonic geodesy, geodetic results were compared to models of geologically current plate motions, which were more precise at that time and whose level of accuracy was understood. Even at that time the geodetic data made a significant contribution. Robbins et al. (1993) compared rates of plate motion from SLR and VLBI to predictions of the geological plate model NUVEL-1 (DeMets et al., 1990), and found a very high correlation between the geodetic and geologic estimates based on magnetic anomalies at spreading centers and transform fault azimuths. However, the geodetic plate motion rates were systematically $6 \pm 1\%$ slower than geologic rates. This discrepancy resulted mainly from errors in the geomagnetic timescale; after recalibration based on the revised timescale, the NUVEL-1A plate motion model agreed with space geodetic estimates (DeMets et al., 1994). The conclusion of this work was that, with perhaps a few exceptions, plate motions averaged over the last ~3 million years and plate motions averaged over the last decade were the same within the measurement precision of the time.

The full capability of space geodesy to measure plate motions was not realized until long time series of GPS data were available, beginning in the mid to late 1990s. The spatial density and coverage of GPS data meant that more plates could be included and the measurements for each plate were more robust. The first global plate motion models based on GPS data alone were Argus and Heflin (1995) and Larson et al. (1997). Larson et al. (1997) found that the motion of the Nazca plate was significantly slower than predicted by NUVEL-1A, a result later confirmed by Norabuena et al. (1998, 1999), who found that the rate of subduction of the Nazca plate beneath South America has declined steadily over the last 20 million years. The GEODVEL model of Argus et al. (2010) is the most complete global geodetic plate motion model to date. GEODVEL combined GPS, VLBI, SLR, and DORIS site

velocities (Figure 1) and presented an estimate of relative plate angular velocities based on these data. The earlier REVEL model (Sella et al., 2002) presented the angular velocities of each plate in ITRF97, and although now somewhat dated it remains useful for realizing a reference frame fixed to a particular plate. Estimates of motions of the major plates in ITRF have also been included in the entries that describe the successive realizations of ITRF (e.g., Altamimi et al., 2007), although these estimates are based on smaller data sets than GEODVEL or REVEL.

The central results of these studies are that the present-day motions of the plate interiors are similar to plate motions averaged over the last ~3 million years, and outside of known regions of tectonic activity, the plate interiors are rigid to a level comparable to the measurement precision of GPS (Calais et al., 2006; Argus et al., 2010). However, for most plate pairs the small differences between GEODVEL and the NUVEL-1A (median difference of 0.063° Myr^{-1}) are nonetheless statistically significant (Argus et al., 2010). The comparison of GEODVEL and the most recent geological model MORVEL (DeMets et al., 2010) reveals smaller but still significant differences. This implies that changes in plate motions of a few to several percent have occurred over the last 3 million years, most notably the reduction in angular speed of the Nazca plate relative to South America. Calais et al. (2006) analyzed internal deformation of the eastern part of the North American plate, and found a weighted RMS residual horizontal velocity of 0.7 mm/year. However, horizontal residuals showed a systematic spatial pattern, interpreted to be the result of glacial isostatic adjustment (GIA). Horizontal deformation from GIA complicates the determination of the angular velocity of the plate itself, so Sella et al. (2007) used a set of 124 sites selected to minimize the effects of GIA and used these sites to determine an updated estimate of the North American plate. The unweighted RMS residuals for these sites were 0.6 mm/year, consistent with the estimated precision of the measurements. The Indian plate may be an exception. Banerjee et al. (2008) estimated 2 ± 1 mm/year of shortening across central India, which could reflect contraction driven by the collision of India with Eurasia.

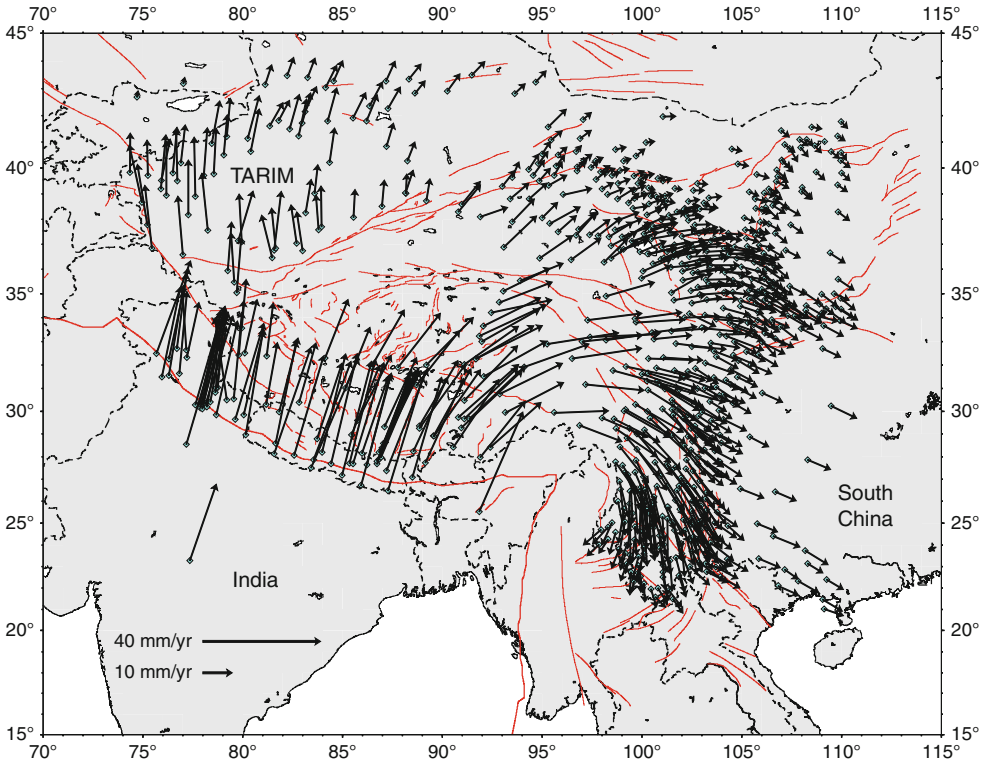
The estimated plate motions from tectonic geodesy are also important for studies of plate boundary zones, which are zones of distributed deformation on the boundary between two or more plates (Figures 1 and 2). The steady interseismic motions of areas within a plate boundary zone between two major plates must be understood in context of the motions of the major plates. Geodetic motions are understood most easily in terms of geologic structures when the velocities of sites are expressed relative to one of the major plates. Plate boundary zones can extend over hundreds or thousands of kilometers and involve many active structures, especially within the continental crust (Thatcher, 2003, 2009). Plate boundary zones typically include rigid blocks or microplates within them, such as Tarim and South China in the broad and complex India – Eurasia boundary zone (Figure 2). Plate boundary zone deformation in Asia and North America (Figure 1) has received a great deal of study, and work on these plate boundary zones requires an integration of the concepts of plate tectonics with the earthquake cycle and interseismic deformation, along with other causes of deformation detailed later in this entry.

Steady "Interseismic Deformation"

Time series of site positions is typically dominated by steady, quasi-linear motion with time, although periodic seasonal variations are also commonly observed (especially in the vertical). This is true not only for sites in the interiors of tectonic plates, but also for many sites within plate boundary zones. As a result, the time series of positions is commonly reduced to a site velocity based on a linear fit and velocities are then used to study the steady motions with time. Offsets in position due to earthquakes and time-dependent motions due to postseismic deformation and other effects will be discussed in later entries. For now, it is simply important to note that the linear fit for velocity needs to take account of any discontinuities or known time-dependent variations in the time series, whether due to motions of the earth or to changes in equipment.

The steady, linear motion of the tectonic plates leads to the concept of an "earthquake cycle," in which tectonic stresses increases linearly with time until a fault failure criterion is reached, at which point an earthquake occurs and the stored stresses are relieved by fault slip (e.g., Scholz, 1990). The basic concept goes back to Reid's elastic rebound hypothesis, in which he proposed that the strain accumulation pattern between earthquakes would be opposite to the strain release pattern during the earthquakes (Reid, 1910). The net result of one full cycle is a block-like offset along with fault, with no strain off the fault – much like what would be observed by geologists (Figure 3). Given linear stress accumulation and an elastic rheology for the earth, strain (deformation) should occur at a constant linear rate between earthquakes. Elastic earthquake cycle models thus imply a long "interseismic" period of linear deformation with time, punctuated by offsets at the times of large earthquakes. Although there is unambiguous observational evidence for transient postseismic deformation following earthquakes, the observed deformation prior to large earthquakes has generally been linear with time within measurement precision, and the concept of "interseismic deformation" remains useful.

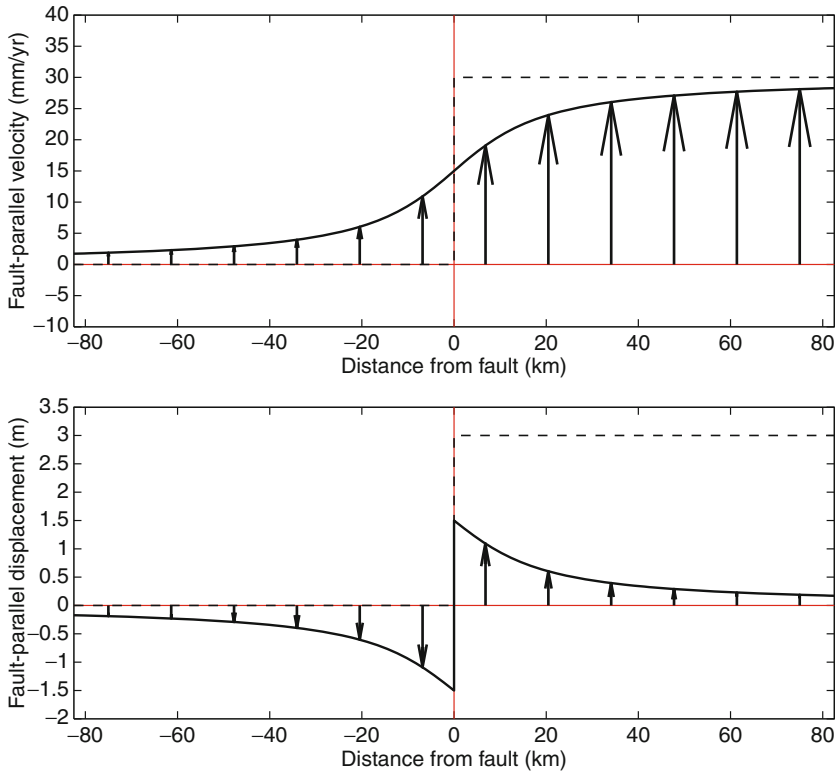
Numerical models for the interseismic period follow directly from the assumptions of the earthquake cycle model. Large earthquakes within continental crust rupture a limited range of depths, from at or near the surface to a lower depth limit that is usually no deeper than 10–20 km. Observations of exhumed deep fault zones show evidence for ductile shear, presumed to occur continuously with time (for a comprehensive summary, see



GPS, Tectonic Geodesy, Figure 2 Velocities of sites in the India-Eurasia plate boundary zone, relative to the Eurasian plate. GPS velocities are from Gan et al. (2007), and faults (*red lines*) are from Taylor and Yin (2009). Rigid rotation of two microplates or small plates, Tarim and South China, is clearly visible in the velocity field at this scale. Most of the motions shown here can be explained by a combination of rigid block motion, including a number of smaller blocks, and elastic deformation that results from the shallow locked zones of block-bounding faults.

Scholz, 1990). Savage and Burford (1970) proposed a simple interseismic model using elastic dislocation theory. In their model, the deeper part of the fault zone creeps continuously at the long-term fault slip rate, while the shallow part of the fault zone remains stuck by friction except in earthquakes. They represented this numerically by an elastic dislocation plane that slips steadily from a locking depth d to infinite depth. The dislocation is embedded in an elastic half-space, which represents the earth from its surface to infinite depth as an elastic body. Alternately, the slip on the fault as a function of depth can be represented by a linear superposition between uniform slip at all depths and backward slip or backslip on the shallow part of the fault. The backslip is a mathematical construct, one component of the linear superposition; the fault does not actually slip backwards. Backslip represents the *slip deficit*, or the difference between the slip that actually occurs over a period of time, and the slip expected from the long-term average rate of fault slip. Slip deficit is proportional to the accumulated stress. In the simplest earthquake cycle model, the accumulated slip deficit over the interval between earthquakes would be equal to the slip in an earthquake. In the two-dimensional case of an infinitely long strike-slip fault, the equations describing the dislocation model are very simple, with the velocity of a site located at a distance x from the fault $V = (s/\pi)$*arctan(x/d), where s is the average slip rate, and d is the locking depth (Figure 1). This velocity is expressed in a frame in which the velocity is zero at the fault and reaches plus or minus half the long-term fault slip rate far from the fault. When the locking depth is shallow, the interseismic strain is concentrated close to the fault, and when it is deep, the strain is distributed over a larger region.

This simple model has proven to be extremely successful in a wide variety of tectonic settings, and the same conceptual model has been applied to normal faulting and to subduction zones (Savage, 1983) in addition to strike-slip faults. In all cases, the critical step in describing interseismic deformation is a superposition between steady fault slip at the long-term fault slip rate, assumed to result in block-like or plate-like motion with all offset occurring as slip on the fault and no strain occurring in the surrounding rock, and the deformation associated with the accumulating slip deficit. When the locking depth varies along strike, the pattern of strain will also vary along strike; in the extreme limit of a locking depth of zero, the fault creeps at the surface at its long-term slip rate. Observations show a wide range of fault locking depths, generally in agreement with depths of maximum earthquake slip or depths of seismicity. Locking depths



GPS, Tectonic Geodesy, Figure 3 Velocities and coseismic displacements expected from a simple elastic earthquake cycle model. The fault in this model is an infinitely long strike-slip fault (*left-lateral*) with a slip rate of 30 mm/year, which has an earthquake after 100 years with 3 m slip. During the interseismic period, the fault is locked from the surface to 15 km depth and slipping steadily below that. The interseismic deformation pattern (*top panel*) is shown by the arrows and solid line, relative to the far field on the left side of the fault. Compared to the uniform block motion that would result if the fault slipped uniformly at all depths (*dashed line*), the elastic response of the earth spreads the fault shear over a broad area. The coseismic displacement (*bottom panel*) is antisymmetric about the fault, with maximum displacement at the fault (each side moves by half the coseismic slip, in opposite directions). The sum of 100 years of interseismic deformation and the coseismic displacement is a uniform block motion, shown by the dotted line.

sometimes vary abruptly in the along-strike direction. Examples of along-strike changes in fault locking behavior include variations in the rate of fault creep along the creeping section of the San Andreas fault in California (Titus et al., 2005), and variations in the extent of the locked zone along the subduction zone of the Aleutian and Sumatran arcs (Cross and Freymueller, 2008; Freymueller et al., 2008; McCaffrey, 2008; Chlieh et al., 2008). The parts of the fault determined (from geodetic data) to be locked by friction and accumulating slip deficit correspond well to the rupture areas of the last set of great earthquakes, and the along-strike boundaries between locked and creeping segments correspond to the limits of seismic slip or major seismic asperities of the last great earthquakes.

The elastic dislocation approach to modeling interseismic deformation has been extended to three dimensions using elastic block modeling (e.g., McCaffrey, 2002; Meade and Hager, 2005a). In this approach, which blends the description of plate motions in terms of angular velocities with an elastic interseismic model, the model domain is first broken up into a set of blocks or microplates, each of which is assumed to be rigid. The boundaries between adjacent blocks are thus the faults on which the relative block motion occurs. The velocity of a GPS site is the sum of two components, the rotation of the block it lies upon (described by the block angular velocity) and the elastic deformation caused by the slip deficit on all faults in the model. The elastic block model enforces self-consistency of block motions, fault slip rates, and elastic deformation, because the fault slip rates are computed from the block angular velocities and the elastic deformation is computed using dislocation theory from the fault slip rates and assumed fault locking depths. Thus, the only estimated parameters in an elastic block model are the block angular velocities, unless the fault locking depths or other geometric parameters are explicitly optimized as well. This approach has been successful in describing deformation in North America, Asia, the Mediterranean and Middle East, and other parts of the world, and has sometimes been combined with an assumption of uniform strain for some blocks rather than

rigidity (McCaffrey, 2002; Nyst and Thatcher, 2004; Meade and Hager, 2005a, b; Reilinger et al., 2006; McCaffrey et al., 2007; Meade, 2007, Loveless and Meade, 2010).

Despite the general success of this approach, there is no way to account for observed time dependence in postseismic deformation within this model, and there are some significant discrepancies between fault slip rates estimated from geodetic data and geologically estimated rates from dated fault offsets. A prominent example of this involves the Garlock Fault and Eastern California Shear Zone in the Mojave Desert of California. The Garlock fault is a left-lateral strike-slip fault that is in a conjugate orientation to the right-lateral San Andreas Fault and the faults of the Eastern California Shear zone. The present geodetic slip rate estimated for the Garlock fault (Meade and Hager, 2005a) is significantly lower than expected based on the recent geological slip rate (McGill et al., 2009), while the reverse is true for the adjacent Eastern California Shear Zone (Peltzer et al., 2001; Oskin et al., 2008). Dolan et al. (2007) and Oskin et al. (2008) suggested that time-dependent behavior of ductile shear zones in the deep crust may result in time-dependent fault slip, which in this case might result in switching between two modes: elevated slip rate on the Eastern California Shear Zone with low slip on the Garlock fault (the present), and elevated slip on the Garlock fault with low slip on the Eastern California Shear Zone (much of the geologically recent past). Time-dependent postseismic deformation not accounted for in the geodetic models might play a role in this discrepancy.

Coseismic deformation

Coseismic displacements are the sudden displacements caused by earthquakes. Coseismic displacements from large earthquakes can extend a long distance from the fault due to the elastic response of the earth (Figure 3). The earliest observations in tectonic geodesy were of the angle changes between triangulation stations caused by coseismic displacements. Today, GPS and InSAR displacements provide precise and spatially dense measurements of coseismic displacements, allowing detailed models of the distribution of slip in earthquakes to be determined. Very large earthquakes are followed by a period of postseismic deformation, during which the deformation pattern is very different from the pre-earthquake pattern (see next section).

One of the earliest important large earthquakes during the modern GPS era was the 1992 Landers earthquake in southern California. This M_W 7.4 strike-slip earthquake ruptured multiple faults with a complex surface rupture pattern (Sieh et al., 1993) and caused immediate displacements of five continuous GPS sites in the region, which had been installed a few years earlier. Displacements from these continuous GPS sites were used to develop slip models for the earthquake (Bock et al., 1993; Blewitt et al., 1993). Later, a much more extensive set of displacements determined from repeated GPS surveys of many survey marks in the region provided almost 100 displacement vectors for use in studying the distribution of slip during the earthquake (Hudnut et al., 1994; Freymueller et al., 1994). The pre-earthquake surveys were of variable precision and were made over a range of time before the earthquake. Many of the important near-field displacement vectors came from survey marks that had been surveyed originally by private land surveyors or by the California Department of Transportation. In addition, the post-earthquake surveys, while more uniform in precision, were carried out over a period of a few weeks to months during which postseismic deformation was occurring. Removing pre- and post-earthquake deformation from the coseismic displacements was, and still is, a significant challenge. Even in the era of dense continuous networks, the problem of separating coseismic and postseismic deformation has been pushed down to shorter timescales, as postseismic deformation is observed to begin immediately after the earthquake rupture terminates (Melbourne et al., 2002; Savage and Langbein, 2008).

The Landers earthquake and the studies that followed it provide several lessons that remain relevant today. Even a small number of geodetic displacements provide significant constraints on the total seismic moment of an earthquake. Six different geodetic studies for the Landers earthquake were published within the first 2 years after the event, and the total moment estimated in all cases was the same within 10% even though some studies used as few as four or five displacements and others used as many as 96. Determining the moment of the earthquake to within 10% means determining the magnitude to within ± 0.1, which is comparable to the best magnitude precision from seismology. More displacements allowed a more detailed model for the distribution of slip on the fault plane, but changed the total moment (proportional to slip times area) only incrementally. The GPS displacements were compatible with offset measurements made by geologists working on the surface rupture, but slip at the surface rupture was not always indicative of the typical slip at depth. In general, the maximum slip was located some distance beneath the surface. The slip distribution was highly variable in space even along parts of the rupture that were geometrically simple. Estimates of the slip distribution from geodesy and seismology were also compatible (Cohee and Beroza, 1994).

The recent sequence of great earthquakes along the Sumatra subduction zone provides a good picture of the present capabilities for studying earthquakes using geodesy (McCaffrey, 2008). Coseismic displacements from the 2004 Sumatra-Andaman earthquake, the first and largest of the sequence, were large enough to be measured more than 2,000 km from the fault, and probably as far as 4,000 km distant (Banerjee et al., 2005). Significant displacements, large enough to be informative about the earthquake slip distribution, were measured throughout Southeast Asia and across the Indian Ocean on the Indian subcontinent. By the time of the subsequent earthquakes in Sumatra, continuous GPS networks extended across

most of the area, and the coseismic displacement fields were measured rapidly and easily (e.g., Hsu et al., 2006; Kreemer et al., 2006). Additional GPS surveys in the affected areas provided additional displacements from a spatially dense network of sites.

The problem of estimating the earthquake slip distribution from surface displacement data is similar to that of modeling interseismic deformation. As in the interseismic case, elastic dislocation theory provides the mathematical basis for predicting surface displacements given slip on a fault and the inverse problem of estimating slip given surface displacement. Most studies have used the simplest elastic model, a uniform elastic half-space, although some studies have included the effect of elastic layering and/or sphericity of the earth. For the largest earthquakes, a model that treats the earth as a layered elastic sphere is necessary (Banerjee et al., 2005). As in the interseismic case, the fault geometry may or may not be known, and the detailed fault geometry must be estimated as part of the inverse problem if is not known from other sources. What is most different about the coseismic slip problem is that spatial variations in the slip always must be estimated and, because the displacements are larger than in the interseismic case, the data have the strength to resolve spatial variations in slip to a greater extent than in the interseismic case.

The most common approach to estimating the slip distribution is to segment the fault plane or planes into a grid of small elements, or subfaults. The slip on each subfault is estimated, subject to spatial smoothing or other regularization to make the problem invertible. As a general rule, it is necessary to discretize the fault into a grid that is small compared to the actual spatial resolution of the data; the weight given to smoothing is then optimized according to some criterion. The reason for this is that actual spatial variations in slip during earthquakes can be quite large over relatively short distances, and a grid that is too coarse can compromise fit to the data significantly by forcing the boundary between a high slip and low slip region to be in the wrong place. An alternate approach to spatial smoothing is to discretize the fault at a fine spacing first, and then use the model resolution matrix to lump together adjacent subfaults that are poorly resolved (Pritchard et al., 2002). In this approach, a single slip parameter refers to a large area of the fault where there are no nearby data that can resolve variations in slip, and where data are dense the ability to resolve short-wavelength variations in slip is retained. Earthquake slip inversions commonly use nonlinear optimization techniques, from least squares with inequality constraints (to constrain the slip direction) to full nonlinear inversions that are required when the fault geometry must be estimated.

Over time, the amount and accuracy of data used in coseismic slip inversions have increased, and improvements in the terrestrial reference frame have made it possible to use absolute displacements, even for distant sites. Studies of the Landers earthquake used nearly 100 displacements, the 2002 Denali earthquake more than 200, and the 2008 Wenchuan earthquake in China more than 400 GPS displacements. Early earthquake slip inversions using GPS data generally used displacements relative to a local site, because the site positions in an absolute frame were not accurate enough. The inversion for the 2002 Denali Fault Earthquake in Alaska (Hreinsdóttir et al., 2006) used absolute displacements in ITRF, even for sites as much as 800 km from the fault, and Banerjee et al. (2005) determined and modeled absolute displacements for the Sumatra-Andaman Earthquake up to 4,000 km away. The greater density and spatial distribution of data for more recent earthquakes allows estimation of the fault geometry along with the slip distribution as well as resolution of greater detail in the slip distribution. GPS displacements can be augmented with displacements from InSAR, surface offsets from geological measurements from correlation of remote sensing images, and seismic data.

Although GPS for tectonic geodesy is most often used in static mode, in which up to a day of data are used to determine a single daily position for each station, GPS can also be used in kinematic mode to measure the ground motions associated with earthquakes. Nikolaidis et al. (2001) first demonstrated that kinematic GPS could record displacements due to seismic waves, but they were limited by having only data recorded once every 30 s. Larson et al. (2003) measured the propagation of surface waves from the 2002 Denali fault earthquake using GPS data recorded at 1 Hz, which demonstrated the utility of kinematic GPS for "GPS Seismology." GPS displacement records are much less precise than records from seismometers, but can record large displacements faithfully, are not subject to distortion in the presence of tilts, and do not need external information to resolve integration constants because they provide a displacement record directly. Miyazaki et al. (2004a) determined a time-dependent slip model for the 2003 Hokkaido Tokachi-oki earthquake in Japan based entirely on kinematic GPS records. This model clearly showed the spatial evolution of slip on the fault, which propagated downdip and to the northwest from the hypocenter. The kinematic GPS records also showed a secondary slip peak to the northeast, which might represent triggered slip. The spatial and temporal resolution of slip models estimated from kinematic GPS data is superior to that of teleseismic recordings, and are comparable to models derived from near-field strong motion recordings (Ji et al., 2004).

Postseismic deformation

Postseismic deformation is the general term used for any kind of transient deformation that follows earthquakes. Because of the large stress changes associated with large earthquakes, the pattern and rate of strain around the fault is often quite different immediately after the earthquake compared to the pre-earthquake time period, and the time evolution of strain depends on the rheology of the crust

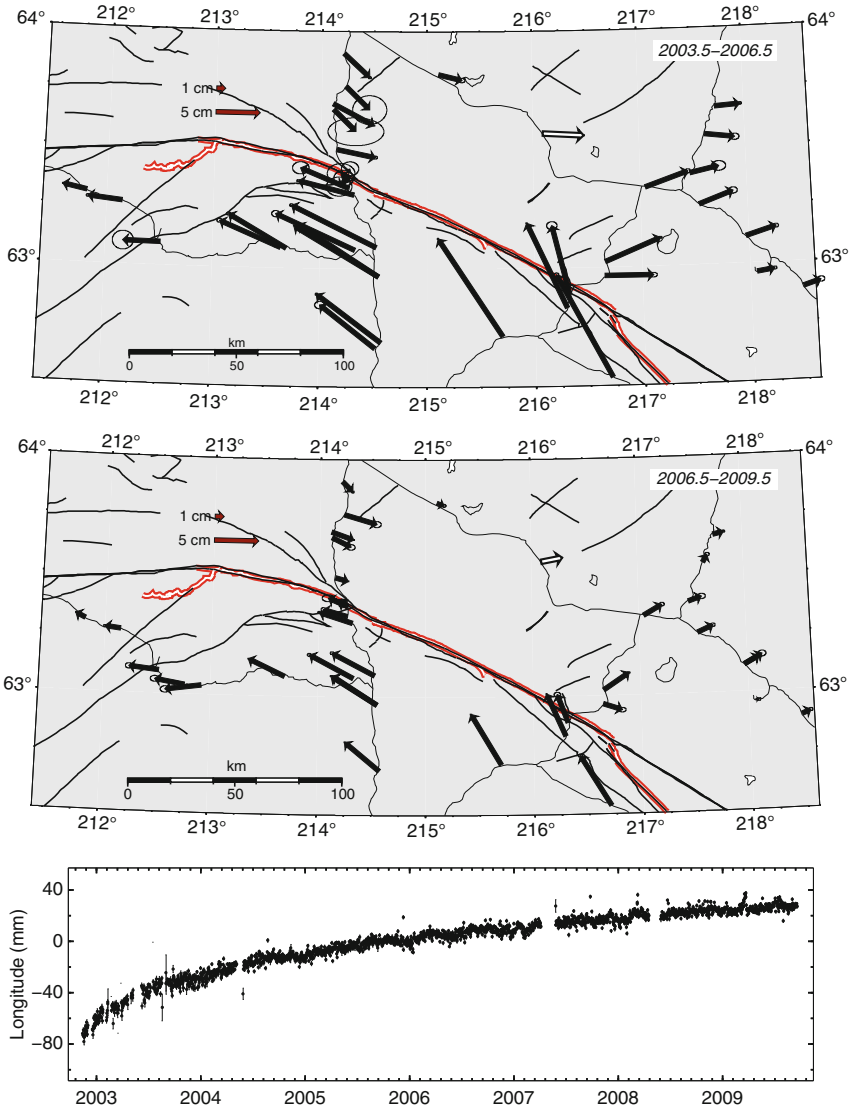
and upper mantle (Bürgmann and Dresen, 2008). Postseismic deformation provides an opportunity to study the dynamics of tectonic deformation and the rheology of the crust and upper mantle, whereas the linear motion in time observed for most of the time between earthquakes allows only a kinematic description. Postseismic deformation can be large and long lasting, although there is great variation from earthquake to earthquake.

Postseismic deformation is thought to result from a superposition of three main physical mechanisms, which cause deformation on different spatial and temporal scales: viscoelastic relaxation of the mantle and possibly lower crust, afterslip on the very shallow or deep parts of the fault zone, and poroelastic relaxation. Poroelastic relaxation, which is deformation driven by fluid flow that relieves pressure differences caused by the earthquake, is generally considered to be important only very close to the fault and near geometric complexities such as fault stepovers and bends, where the dilatational strain is the largest (and thus fluid flow changes will be most significant). Shallow afterslip is a frictional process, first observed after the 1966 Parkfield earthquakes (Smith and Wyss, 1968), and observed after many subsequent earthquakes, including subduction zone events (Hsu et al., 2006; Kreemer et al., 2006). Depending on the rheology of deep fault zones, deep afterslip may be a frictional process, a ductile shear process, or some combination of the two. Deep afterslip is generally found immediately downdip of the coseismic rupture zone. Close to the fault, postseismic displacements are usually significantly smaller than coseismic displacements, reflecting the smaller amount of slip involved and the generally deeper deformation source. The 2004 Parkfield earthquake was an exception to this rule, because coseismic slip at shallow depth was small, while shallow postseismic afterslip was large (Johanson et al., 2006). However, in the far field, the longer spatial wavelength of the postseismic signal that results from its deeper or spatially distributed source means that far-field postseismic displacements can be a significant fraction of, or even comparable to, the coseismic displacements.

The various mechanisms of postseismic deformation can produce similar deformation at the surface, especially in the horizontal components, and the predicted displacements from different mechanisms can be similar enough that controversy over postseismic deformation models can linger for years, even for well-studied earthquakes. The ambiguity between postseismic source models is enhanced because GPS data are always irregularly sampled in space, while InSAR data may be continuous in space, but do not measure 3D displacements. There is not yet a consensus regarding the physical model for afterslip that would allow accurate prediction of afterslip given only the coseismic slip model; this means that afterslip is usually estimated in a manner similar to the coseismic slip model, whereas viscoelastic relaxation is predicted from the coseismic slip distribution and an (estimated or inferred) earth model that describes the rheology. Unfortunately, the appropriate rheological model varies from place to place, and parameters such as mantle viscosity are both heterogeneous and poorly constrained in value (Bürgmann and Dresen, 2008). As a result of these factors, postseismic studies often make assumptions that certain displacements result from only the effect of one mechanism, or attempt to use the spatial or temporal character of the observed displacements to constrain the mechanisms separately.

The 2002 Denali fault earthquake provides a useful example. Ten continuous GPS sites were set up within a few weeks of the earthquake and nearly 100 sites were surveyed repeatedly over the next several years, providing a rich data set in space and time (Figure 4). Both the spatial and temporal variations in postseismic displacements require more than one mechanism to explain them. All sites show an initially rapid rate of deformation that decayed over time. In general, site velocities averaged over the first 1–2 years were ~20 times faster than the pre-earthquake rates. Even several years after the earthquake, average velocities remain several times higher than the pre-earthquake rates at many sites. No single simple relaxation function of time, such as an exponential or logarithmic relaxation, can explain the temporal variations of the time series to within the scatter in the measurements. However, a combination of relaxation functions can explain the temporal decay of the displacements quite well, as long as the model fit to the time series has both a short relaxation component (characteristic decay time of ~months) and a long relaxation component (characteristic decay time of a few years). In addition, several sites close to the fault (within a few tens of km) exhibited an initial phase of extremely rapid displacements, up to several millimeters per day, which decayed over a period of a few weeks with a characteristic decay time on the order of days. This component was not sampled uniformly enough to develop models for it.

Postseismic models for the 2002 Denali fault earthquake have called upon a variety of mechanisms to explain the deformation. Pollitz (2005) used data from a limited set of sites over the first 16 months after the earthquake and a model that assumed all deformation resulted from viscoelastic of a material obeying a Burger's body transient viscosity. Freed et al. (2006a) showed that no single mechanism could explain the observed displacements, and proposed a model based on afterslip within the lithosphere (approximated as a narrow low viscosity shear zone) and viscoelastic relaxation in the asthenosphere. Freed et al. (2006b) analyzed the time dependence at far-field sites and concluded that a model with a nonlinear (power-law) viscosity in the mantle, consistent with lab measurements of the rheology of olivine, explained these data better than a model with a linear Newtonian viscosity. Johnson et al. (2009) estimated a combined afterslip and viscoelastic model, using an assumed frictional model. However, none of these models have been very successful in predicting displacements into the future, beyond the time span of the data



GPS, Tectonic Geodesy, Figure 4 Postseismic displacements following the November 2002 Denali fault earthquake. Displacements are shown for the time periods 2003.5–2006.5 (*top panel*) and 2006.5–2009.5 (*middle panel*). The pre-earthquake motions have been subtracted from both figures, so the displacements represent the transient component of motion over those time intervals. Pre-earthquake velocities relative to North America were less than 1 cm/year over most of the area shown, and postseismic velocities are several times faster, and diminish in rate over time (*bottom panel*). The time series of eastward motion of the site JANL (*white vector*) relative to North America illustrates the nonlinear character of postseismic deformation. Although the time series appears to be nearly linear several years after the earthquake, there is actually significant curvature of the time series over the entire time period.

used to constrain them (Freymueller et al., 2009). These model shortcomings may result from deficiencies in the estimated coseismic slip model, from a failure to separate the different mechanisms correctly, or from additional processes not considered in the models, possibly including changes in the slip on the subduction thrust in southern Alaska caused by the Denali fault earthquake stress changes (Johnson et al., 2009).

The general pattern of postseismic deformation seen in the 2002 Denali fault earthquake has been observed in many other cases, from moderately large crustal events to great subduction zone earthquakes. However, individual earthquakes have displayed a puzzling variety of postseismic responses. For some large subduction earthquakes, postseismic transients seem to have decayed away after only a few years, while in other cases small transient signals have lingered for decades (Melbourne et al., 2002). After the 1995 M_W 8.1 Antofagasta (Chile) earthquake, there was significant afterslip for a few years but apparently little viscoelastic relaxation of the mantle

because postseismic deformation rates returned to pre-earthquake rates within a few years (Melbourne et al., 2002). The 2003 Hokkaido Tokachi-oki earthquake caused afterslip downdip of the coseismic rupture, which also extended at depth for a significant lateral distance north of the rupture zone (Miyazaki et al., 2004b). In the 2006–2007 Kuril Island doublet, a subduction thrust earthquake followed by a similar-sized outer rise event, afterslip appears to have been limited to the first several months after the thrust earthquake (Steblov et al., 2008; Kogan et al., submitted manuscript). The cause of this wide range of behavior is not understood at this time. For the largest earthquakes, significant postseismic deformation clearly can last for decades (Cohen and Freymueller, 2004; Wang, 2007; Suito and Freymueller, 2009). Postseismic deformation from the great 1960 Chile and 1964 Alaska earthquakes remains very large (rates of 1–2 cm/year), decades after these large events (Wang, 2007; Suito and Freymueller, 2009).

Because of the long duration of postseismic transient deformation and the short length of precise geodetic measurements in most places, it is not clear whether postseismic transients decay away to a relatively steady background deformation rate, or whether deformation rates are never truly steady in time. Answering this question is very important, because the answer will reveal a great deal about the dynamics of earth deformation. Decay to a steady background rate suggests that deformation in the mantle and lower crust associated with fault motion is localized due to lateral variations in properties or nonlinear viscoelasticity. The long-term fault motion rate and background deformation rate would represent a balance between tectonic forces and the effective viscosity of the weak zone beneath the fault. On the other hand, a uniform linear viscoelastic mantle requires that shear in the mantle be distributed broadly rather than localized beneath faults. In this case, deformation at the surface would never be steady in time except when the earthquake recurrence time and mantle relaxation time are roughly equal, and transient deformations would always extend a great distance away from every active fault, even a long time after the last major earthquake (Savage and Prescott, 1978). The apparently linear "interseismic" phase would be essentially an illusion caused by our short measurement history. However, our ability to measure steady rigid plate motions that are very close to long-term geological plate motion estimates argues in favor of some degree of localization of strain in the mantle beneath faults, as does the observation that interseismic motions are localized around faults like the North Anatolian Fault in Turkey regardless of the time since the last earthquake (Hearn et al., 2009). This is also supported by the frequent consistency of geologic estimates of fault slip rates and geodetic estimates based on simple linear models. However, this question is not yet settled and probably will remain a focus of research for some time. Although some progress has been made in the theoretical models, a convincing and complete model for the earthquake cycle that incorporates postseismic and interseismic deformation and agrees with data has not yet been developed.

Transient fault movements and more complex models of the earthquake cycle

Faults also exhibit a range of transient slip behavior, possibly similar to postseismic afterslip, but not necessarily caused by identifiable stress changes such as those caused by large earthquakes. The most notable examples are slow slip events (SSE) on subduction thrusts, also called episodic tremor and slip (ETS) events because the slip is usually accompanied by the emission of seismic tremor (Rogers and Dragert, 2003). There are also several known examples of triggered creep or changes in creep rate, sometimes caused by nearby earthquakes on a different fault. For example, the 1989 Loma Prieta earthquake in California triggered shallow creep on a subparallel thrust fault system (Bürgmann et al., 1997) and caused changes in creep rate on sections of the nearby Hayward and Calaveras faults (Lienkaemper et al., 1997). In the latter case, the earthquake caused a reduction of stress on these nearby creeping faults and some sections responded by ceasing to creep for a period of time, and later restarted, while other sections continued to creep at a reduced rate. These transient fault movements are further evidence that the steady, elastic earthquake cycle model is a significant simplification of reality: even between earthquakes there can be small variations in deformation.

SSE and ETS events were first observed in Japan (Heki et al., 1997) and at the Cascadia subduction zone (Dragert et al., 2001) in North America. A large number of these events have now been identified in Japan, Cascadia, Alaska, Mexico, the Ryukyu arc, and New Zealand, and in most cases, the slip events have been accompanied by emission of seismic tremor (Schwartz and Rokosky, 2007). These events come in a range of sizes, many with slip and area equivalent to magnitude 6–7.5 earthquakes. Small tremor bursts are also observed, as are smaller slip events too small to detect with GPS (but observed on strainmeters or tiltmeters). These mostly aseismic slip events, including small bursts of tremor that do not have resolvable geodetic signals, appear to follow a self-similar (power-law) scaling relation somewhat like regular earthquakes, but with a different power-law exponent (Ide et al., 2007).

The recognition of these transient events, along with the development of models that explained seasonal loading variations (Heki, 2001) has resulted in a transformation of thought about the steadiness in time of site motions. Prior to the recognition of SSEs or ETS events, the typical assumption in tectonic geodesy was that all deviations from a linear trend represented noise in the data, except when there was a known cause of displacement such as an earthquake or postseismic deformation. Today, deviations from a linear trend are likely to be considered signal, although it remains uncertain how to reliably distinguish between transient tectonic signal and

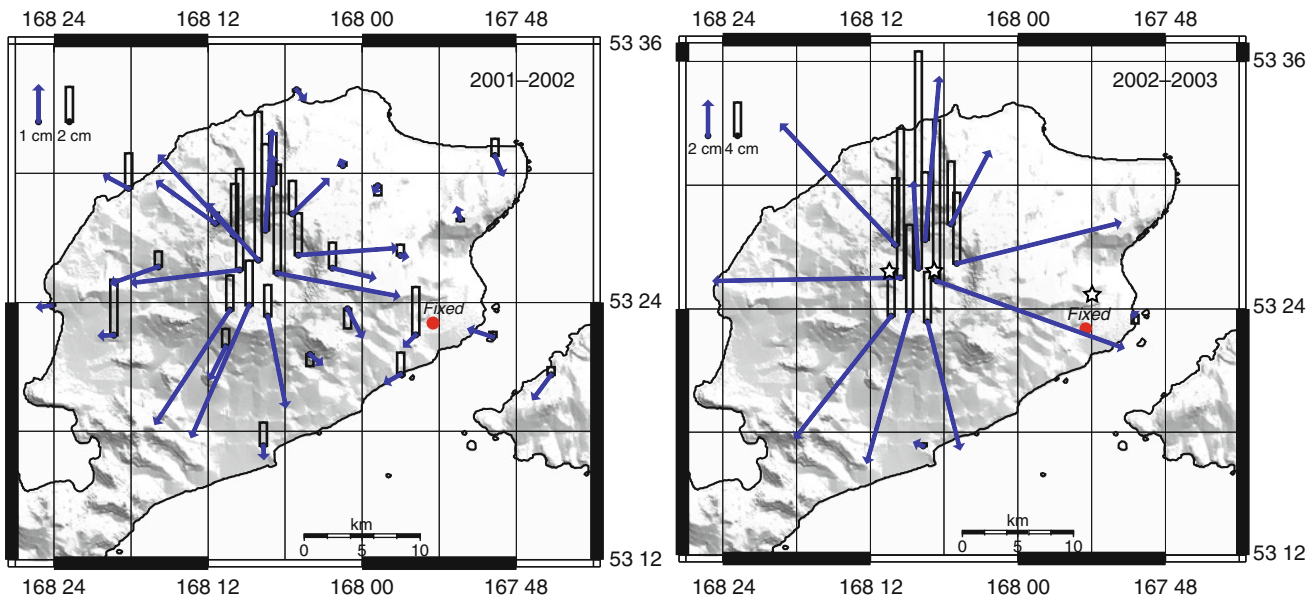
noise. The most successful attempts to do so rely on the spatial coherence of deformation signals, especially when the deformation can be linked to a source model, such as slip on a fault of known geometry. However, noise in geodetic time series can be spatially correlated as well, and the separation of small transient signals from noise remains an area of active research. A particular area of emphasis today is in the automatic detection of transient events from observed time series.

Volcano deformation

Unlike tectonic deformation, which results mainly from slip on faults (shear sources), volcanic deformation results mainly from volumetric or dilatational sources, such as the expansion or contraction of a dike or sill or other magma body. Injection of magma into the subsurface through intrusion or removal of magma from the subsurface as it is erupted cause pressure changes that drive volcanic deformation, and result in deformation at the surface due to the elasticity of the earth. Volcanic deformation commonly results in horizontal displacements that are oriented radially away from or toward the volcano, and can produce significant subsidence or uplift (Figure 5). Some volcanoes have produced very large deformation signals associated with the intrusion of magma before eruption, while others have produced only small deformation signals even as the result of large eruptions. The heterogeneity of behavior may be a result in part of magma properties that vary with magma composition and gas content, or because of variations in eruption style and triggering mechanisms.

The simplest volcanic source model is a spherical point pressure source, usually called the Mogi model after the Japanese scientist who first applied it in the 1950s (Mogi, 1958). This model describes the deformation at the surface caused by a pressure change in a small spherical body in an elastic half-space; the point source approximation actually holds even for relatively large bodies. The Mogi model has provided a surprisingly good fit to observed displacements from volcanic sources. Models for finite spherical or elliptical sources have also been developed. The finite spherical source is rarely used because the size of the source usually cannot be constrained by data, but elliptical sources produce a different pattern of displacements due to the difference from spherical geometry. Dikes and sills, planar bodies of magma that are associated with the opening and filling of planar cracks, can also produce significant deformation. These sources can be modeled using the same dislocation theory that is used for shear faulting, as the most commonly used models include the equations for opening-mode deformation.

Volcanic deformation is inherently time-dependent, and volcanoes can display episodes of both inflation (increasing pressure) and deflation (decreasing pressure). Deflationary episodes probably result from a pressure drop caused by the leaking out of exsolved gases or from the depressurization of a hydrothermal system. During eruption, most volcanoes show deformation associated with decreasing pressure, although in a few cases there have



GPS, Tectonic Geodesy, Figure 5 Displacements from Okmok volcano, eastern Aleutian arc, Alaska, showing significant volcanic deformation. The left panel shows displacements from 2001–2002, and the right panel from 2002–2003. The scale for vertical motion is different from that of the horizontal, and there is a factor of two differences in scale between the two plots, a result of the much larger deformation in 2002–2003. The blue arrows are horizontal displacements, and the white bars show vertical displacements. Data are from Miyagi et al. (2004). The deformation results from the pressurization of a magma body located about 3 km below the surface.

been observations of simultaneous deflation at depth and inflation close to the surface. One example of this is the 2000 eruption of Usu volcano in Japan, for which deformation and gravity change data were fit best by a model with a deflation source at ~4 km depth, connected to the surface by an intruded fissure (Jousset et al., 2003). If the magma is assumed to be incompressible, which is a good assumption for a magma that contains no exsolved gas bubbles, then the volume of magma added or removed to the subsurface is proportional to the pressure change in the reservoir or the opening or closing of a dike or sill, and thus the deformation magnitude is proportional to magma volume. If the magma is compressible, then the deformation magnitude will be a function of the magma volume change, magma compressibility, and the elastic properties of the surrounding rock.

Unlike the tectonic problem, it seems that a simple "eruption cycle" model does not apply to volcanoes. The idea is appealing: magma would accumulate in the subsurface until the pressure reaches some critical level, which causes failure of the overlying rock, opening of a conduit to the surface, and then eruption. However, observations show that the magma supply is not uniform in time (Dvorak and Dzurisin, 1993), while tectonic loading is, and successive eruptions at the same volcano can be quite different in volume and style. Even at volcanoes that erupt frequently, there does not appear to be a clear relationship between the volume of magma that accumulates between eruptions and the volume that erupts. This reflects the fact that the crust beneath active volcanoes may be occupied by a number of bodies of crystal mush or partially solidified magma, and also the importance of gases and volatiles in triggering magma rise and eruption. A small intrusion of fresh, hot magma into a cooler body can trigger exsolution of dissolved volatiles and remobilization of the magma, causing a pressure increase that may be capable of triggering an eruption.

Similarly, although rapid deformation at a volcano is an indication that the volcanic system is active, it is not clear whether episodes of deformation can reliably be treated as short-term eruption precursors. Some large calderas, such as Campi Flegrei in Italy, Yellowstone and Long Valley in the USA, and Rabaul in Papua New Guinea, have experienced episodes of substantial inflation, deflation or both over a period of many years (Dvorak and Mastrolorenzo, 1991; Dvorak and Dzurisin, 1997). Some of this deformation may result from pressure changes in hydrothermal systems associated with the volcano, rather than from changes in the magma reservoirs themselves. Okmok volcano in the eastern Aleutian arc (USA) experienced substantial accumulation of new magma at shallow depth (Figure 5) following its 1997 eruption, causing as much as several decimeters of displacement at the surface (Lu et al., 2005; Fournier et al., 2009), but its 2008 eruption was preceded by a period of ~3 years of relative quiescence. Mt. Etna in Italy is another volcano that has shown significant magma accumulation between eruptions, although not necessarily as a short-term precursor to eruption. Other recent eruptions had only subtle or nonexistent deformation prior to the eruption, including the eruptions of St. Helens in 2004, Augustine in 2006, and Redoubt in 2009 in the USA (Cervelli et al., 2006; Lisowski et al., 2008). However, significant precursory deformation very close to the eruptive vent certainly must happen. A notable example came from a continuous GPS site that was installed on Stromboli volcano (Italy), which showed dramatic displacements in the minutes before it was destroyed by an eruption from a new vent located a very short distance away (Mattia et al., 2004). Such deformation, associated with the final rise of magma to the surface, is very localized.

On the other hand, Kilauea volcano in Hawaii shows distinctive and reliable patterns of deformation prior to almost every significant eruption episode. Most magma erupted from Kilauea in recent decades has been emitted from vents on one of its rift zones, and since 1983 activity has been centered on the flank vent called Pu'u 'O'o. Eruptions of Kilauea are preceded by inflation at the volcano's summit, which then deflates as a pulse of magma that is intruded into the rift zone and proceeds down-rift until it reaches the flank vent (Dvorak and Dzurisin, 1993; Dvorak and Dzurisin, 1997). There is some indication that basaltic volcanoes like Kilauea and Okmok may produce larger and more predictable deformation patterns than more silicic volcanoes like the others mentioned in this section.

Glacial isostatic adjustment and sea level studies

The Earth also deforms in response to changes in surface loads, for example, changes in the mass of glaciers, ice fields, and ice sheets on the surface. The best-known example of this deformation is the dramatic uplift that has occurred in the regions formerly covered by the continental ice sheets in North America and Europe. The ongoing uplift in these regions, due to glacial isostatic adjustment (GIA), is dominated by the viscoelastic response of the mantle to the unloading as the ice sheets collapsed after the last glacial maximum (LGM), about 23,000 years ago. While the ice sheets were growing, the mass of the ice depressed the surface and induced stresses that caused the mantle to flow out of the way. Upon collapse of the ice sheets the process was reversed and mantle is still flowing back at present. In addition to the long-lasting effects of the great ice sheets, deformation due to present-day ice mass changes can be observed as well, and can cause even more rapid deformation even though the ice mass changes are much smaller.

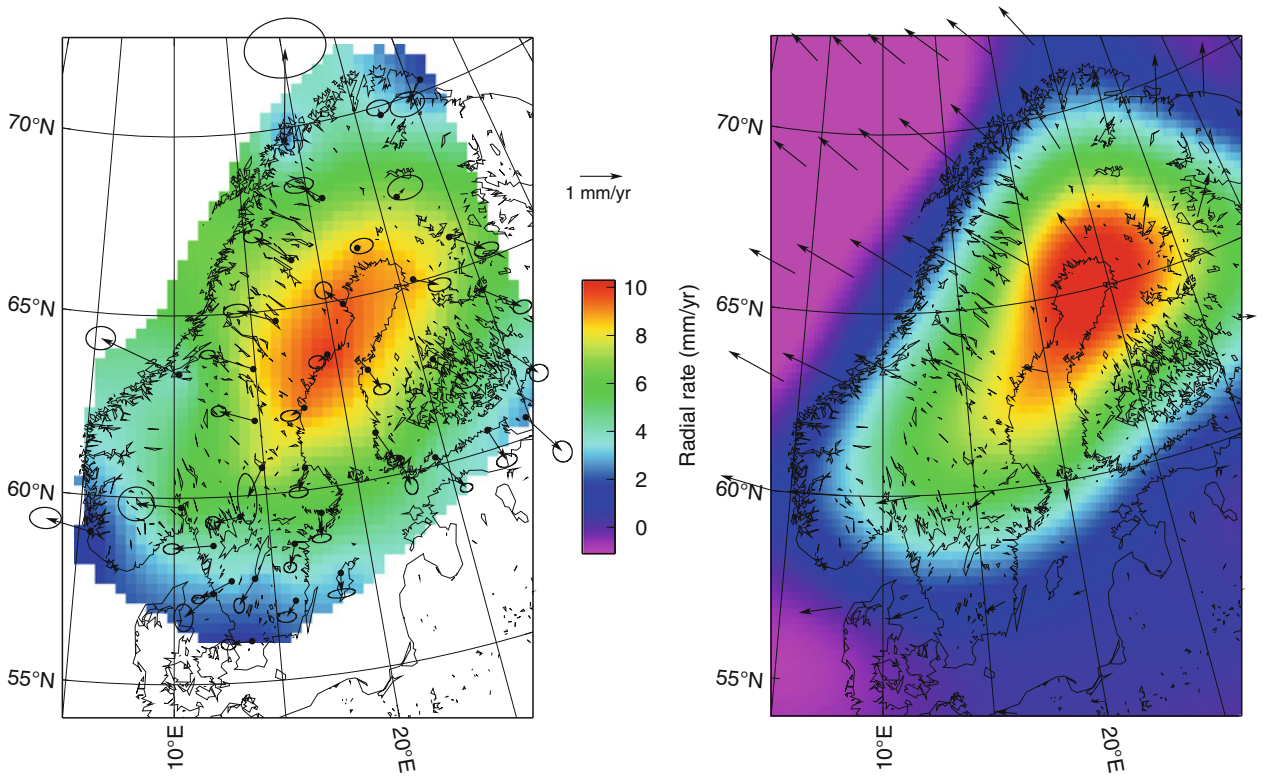
Displacements from load changes are predominantly vertical. Horizontal displacements can be significant relative to geodetic measurement errors, but are usually several times smaller than the vertical displacements. Horizontal displacements are more sensitive than the vertical displacements to assumptions made in the computations (compressibility vs incompressibility, radially symmetric earth structure vs lateral variations, etc.). Thus,

while model predictions for the vertical motions are in general agreement, the best models for the horizontal motions remain a matter of considerable debate.

Conceptually, it is useful to break up the deformation due to load changes into two separate components. The first component is the elastic response, which is instantaneous upon a change in load. After the elastic response at the time of the load change, stresses in the viscoelastic mantle relax with time, causing subsequent displacements until the final relaxed response is reached. The theory for computing the elastic response of the earth dates back to Love in the early twentieth century; Farrell (1972) used this theory to compute the response of a realistic earth model to both point loads and disk loads. Farrell's results are commonly used to compute the elastic response. The viscoelastic problem results in time-dependent deformation, which can be computed in a variety of ways, either in terms of viscoelastic Love numbers or viscoelastic Green's functions for point or finite loads. However it is done, in the viscoelastic problem the time-dependent displacements or displacement rates are computed and include both the elastic and viscous components.

The BIFROST project in Scandinavia was initiated in the early 1990s to study GIA resulting from the post-LGM collapse of the Fennoscandian ice sheet (Johansson et al., 2002; Milne et al., 2004). This project demonstrated the capability of GPS to measure accurate vertical motions. A network of continuous GPS sites was established that encompassed the entire region surrounding the Baltic Sea. The observations revealed the expected pattern of deformation, with the largest uplift rates (12–13 mm/year) located on the northwest margin of the Baltic Sea where the LGM ice sheet had its maximum thickness (Figure 6). Overall, the contours of constant uplift rate are elliptical in shape. Horizontal motions are no larger than 2–3 mm/year and generally are oriented away from the center of the former ice sheet. The GPS observations provided important new constraints on the viscosity structure of the mantle (Milne et al., 2001).

There have been fewer geodetic studies of GIA from the Laurentide ice sheet in North America. The number and spatial density of continuous GPS sites in that part of North America is much smaller than in Scandinavia. However, reliable estimates of the vertical GIA signal have been determined from repeat surveys of sites of the Canadian Base Network and from the available continuous GPS sites (Sella et al., 2007). Peak uplift rates are 10 mm/year near Hudson's Bay and significant uplift is



GPS, Tectonic Geodesy, Figure 6 Observations and model predictions for GIA in the Fennoscandian region, after Milne et al. (2001) using the BIFROST solution of Lidberg et al. (2007) (*left panel*). Horizontal velocities relative to the Eurasian plate (arrows, with 95% confidence ellipses), and vertical velocities (background color) (*right panel*). Model predictions based on ICE-3G with lithosphere thickness 120 km, upper mantle viscosity 5e20 Pa s, and lower mantle 8e21 Pa s. This figure was provided by James L. Davis, Lamont-Doherty Earth Observatory, Columbia University.

observed over most of northern and northeastern North America. The horizontal motions are much smaller, generally at the 1–2 mm/year level, and include motions directed both toward and away from the former ice sheet. The scale of the GIA deformation in North America poses an interesting problem: the estimates of horizontal motion relative to North America depend on the estimated motion of the North American plate in ITRF, but it is difficult to measure the motion of North America without having that estimate biased by horizontal GIA. As a result, independent estimates of the motion of the North American plate tend to differ by more than would be expected from data noise alone.

The most rapid uplift rates from GIA come not from the areas formerly covered by the great ice sheets, but from regional ice masses (glaciers and ice fields) that recently have been losing mass. Ice mass loss in Alaska and surrounding parts of Canada in North America and in Patagonia in South America has been very large since the end of the Little Ice Age (LIA) ~200 years ago, large enough to account for 10–20% of twentieth century global sea level rise (Dyurgerov and Meier, 2005; Larsen et al., 2007; Lemke et al., 2007). In both of these regions, measured uplift rates exceed 30 mm/year at their peak, and significant uplift has been observed in a broad area surrounding the regional ice fields. In Alaska, the observed uplift depends significantly on both the ongoing regional ice mass loss and the nineteenth century deglaciation of Glacier Bay (Larsen et al., 2005). The former source has important elastic and viscoelastic components, while the latter is purely a viscoelastic response. Overall, depending on the place, the contribution to the uplift rate from the present ongoing mass loss ranges from ~1/3 to ~1/2 of the entire signal, and the uplift rate from present mass loss can be as large as ~12 mm/year (Sato et al., in press).

Mass loss in the Greenland and Antarctic ice sheets has begun or accelerated within the last decade and the changes in mass loss rate are reflected in changes in the vertical motions. Southeast Greenland made the transition from no or minimal uplift to significant uplift in the early 2000s, while the onset of uplift in northwest Greenland began a few years later (Khan et al., 2008; Khan et al., 2010). The uplift history measured by GPS is in agreement with the mass loss history estimated by gravity observations of the GRACE satellite mission and is corroborated by direct observation and measurement of glacier mass balance in the area. The prospect of observing changes in Antarctic and Greenland ice sheets has prompted the installation of new networks of continuous GPS sites, which have begun to produce data that will allow researchers to "weigh" the ice sheets as they change over time.

As glaciers, icefields, and ice sheets lose mass, most of the water ends up in the ocean where it raises sea level. Tide gauges measure relative sea level, which is the level of the ocean surface relative to the land surface. Estimates of global sea level rise depend on averaging over a global set of tide gauges where uplift or subsidence of the land is negligible, but more data could be used to analyze sea level variations if land uplift rates at tide gauges were precisely known. Wöppelman et al. (2009) used GPS measurements at a globally distributed set of tide gauges to correct for vertical motions of the land and estimated an average global sea level rise of 1.6 ± 0.2 mm/year, assuming that the vertical motions observed by GPS over the last decade are representative of the last 100 years. This estimate agrees with previous estimates for twentieth century sea level rise.

Sea level does not rise uniformly around the globe when there is a significant redistribution of mass resulting from the melting of ice. Spatial variations in sea level rise result from changes in the geoid caused by the mass redistribution; sea level follows the geoid, so if the geoid height is increasing over a section of ocean, then local sea level would rise in a geocentric reference frame (Mitrovica et al., 2001; Tamisiea et al., 2001). The difference between relative sea level variations measured by tide gauges and vertical motions measured by GPS is a combination of vertical motion of the land, sea level rise due to the addition of water to the oceans, ocean thermal expansion, and the local geoid change. Geoid rate changes at the multiple mm/year level are predicted for areas of significant regional mass loss, such as Alaska and Patagonia, and significant geoid height trends have been measured using GRACE (Tamisiea et al., 2001; Tamisiea et al., 2005). Much larger variations in the geoid would appear in the future if the Greenland and Antarctic ice sheets melt significantly (Mitrovica et al., 2009). Future studies will integrate data and models for all of these effects to provide a complete and self-consistent picture of sea level, gravity change, and height change around the world's coastlines.

Summary

High-precision geodesy is an important tool for the study of active tectonics and other causes of crustal deformation. Motion and deformation of points on the surface of the earth result from a combination of tectonic plate motions, elastic deformation from the slip deficit associated with faults that are locked by friction, and coseismic and postseismic displacements. Deformation caused by pressure changes within volcanic systems is important near volcanoes, and in some areas there is additional deformation caused by past and present surface load changes, most importantly from the melting of ice in glaciers, ice fields, and ice sheets. Deformation due to several or all of these sources may be superposed in some locations, such as southern Alaska and Patagonia. Surface displacements can be related to the underlying sources using elastic or viscoelastic deformation models, the most important of which are based on dislocation theory.

Bibliography

Altamimi, Z., Collilieux, X., Legrand, J., Garayt, B., and Boucher, C., 2007. ITRF2005: a new release of International Terrestrial Reference Frame based on time series of station positions and Earth Orientation Parameters. *Journal of Geophysical Research*, **112**, B004949, doi:10.1029/2007JB004949.

Argus, D. F., and Heflin, M. B., 1995. Plate motion and crustal deformation estimated with geodetic data from the Global Positioning System. *Geophysical Research Letters*, **22**, 1973–1976.

Argus, D. F., Gordon, R. G., Heflin, M. B., Ma, C., Eanes, R. J., Willis, P., Peltier, W. R., and Owen, S. E., 2010. The angular velocities of the plates and the velocity of Earth's centre from space geodesy. *Geophysical Journal International*, **181**, 1–48, doi:10.1111/j.1365-246X.2009.04463.x.

Banerjee, P., Pollitz, F. F., and Bürgmann, R., 2005. The size and duration of the Sumatra-Andaman earthquake from far-field static offsets. *Science*, **308**, 1769–1772.

Banerjee, P., Bürgmann, R., Nagarajan, B., and Apel, E., 2008. Intraplate deformation of the Indian subcontinent. *Geophysical Research Letters*, **35**, L18301, doi:10.1029/2008GL035468.

Blewitt, G., Heflin, M., Hurst, K., Jefferson, D., Webb, F., and Zumberge, J., 1993. Absolute Far-Field Displacements from the 28 June 1992 Landers earthquake sequence. *Nature*, **361**, 340–342.

Bock, Y., Agnew, D. C., Fang, P., Genrich, J. F., Hager, B. H., Herring, T. A., Hudnut, K. W., King, R. W., Larsen, S., Minister, J. B., Stark, K., Wdowinski, S., and Wyatt, F. K., 1993. Detection of crustal deformation from the Landers earthquake using continuous geodetic measurements. *Nature*, **361**, 337–340.

Bonafede, M., Strehlau, J., and Ritsema, A. R., 1992. Geophysical and structural aspects of fault mechanics – a brief historical overview. *Terra Nova*, **4**, 458–463.

Bürgmann, R., and Dresen, G., 2008. Rheology of the lower crust and upper mantle: evidence from rock mechanics, geodesy, and field observations. *Annual Review of Earth and Planetary Sciences*, **36**, 531–567, doi:10.1146/annurev.earth.36.031207.124326.

Bürgmann, R., Segall, P., Lisowski, M., and Svarc, J. L., 1997. Postseismic strain following the 1989 Loma Prieta earthquake from GPS and leveling measurements. *Journal of Geophysical Research*, **102**, 4933–4955.

Calais, E., Han, J. Y., DeMets, C., and Nocquet, J. M., 2006. Deformation of the North American plate interior from a decade of continuous GPS measurements. *Journal of Geophysical Research*, **111**, B06402, doi:10.1029/2005JB004253.

Cervelli, P., Fournier, T., Freymueller, J. T., and Power, J., 2006. Ground deformation associated with the precursory unrest and early phases of the January 2006 eruption of Augustine Volcano, Alaska. *Geophysical Research Letters*, **33**, L18304, doi:10.1029/2006GL027219.

Chlieh, M., Avouac, J. P., Sieh, K., Natawidjaja, D. H., and Galetzka, J., 2008. Heterogeneous coupling of the Sumatran megathrust constrained by geodetic and paleogeodetic measurements. *Journal of Geophysical Research*, **113**, B05305, doi:10.1029/2007JB004981.

Cohee, B. P., and Beroza, G. C., 1994. Slip distribution of the 1992 Landers earthquake and its implication for earthquake source mechanics. *Bulletin. Seismological Society of America*, **84**, 646–659.

Cohen, S. C., and Freymueller, J. T., 2004. Crustal deformation in Southcentral Alaska: the 1964 Prince William sound earthquake subduction zone. *Advances in Geophysics*, **47**, 1–63.

Cross, R. S., and Freymueller, J. T., 2008. Evidence for and implications of a Bering plate based on geodetic measurements from the Aleutians and western Alaska. *Journal of Geophysical Research*, **113**, B07405, doi:10.1029/2007JB005136.

DeMets, C., Gordon, R., Argus, D., and Stein, S., 1990. Current plate motions. *Geophysical Journal International*, **101**, 425–478.

DeMets, C., Gordon, R. G., Argus, D. F., and Stein, S., 1994. Effect of recent revisions to the geomagnetic reversal time-scale on estimates of current plate motions. *Geophysical Research Letters*, **21**, 2191–2194.

DeMets, C., Gordon, R. G., and Argus, D. F., 2010. Geologically current plate motions. *Geophysical Journal International*, **181**, 1–80, doi:10.1111/j.1365-246X.2009.04491.x.

Dolan, J. F., Bowman, D. D., and Sammis, C. G., 2007. Long range and long-term fault interactions in Southern California. *Geology*, **35**, 855–858, doi:10.1130/G23789A.1.

Dragert, H., Wang, K., and James, T. S., 2001. A silent slip even on the deeper Cascadia subduction interface. *Science*, **292**, 1525–1528, doi:10.1126/science.1060152.

Dvorak, J. J., and Dzurisin, D., 1993. Variations in magma supply rate at Kilauea Volcano, Hawaii. *Journal of Geophysical Research*, **98**, 22255–22268.

Dvorak, J. J., and Dzurisin, D., 1997. Volcano geodesy: the search for magma reservoirs and the formation of eruptive vents. *Reviews of Geophysics*, **35**, 343–384.

Dvorak, J. J., and Mastrolorenzo, G., 1991. The mechanisms of recent vertical crustal movements in Campi Flegrei caldera, southern Italy. *Geological Society of America*, Special Paper **263**, 47 pp.

Dyurgerov, M. B., and Meier, M. F., 2005. Glaciers and the changing Earth system: a 2004 snapshot. INSTAAR Occasional Paper 58, Boulder: Institute of Arctic and Alpine Research, University of Colorado

Farrell, W. E., 1972. Deformation of the Earth by surface loads. *Reviews of Geophysics*, **10**, 761–797.

Fournier, T., Freymueller, J. T., and Cervelli, P., 2009. Tracking magma volume recovery at Okmok Volcano using GPS and an unscented Kalman filter. *Journal of Geophysical Research*, **114**, B02405, doi:10.1029/2008JB005837.

Freed, A. M., Bürgmann, R., Calais, E., Freymueller, J., and Hreinsdóttir, S., 2006a. Implications of deformation following the 2002 Denali, Alaska earthquake for postseismic relaxation processes and lithospheric rheology. *Journal of Geophysical Research*, **111**, B01401, doi:10.1029/2005JB003894.

Freed, A., Burgmann, R., Calais, E., and Freymueller, J., 2006b. Stress-dependent power-law flow in the upper mantle following the 2002 Denali, Alaska, earthquake. *Earth and Planetary Science Letters*, **252**, 481–489.

Freymueller, J. T., King, N. E., and Segall, P., 1994. The co-seismic slip distribution of the Landers earthquake. *Bulletin of the Seismological Society of America*, **84**, 646–659.

Freymueller, J. T., Woodard, H., Cohen, S., Cross, R., Elliott, J., Larsen, C., Hreinsdóttir, S., and Zweck, C., 2008. Active deformation processes in Alaska, based on 15 years of GPS measurements. In Freymueller, J. T., Haeussler, P. J., Wesson, R., and Ekstrom, G. (eds.), *Active Tectonics and Seismic Potential of Alaska*. Washington: American Geophysical Union. AGU Geophysical Monograph, Vol. 179, pp. 1–42.

Freymueller, J. T., Freed, A. M., Johnson, K. M., Bürgmann, R., Calais, E., Pollitz, F. F., and Biggs, J., 2009. Denali fault earthquake postseismic deformation models. *Eos Transactions of the AGU*, 90, Fall meeting supplement, Abstract G34A-05.

Gan, W., Zhang, P., Shen, Z.-K., Niu, Z., Wang, M., Wan, Y., Zhou, D., and Cheng, J., 2007. Present-day crustal motion within the Tibetan Plateau inferred from GPS measurements. *Journal of Geophysical Research*, **112**, B08416, doi:10.1029/2005JB004120.

Gendt, G., Montag, H., and Dick, G., 1993. Plate kinematics in a global and European scale by LAGEOS laser ranging data from 1983 to 1990. In Smith, D. E., and Turcotte, D. L. (eds.), *Contributions of Space Geodesy to Geodynamics: Crustal Dynamics*. Washington: American Geophysical Union. AGU Geophysical Monograph, Vol. 23, pp. 311–330.

Gilbert, G. K., 1884. A theory of the earthquakes of the Great Basin with a practical application. *American Journal of Science*, **27**, 49–53. reprinted from the *Salt lake Tribune* of 20 Sept. 1883.

Gordon, D., Ma, C., and Ryan, J. W., 1993. Results from the CDP mobile VLBI program in the western United States. In Smith,

D. E., and Turcotte, D. L. (eds.), *Contributions of Space Geodesy to Geodynamics: Crustal Dynamics*. Washington: American Geophysical Union. AGU Geophysical Monograph, Vol. 23, pp. 131–138.

Hearn, E. H., McClusky, S., Ergintav, S., and Reilinger, R. E., 2009. Izmit earthquake postseismic deformation and dynamics of the North Anatolian Fault Zone. *Journal of Geophysical Research*, **114**, B08405, doi:10.1029/2008JB006026.

Heki, K., 2001. Seasonal modulation of interseismic strain buildup in northeastern Japan driven by snow loads. *Science*, **293**, 89–92.

Heki, K., Miyazaki, S., and Tsuji, H., 1997. Silent fault slip following an interplate thrust earthquake at the Japan Trench. *Nature*, **386**, 595–598.

Hreinsdottir, S., Freymueller, J. T., Bürgmann, R., and Mitchell, J., 2006. Coseismic deformation of the 2002 Denali fault earthquake: insights from GPS measurements. *Journal of Geophysical Research*, **111**, B03308, doi:10.1029/2005JB003676.

Hsu, Y. J., Simons, M., Avouac, J.-P., Galetzka, J., Sieh, K., Chlieh, M., Natawidjaja, D., Prawirodirdjo, L., and Bock, Y., 2006. Frictional afterslip following the 2005 Nias-Simeulue earthquake, Sumatra. *Science*, **312**, 1921–1926, doi:10.1126/science.1126960.

Hudnut, K. W., Bock, Y., Cline, M., Fang, P., Feng, Y., Freymueller, J., Ge, X., Gross, W. K., Jackson, D., Kim, M., King, N. E., Langbein, J., Larsen, S. C., Lisowski, M., Shen, Z.-K., Svarc, J., and Zhang, J., 1994. Co-seismic displacements of the 1992 Landers earthquake sequence. *Bulletin of the Seismological Society of America*, **84**, 625–645.

Ide, S., Beroza, G. C., Shelly, D. R., and Uchide, T., 2007. A scaling law for slow earthquakes. *Nature*, **447**, 76–79, doi:10.1038/nature05780.

Ji, C., Larson, K. M., Tan, Y., Hudnut, K. W., and Choi, K., 2004. Slip history of the 2003 San Simeon earthquake constrained by combining 1-Hz GPS, strong motion, and teleseismic data. *Geophysical Research Letters*, **31**, L17608, doi:10.1029/2004GL020448.

Johanson, I. A., Fielding, E. J., Rolandone, F., and Bürgmann, R., 2006. Coseismic and postseismic slip of the 2004 Parkfield earthquake from space-geodetic data. *Bulletin of the Seismological Society of America*, **96**, 269–282.

Johansson, J. M., et al., 2002. Continuous GPS measurements of postglacial adjustment in Fennoscandia 1. Geodetic results. *Journal of Geophysical Research*, **107**, 2157, doi:10.1029/2001JB000400.

Johnson, K., Bürgmann, R., and Freymueller, J. T., 2009. Coupled afterslip and viscoelastic flow following the 2002 Denali Fault, Alaska earthquake. *Geophysical Journal International*, **176**, 670–682, doi:10.1111/j.1365-246X.2008.04029.x.

Jousset, P., Mori, H., and Okada, H., 2003. Elastic models for the magma intrusion associated with the 2000 eruption of Usu Volcano, Hokkaido, Japan. *Journal of Volcanology and Geothermal Research*, **125**, 81–106, doi:10.1016/S0377-0273(03)00090-8.

Khan, S. A., Wahr, J., Leuliette, E., van Dam, T., Larson, K. M., and Francis, O., 2008. Geodetic measurements of postglacial adjustments in Greenland. *Journal of Geophysical Research*, **113**, B02402, doi:10.1029/2007JB004956.

Khan, S. A., Wahr, J., Bevis, M., Velicogna, I., and Kendrick, E., 2010. Spread of ice mass loss into northwest Greenland observed by GRACE and GPS. *Geophysical Research Letters*, **37**, L06501, doi:10.1029/2010GL042460.

Kreemer, C., Blewitt, G., and Maerten, F., 2006. Co- and postseismic deformation of the 28 March 2005 Nias M_w 8.7 earthquake from continuous GPS data. *Geophysical Research Letters*, **33**, L07307, doi:10.1029/2005GL025566.

Larsen, C. F., Motyka, R. J., Freymueller, J. T., Echelmeyer, K. A., and Ivins, E. R., 2005. Rapid viscoelastic uplift in southeast Alaska caused by post-Little Ice Age glacial retreat. *Earth and Planetary Science Letters*, **237**, 548–560.

Larsen, C. F., Motyka, R. J., Arendt, A. A., Echelmeyer, K. A., and Geissler, P. E., 2007. Glacier changes in southeast Alaska and northwest British Columbia and contribution to sea level rise. *Journal of Geophysical Research*, **112**, F01007, doi:10.1029/2006JF000586.

Larson, K., Freymueller, J., and Philipsen, S., 1997. Global plate velocities from the Global Positioning System. *Journal of Geophysical Research*, **102**, 9961–9981.

Larson, K. M., Bodin, P., and Gomberg, J., 2003. Using 1-Hz GPS data to measure deformations caused by the Denali Fault earthquake. *Science*, **300**, 1421–1424.

Lemke, P., Ren, J., Alley, R. B., Allison, I., Carrasco, J., Flato, G., Fujii, Y., Kaser, G., Mote, P., Thomas, R. H., and Zhang, T., 2007. Observations: changes in snow, ice and frozen ground. In Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor, M., and Miller, H. L. (eds.), *Climate Change 2007: The Physical Science Basis*. Cambridge/New York: Cambridge University Press. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change.

Lidberg, M., Johansson, J. M., Scherneck, H.-G., and Davis, J. L., 2007. An improved and extended GPS-derived 3D velocity field of the glacial isostatic adjustment (GIA) in Fennoscandia. *Journal of Geodesy*, **81**, 213–230.

Lienkaemper, J. J., Galehouse, J. S., and Simpson, R. W., 1997. Creep response of the Hayward fault to stress changes caused by the Loma Prieta earthquake. *Science*, **276**, 2014–2016, doi:10.1126/science.276.5321.2014.

Lisowski, M., Dzurisin, D., Denlinger, R. P., and Iwatsubo, E. Y., 2008. Analysis of GPS-measured deformation associated with the 2004–2006 dome-building eruption of Mount St. Helens, Washington. In Sherrod, D. R., Scott, W. E., and Stauffer, P. H. (eds.), A Volcano Rekindled; the Renewed Eruption of Mount St. Helens, 2004–2006: U.S. Geological Survey Professional Paper 1750, 856 p., DVD-ROM [http://pubs.usgs.gov/pp/1750/].

Loveless, J. P., and Meade, B. J., 2010. Geodetic imaging of plate motions, slip rates, and partitioning of deformation in Japan. *Journal of Geophysical Research*, **115**, B02410, doi:10.1029/2008JB006248.

Lu, Z., Masterlark, T., and Dzurisin, D., 2005. Interferometric synthetic aperture radar study of Okmok volcano, Alaska, 1992–2003: Magma supply dynamics and postemplacement lava flow deformation. *Journal of Geophysical Research*, **110**, B02403, doi:10.1029/2004JB003148.

Ma, C., Sauber, J., Bell, L., Clark, T., Gordon, D., Himwich, W., and Ryan, J., 1990. Measurement of Horizontal Motions in Alaska Using Very Long Baseline Interferometry. *Journal of Geophysical Research*, **95**, 21991–22011.

Mattia, M., Rossi, M., Guglielmino, F., Aloisi, M., and Bock, Y., 2004. The shallow plumbing system of Stromboli Island as imaged from 1 Hz instantaneous GPS positions. *Geophysical Research Letters*, **31**, L24610.

McCaffrey, R., 2002. Crustal block rotations and plate coupling. In Stein, S., and Freymueller, J. (eds.), *Plate Boundary Zones*. Washington: American Geophysical Union. AGU Geodynamics Series, Vol. 30, pp. 101–122.

McCaffrey, R., 2008. The tectonic framework of the Sumatran subduction zone. *Annual Reviews of Earth and Planetary Sciences*, **37**, 345–366.

McCaffrey, R., Qamar, A. I., King, R. W., Wells, R., Khazaradze, G., Williams, C. A., Stevens, C. W., Vollick, J. J., and Zwick, P. C., 2007. Fault locking, block rotation and crustal deformation in the Pacific Northwest. *Geophysical Journal International*, **169**, 1315–1340, doi:10.1111/j.1365-246X.2007.03371.x.

Argus, D. F., and Heflin, M. B., 1995. Plate motion and crustal deformation estimated with geodetic data from the Global Positioning System. *Geophysical Research Letters*, **22**, 1973–1976.

Argus, D. F., Gordon, R. G., Heflin, M. B., Ma, C., Eanes, R. J., Willis, P., Peltier, W. R., and Owen, S. E., 2010. The angular velocities of the plates and the velocity of Earth's centre from space geodesy. *Geophysical Journal International*, **181**, 1–48, doi:10.1111/j.1365-246X.2009.04463.x.

Banerjee, P., Pollitz, F. F., and Bürgmann, R., 2005. The size and duration of the Sumatra-Andaman earthquake from far-field static offsets. *Science*, **308**, 1769–1772.

Banerjee, P., Bürgmann, R., Nagarajan, B., and Apel, E., 2008. Intraplate deformation of the Indian subcontinent. *Geophysical Research Letters*, **35**, L18301, doi:10.1029/2008GL035468.

Blewitt, G., Heflin, M., Hurst, K., Jefferson, D., Webb, F., and Zumberge, J., 1993. Absolute Far-Field Displacements from the 28 June 1992 Landers earthquake sequence. *Nature*, **361**, 340–342.

Bock, Y., Agnew, D. C., Fang, P., Genrich, J. F., Hager, B. H., Herring, T. A., Hudnut, K. W., King, R. W., Larsen, S., Minister, J. B., Stark, K., Wdowinski, S., and Wyatt, F. K., 1993. Detection of crustal deformation from the Landers earthquake using continuous geodetic measurements. *Nature*, **361**, 337–340.

Bonafede, M., Strehlau, J., and Ritsema, A. R., 1992. Geophysical and structural aspects of fault mechanics – a brief historical overview. *Terra Nova*, **4**, 458–463.

Bürgmann, R., and Dresen, G., 2008. Rheology of the lower crust and upper mantle: evidence from rock mechanics, geodesy, and field observations. *Annual Review of Earth and Planetary Sciences*, **36**, 531–567, doi:10.1146/annurev.earth.36.031207.124326.

Bürgmann, R., Segall, P., Lisowski, M., and Svarc, J. L., 1997. Postseismic strain following the 1989 Loma Prieta earthquake from GPS and leveling measurements. *Journal of Geophysical Research*, **102**, 4933–4955.

Calais, E., Han, J. Y., DeMets, C., and Nocquet, J. M., 2006. Deformation of the North American plate interior from a decade of continuous GPS measurements. *Journal of Geophysical Research*, **111**, B06402, doi:10.1029/2005JB004253.

Cervelli, P., Fournier, T., Freymueller, J. T., and Power, J., 2006. Ground deformation associated with the precursory unrest and early phases of the January 2006 eruption of Augustine Volcano, Alaska. *Geophysical Research Letters*, **33**, L18304, doi:10.1029/2006GL027219.

Chlieh, M., Avouac, J. P., Sieh, K., Natawidjaja, D. H., and Galetzka, J., 2008. Heterogeneous coupling of the Sumatran megathrust constrained by geodetic and paleogeodetic measurements. *Journal of Geophysical Research*, **113**, B05305, doi:10.1029/2007JB004981.

Cohee, B. P., and Beroza, G. C., 1994. Slip distribution of the 1992 Landers earthquake and its implication for earthquake source mechanics. *Bulletin. Seismological Society of America*, **84**, 646–659.

Cohen, S. C., and Freymueller, J. T., 2004. Crustal deformation in Southcentral Alaska: the 1964 Prince William sound earthquake subduction zone. *Advances in Geophysics*, **47**, 1–63.

Cross, R. S., and Freymueller, J. T., 2008. Evidence for and implications of a Bering plate based on geodetic measurements from the Aleutians and western Alaska. *Journal of Geophysical Research*, **113**, B07405, doi:10.1029/2007JB005136.

DeMets, C., Gordon, R., Argus, D., and Stein, S., 1990. Current plate motions. *Geophysical Journal International*, **101**, 425–478.

DeMets, C., Gordon, R. G., Argus, D. F., and Stein, S., 1994. Effect of recent revisions to the geomagnetic reversal time-scale on estimates of current plate motions. *Geophysical Research Letters*, **21**, 2191–2194.

DeMets, C., Gordon, R. G., and Argus, D. F., 2010. Geologically current plate motions. *Geophysical Journal International*, **181**, 1–80, doi:10.1111/j.1365-246X.2009.04491.x.

Dolan, J. F., Bowman, D. D., and Sammis, C. G., 2007. Long range and long-term fault interactions in Southern California. *Geology*, **35**, 855–858, doi:10.1130/G23789A.1.

Dragert, H., Wang, K., and James, T. S., 2001. A silent slip even on the deeper Cascadia subduction interface. *Science*, **292**, 1525–1528, doi:10.1126/science.1060152.

Dvorak, J. J., and Dzurisin, D., 1993. Variations in magma supply rate at Kilauea Volcano, Hawaii. *Journal of Geophysical Research*, **98**, 22255–22268.

Dvorak, J. J., and Dzurisin, D., 1997. Volcano geodesy: the search for magma reservoirs and the formation of eruptive vents. *Reviews of Geophysics*, **35**, 343–384.

Dvorak, J. J., and Mastrolorenzo, G., 1991. The mechanisms of recent vertical crustal movements in Campi Flegrei caldera, southern Italy. *Geological Society of America*, Special Paper **263**, 47 pp.

Dyurgerov, M. B., and Meier, M. F., 2005. Glaciers and the changing Earth system: a 2004 snapshot. INSTAAR Occasional Paper 58, Boulder: Institute of Arctic and Alpine Research, University of Colorado

Farrell, W. E., 1972. Deformation of the Earth by surface loads. *Reviews of Geophysics*, **10**, 761–797.

Fournier, T., Freymueller, J. T., and Cervelli, P., 2009. Tracking magma volume recovery at Okmok Volcano using GPS and an unscented Kalman filter. *Journal of Geophysical Research*, **114**, B02405, doi:10.1029/2008JB005837.

Freed, A. M., Bürgmann, R., Calais, E., Freymueller, J., and Hreinsdóttir, S., 2006a. Implications of deformation following the 2002 Denali, Alaska earthquake for postseismic relaxation processes and lithospheric rheology. *Journal of Geophysical Research*, **111**, B01401, doi:10.1029/2005JB003894.

Freed, A., Burgmann, R., Calais, E., and Freymueller, J., 2006b. Stress-dependent power-law flow in the upper mantle following the 2002 Denali, Alaska, earthquake. *Earth and Planetary Science Letters*, **252**, 481–489.

Freymueller, J. T., King, N. E., and Segall, P., 1994. The co-seismic slip distribution of the Landers earthquake. *Bulletin of the Seismological Society of America*, **84**, 646–659.

Freymueller, J. T., Woodard, H., Cohen, S., Cross, R., Elliott, J., Larsen, C., Hreinsdóttir, S., and Zweck, C., 2008. Active deformation processes in Alaska, based on 15 years of GPS measurements. In Freymueller, J. T., Haeussler, P. J., Wesson, R., and Ekstrom, G. (eds.), *Active Tectonics and Seismic Potential of Alaska*. Washington: American Geophysical Union. AGU Geophysical Monograph, Vol. 179, pp. 1–42.

Freymueller, J. T., Freed, A. M., Johnson, K. M., Bürgmann, R., Calais, E., Pollitz, F. F., and Biggs, J., 2009. Denali fault earthquake postseismic deformation models. *Eos Transactions of the AGU*, 90, Fall meeting supplement, Abstract G34A-05.

Gan, W., Zhang, P., Shen, Z.-K., Niu, Z., Wang, M., Wan, Y., Zhou, D., and Cheng, J., 2007. Present-day crustal motion within the Tibetan Plateau inferred from GPS measurements. *Journal of Geophysical Research*, **112**, B08416, doi:10.1029/2005JB004120.

Gendt, G., Montag, H., and Dick, G., 1993. Plate kinematics in a global and European scale by LAGEOS laser ranging data from 1983 to 1990. In Smith, D. E., and Turcotte, D. L. (eds.), *Contributions of Space Geodesy to Geodynamics: Crustal Dynamics*. Washington: American Geophysical Union. AGU Geophysical Monograph, Vol. 23, pp. 311–330.

Gilbert, G. K., 1884. A theory of the earthquakes of the Great Basin with a practical application. *American Journal of Science*, **27**, 49–53. reprinted from the *Salt lake Tribune* of 20 Sept. 1883.

Gordon, D., Ma, C., and Ryan, J. W., 1993. Results from the CDP mobile VLBI program in the western United States. In Smith,

Sella, G. F., Stein, S., Dixon, T. H., Craymer, M., James, T. S., Mazzotti, S., and Dokka, R. K., 2007. Observation of glacial isostatic adjustment in "stable" North America with GPS. *Geophysical Research Letters*, **34**, L02306, doi:10.1029/2006GL027081.
Sieh, K., Jones, L., Hauksson, E., Hudnut, K., Eberhart-Phillips, D., Heaton, T., Hough, S., Hutton, K., Kanamori, H., Lilje, A., Lindvall, S., McGill, S., Mori, J., Rubin, C., Spotila, J. A., Stock, J., Thio, H., Treiman, J., Wernicke, B., and Zachariasen, J., 1993. Near-field investigations of the Landers earthquake sequence, April to July, 1992. *Science*, **260**, 171–176.
Smith, S. W., and Wyss, M., 1968. Displacement on the San Andreas fault subsequent to the 1966 Parkfield earthquake. *Bulletin of the Seismological Society of America*, **58**, 1955–1973.
Smith, D., Kolenkiewicz, R., Dunn, P., Robbins, J., Torrence, M., Klosko, S., Williamson, R., Pavlis, E., Douglas, N., and Fricke, S., 1990. Tectonic motion and deformation from satellite laser ranging to LAGEOS. *Journal of Geophysical Research*, **95**, 22013–22041.
Steblov, G. M., Kogan, M. G., Levin, B. V., Vasilenko, N. F., Prytkov, A. S., and Frolov, D. I., 2008. Spatially linked asperities of the 2006–2007 great Kuril earthquakes revealed by GPS. *Geophysical Research Letters*, **35**, L22306, doi:10.1029/2008GL035572.
Stein, S., 1993. Space geodesy and plate motions. In Smith, D. E., and Turcotte, D. L. (eds.), *Contributions of Space Geodesy to Geodynamics: Crustal Dynamics*. Washington: American Geophysical Union. AGU Geophysical Monograph, Vol. 23, pp. 5–20.
Suito, H., and Freymueller, J. T., 2009. A viscoelastic and afterslip postseismic deformation model for the 1964 Alaska earthquake. *Journal of Geophysical Research*, **114**, B11404, doi:10.1029/2008JB005954.
Tamisiea, M. E., Mitrovica, J. X., Milne, G. A., and Davis, J. L., 2001. Global geoid and sea level changes due to present-day ice mass fluctuations. *Journal of Geophysical Research*, **106**, 30849–30863.
Tamisiea, M. E., Leuliette, E. W., Davis, J. L., and Mitrovica, J. X., 2005. Constraining hydrological and cryospheric mass flux in southeastern Alaska using space-based gravity measurements. *Geophysical Research Letters*, **32**, L20501, doi:10.1029/2005GL023961.
Taylor, M., and Yin, A., 2009. Active structures of the Himalayan-Tibetan orogen and their relationships to earthquake distribution, contemporary strain field, and Cenozoic volcanism. *Geosphere*, **5**, 199–214, doi:10.1130/GES00217.1.
Thatcher, W., 2003. GPS constraints on the kinematics of continental deformation. *International Geology Review*, **45**, 191–212.
Thatcher, W., 2009. How the continents deform: the evidence from tectonic geodesy. *Annual Reviews of Earth and Planetary Sciences*, **17**, 237–262, doi:10.1146/annurev.earth.031208.100035.
Titus, S. J., Tikoff, B., and DeMets, C., 2005. Slip rate estimates on the creeping segment of the San Andreas fault, California: Evidence for distributed transcurrent motion. *Geology*, **33**, 205–208.
Wang, K., 2007. Elastic and viscoelastic models of crustal deformation in subduction earthquake cycles. In Dixon, T. H., and Moore, J. C. (eds.), *The Seismogenic Zone of Subduction Thrust Faults*. New York: Columbia University Press, pp. 540–575.
Wilson, P., and Reinhart, E., 1993. The WEGENER-MEDLAS project preliminary results on the determination of the geokinematics of the eastern Mediterranean. In Smith, D. E., and Turcotte, D. L. (eds.), *Contributions of Space Geodesy to Geodynamics: Crustal Dynamics*. Washington: American Geophysical Union. AGU Geophysical Monograph, Vol. 23, pp. 299–310.
Wöppelmann, G., Letetrel, C., Santamaria, A., Bouin, M.-N., Collilieux, X., Altamimi, Z., Williams, S. D. P., and Miguez, B. M., 2009. Rates of sea-level change over the past century in a geocentric reference frame. *Geophysical Research Letters*, **36**, L12607, doi:10.1029/2009GL038720.

Cross-references

Geodesy, Ground Positioning and Leveling
Geodesy, Networks and Reference Systems
GPS, Data Acquisition and Analysis
SAR Interferometry
Satellite Laser Ranging
Very Long Baseline Interferometry

GRAVIMETERS

Andrew Hugill
Toronto, ON, Canada

Synonyms

Gravity meters, relative

Definition

An instrument which measures differences in gravity. Units: Practical units are mGal (10^{-6} g) and μGal (10^{-9} g). 1 Gal = 1 cm/s^2.

Introduction

The first relative gravity measurements were made in 1672 by the French astronomer, Richer, who noted a difference in the period of a pendulum when observed in Paris and close to the equator in Cayenne, South America. For the next two centuries, relative gravity measurements, mainly for scientific purposes, were made with pendulums. In 1893, Threlfall and Pollock made the first successful spring-mass gravimeter in Sydney, Australia (Threlfall and Pollock, 1900) using the newly discovered quartz fibers. From the middle of the twentieth century up until the present nearly all relative gravity measurements have been made with instruments based on a mass suspended by a quartz or metal elastic system. Instruments in which the mass is levitated in the field of a superconducting magnet and instruments based on inertial-grade accelerometers are also used in some applications.

Absolute gravity meters, which are sometimes used to make relative gravity measurements (e.g., Brady et al., 2008), obtain a value for gravity through measurement of the quantities distance and time without reference to other gravity values. These instruments are beyond the scope of this article and are described in the "Gravity Measurements, Absolute" entry of this encyclopedia.

Relative gravimeters can be classified into three types according to the mode in which they are operated, with each mode of operation having different performance requirements:

1. Survey: Measurement of relative gravity as a function of position on, or under, the surface of the earth or on the sea floor.
2. Stationary: Continuous measurement of changes in gravity or low-frequency earth motion as a function

of time at a fixed location on, or under, the surface of the earth or on the sea floor.
3. Dynamic: Measurement of relative gravity as a function of position from a moving platform such as an aircraft or a ship.

Performance requirements

Basic performance requirements for relative gravimeters are listed in Table 1. These must be met under a range of environmental conditions, including movement of the sensor between and during readings. Dynamic instruments have to be able to produce data while moving and subject to relatively large accelerations, whereas the defining characteristic of survey instruments is the capability to withstand acceleration and shocks during transport without sustaining changes in the reading value.

Survey instruments can be subdivided into borehole, sea-floor, and land gravimeters all with the same basic performance requirements. However, borehole and sea-floor instruments must have remote reading capabilities and be packaged to withstand high pressure. In addition, borehole instruments have size restrictions and have to operate at higher temperatures. Survey gravimeters have a useable sensitivity of less than 5 μGal (e.g., Zumberge et al., 2008; Merlet et al., 2008). A total range of 7,000 mGal covers the gravity field on the surface of the earth. A minimum continuous operating range of around 50 mGal is needed to provide useful measurement capability.

Stationary gravimeters can be divided into those which require very low noise performance and which are usually located in an observatory and portable instruments. Most applications for observatory instruments (Crossley et al., 1999) are in the frequency range from dc up to approximately 7 mHz which is the maximum frequency at which the seismic normal modes are resolvable (Ingate and Berger, 2004). Minimum earth noise in this band is around $3 \times 10^{-2}\mu\text{Gal}/\sqrt{\text{Hz}}$ (Ingate and Berger, 2004) which represents the minimum useful instrument sensitivity. Superconducting instruments and the metal spring instruments which approach this sensitivity (Banka and Crossley, 1999) have a range of a few mGal or less and are saturated by surface waves of large earthquakes (e.g., Mansinha, 1992; Agnew et al., 1986). Applications for portable instruments require a larger range and less sensitivity and include volcanic monitoring, earth tide, and ocean loading studies, groundwater monitoring, and earthquake monitoring.

Useable sensitivity for airborne gravity meters is limited by GPS precision as described in the *Sensor motion and orientation* section below. With a 100 s low pass filter, this limit is around 0.5 mGal (Argyle et al., 2000).

Technology

Sensor technology

Relative gravimeters are based on one of the four core technologies listed in the top row of Table 2. Technologies have characteristics which better suit them to particular types of gravity meter as discussed below.

Fused quartz

Fused quartz sensors perform well in survey gravimeters because quartz has high strength, remains almost perfectly elastic up to its breaking stress, and can be welded into compact structures which have exceptional mechanical integrity. Sensors are free of tares when transported and do not have to be clamped between readings, making them accurate and quick and easy to operate. The relatively high long-term drift of quartz springs is linear enough to be compensated electronically.

Metal

Metal spring-mass sensors have very low long-term drift and low temperature coefficient, making them well suited to stationary measurements. However, they are more susceptible to tares in response to transportation because the elastic properties of the materials in their elastic suspension are inferior to those of quartz. Also they are assembled from a number of components fixed together, so they do not have the integrity of a welded quartz structure. These inherent disadvantages have been overcome by clamping the mass during transportation and careful handling, but at the expense of added manufacturing complexity, requirement for more skilful operators and lower productivity in the field. Historically, survey gravimeters based on metal spring-mass sensors were used extensively in survey operations because of their low long-term drift and high accuracy, especially for establishing gravity networks and other applications which require measurement of gravity at widely separated points. Absolute gravimeters are now used to make large-scale-network-type

Gravimeters, Table 1 Gravimeter performance requirements

	Performance			
			Range (mGal)	
Type of gravimeter	Signal bandwidth	Sensitivity (μGal)	Total	Continuous (min)
Survey	dc	1–10	7,000	50
Stationary: Observatory	dc − 10 mHz	$3 \times 10^{-2}/\sqrt{\text{Hz}}$	7,000	2
Stationary: Portable	dc − 1 Hz	1	7,000	50
Dynamic	dc − 10 mHz	500	10,000	10,000

Gravimeters, Table 2 Gravity sensor core technology

Technology / Type of gravimeter	Spring/mass		Superconducting	Inertial-grade accelerometer
	Quartz	Metal		
Survey	▨	▨		
Stationary		▨	▨	
Dynamic		▨		▨

measurements and survey instruments with metal spring-mass sensors are being replaced for most other applications by modern quartz instruments which are easier to manufacture and operate, and make extensive use of electronics to obtain similar or better performance.

Metal spring-mass gravimeters were first used for dynamic applications in the late 1950s (Dehlinger, 1978). Since then they have been continually refined and modified for this application and are in widespread use today (e.g., Williams and MacQueen, 2001).

Superconducting

The superconducting gravity sensor is described by one of its inventers (Goodkind, 1999) as "a spring type gravimeter in which the mechanical spring is replaced by a magnetic levitation of a superconducting sphere in the field of superconducting, persistent current coils." Because these currents are very stable, the sensor has extremely low drift which makes it well suited for stationary operation. It is unsuited for survey or dynamic operation as the mass is insufficiently constrained to continue operation in the presence of large accelerations.

Inertial-grade accelerometers

These are devices that have been developed for inertial navigation systems so have excellent dynamic performance on moving platforms and are very compact. They have sufficient sensitivity for dynamic operations, but not for survey or stationary operations. An example of an inertial-grade accelerometer is the Q Flex QA3000 which has a volume of approximately 1 cu. inch and is built around a chemically machined quartz suspension (Foote and Grindeland, 1992). Another example is the Bell Aerospace accelerometer (Torge, 1989) which has similar dimensions.

Design principles

A generalized signal flow diagram for relative gravimeters is shown in Figure 1. The proof mass is suspended so that it moves in a well-defined trajectory in response to a change in the gravitational force on it. The proof mass and suspension are termed the gravity sensor. The gravity reading is derived from the input to the calibrated feedback actuator that returns the output of the displacement sensor to the reference point. Feedback force is applied to the proof mass through an electric field or electromagnetic actuator and/or through a change to the suspension geometry.

Gravity sensor natural period, sensitivity, and seismic noise

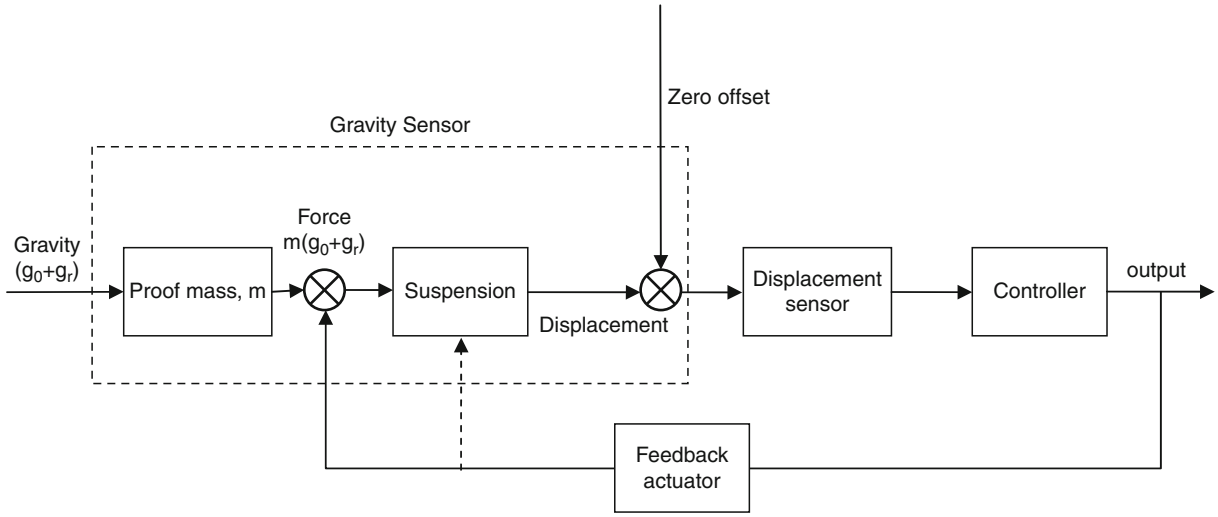
The design principles of the gravity sensor determine important performance characteristics of the gravimeter that are largely independent of the technology used in construction.

The gravitational force on the mass of a spring-mass gravity sensor is balanced by the elastic force generated by the extension of a quartz or metal spring. The extended length of the spring is given by $l = \frac{mg}{k} + l_u$ where k is the spring constant and l_u is the unstressed length of the spring. A change in gravity from a value g_0 to $(g_0 + \Delta g)$ produces a change in spring length from l_0 to $(l_0 + \Delta l)$ where $\Delta l/(l_0 - l_u) = \Delta g/g_0$. As Δl is proportional to $(l_0 - l_u)$ this simple spring-mass system is termed a proportional sensor, or to use the language of seismometry, a short-period sensor.

For a practical sensor with $l_0 - l_u = 2$ cm and a change in gravity of 0.01 mGal (1 Gal = 1 cm/s^2), $\Delta g/g_0 = 10^{-8}$ and the displacement which must be detected is $\Delta l = 2 \times 10^{-10}$ m. This system responds equally to gravity and all acceleration up to its natural frequency $f_n = \frac{1}{2\pi}\sqrt{\frac{k}{m}} = \frac{1}{2\pi}\sqrt{\frac{g_0}{(l_0 - l_u)}} = 3.5\,\text{Hz}$, so unwanted signals, including microseismic and cultural noise, would need to be filtered from the output to recover the gravity signal.

It was not until the 1970s and 1980s that electronics were available to detect sub-nm displacements and effectively filter signals in a portable instrument. The Scintrex CG-3 gravity meter, introduced in 1987, and its successor the CG-5 are based on a proportional spring-mass system and use capacitive displacement sensing and digital filtering to achieve the required level of sensitivity and noise reduction (Hugill, 1990).

Before the advent of modern electronics, many mechanical design concepts were developed (e.g., Hugill, 1984) to increase sensitivity and reduce natural frequency, and thus noise response, of sensors enabling detection of gravity-induced mass displacement by eye, through a microscope. Sensors with increased mechanical sensitivity are termed "astatic" and are analogous to long-period seismometers.



Gravimeters, Figure 1 Relative gravimeter functional diagram.

The most successful of these was the LaCoste suspension based on "zero-length spring" with an unstressed length $l_u = 0$ (LaCoste, 1934). Sensitivity is adjustable and is usually set up with sufficient displacement sensitivity for optical detection by eye with a microscope and a natural frequency low enough for the instrument to be insensitive to seismic noise. Worden-type and LaCoste and Romberg gravimeters, which were introduced in the 1940s and 1950s, are based on the LaCoste suspension.

The sensor in the superconducting gravimeter (Goodkind, 1999) is also astatised with its sensitivity being increased by adjusting the relative current settings in two collinear coils to shape the magnetic field gradient in the vicinity of the levitated proof mass. Superconducting sensors are usually adjusted to operate with a period of 5 s.

Displacement sensing and feedback

Most modern gravimeters use capacitive displacement sensing and either electromagnetic or electrostatic feedback. Electrostatic feedback allows sensor design to be simplified as the same electrodes can be employed for sensing, using high frequency excitation, and feedback at lower frequencies (e.g., Block and Moore, 1966; Hugill, 1984). Examples of instruments using capacitive displacement sensing and electrostatic feedback include the Scintrex CG-3 and CG-5 (Hugill, 1990) and the micro-g LaCoste gPhone (Micro-g LaCoste, 2010).

Electromagnetic feedback has the advantage of inherent linearity and a higher force capability. Instruments with capacitive displacement sensing and electromagnetic feedback include the Bodenseewerke GSS30 (Torge, 1989), the superconducting gravimeter (Goodkind, 1999) and inertial-grade accelerometers. In the latter, feedback supports the full gravitational force on the mass so the spring suspension only provides constraint of the proof mass motion (e.g., Foote and Grindeland, 1992) which is equivalent to having no zero offset in Figure 1.

In most gravimeters with the LaCoste suspension the top of the mainspring is moved to null the mass. This is done in Worden-type meters through a gearbox, screw thread, and auxiliary springs. In the LaCoste and Romberg g meter (LaCoste and Romberg, 1945), the top of the spring is connected to a lever system which is moved by the operator through a calibrated gearbox and screw thread. The required displacement for a change in gravity of 0.01 mGal is 0.25 nm (Torge, 1989). Electrostatic feedback systems are installed in gravimeters with metal LaCoste suspensions as an option for survey instruments to make reading easier and more accurate (e.g., Webring et al., 2004) and to enable high sensitivity automatic reading for stationary instruments (e.g., Block and Moore, 1966).

Error sources

Error sources can be divided into three broad groups: those generated by the gravimeter or intrinsic error sources, those resulting from the response of the instrument to external inputs and those resulting from the effects of transportation.

Intrinsic error sources

Scale factor changes

Changes in the scale factor of gravimeters with quartz sensors can be as high as 10^{-3} in the first year after manufacture; subsequent changes are of the order of 10^{-4} per year (Zumberge et al., 2008; Carbone and Rymer, 1999). Changes in scale factor of around 10^{-3} have also been reported in metal spring sensors (Carbone and Rymer, 1999). In contrast to the smooth drift in the scale factor of quartz sensors, the reported changes in the metal spring

sensor were abrupt and separated by long periods of stability.

As the scale factor, k, relates a reading differences $(R_2 - R_1)$ to the gravity change $(g_2 - g_1)$ through the expression $(g_2 - g_1) = k(R_2 - R_1)$ errors due to changes in scale factor are proportional to gravity difference, so have most effect on measurements made by survey gravimeters. Dynamic instruments are less influenced because the sensitivity requirement is lower. Good field practise, including routine calibration and use of multiple instruments for critical measurements, prevents degradation of survey results.

Instrument noise

Thermal noise determines the detection limit of the sensor and is given by (e.g., Richter et al., 1995)

$$a_t^2 = \frac{25 k_b T}{Q m T_n} \left(\frac{(\mathrm{ms}^{-2})^2}{\mathrm{Hz}} \right)$$

where a_t is the rms noise equivalent acceleration spectral density, k_b is Boltzman's constant, T is absolute temperature, Q is the quality factor, m is the mass of the proof mass and T_n is the natural period. For a metal spring LaCoste and Romberg earth tide gravimeter operating at 323 K with $Q = 1$, $m = 80$ g, and $T_n = 20$ s, the thermal noise is $a_t^2 = 7 \times 10^{-20}$ $\mathrm{m^2 s^{-4}/Hz}$ (Richter et al., 1995) and for a superconducting gravity sensor operating at 4 K with $Q = 0.14$, $m = 4$ g and $T_n = 5$ s, the thermal noise is $a_t^2 = 5 \times 10^{-19}$ $\mathrm{m^2 s^{-4}/Hz}$ and could be a factor of 3 higher, depending on the extent to which the horizontal modes couple into the vertical (Van Camp, 1999).

The thermal noise limit is therefore a significant design parameter for stationary gravimeters measuring in the frequency band dc –7 mHz where earth noise power spectral density can be lower than 10^{-19} $\mathrm{m^2 s^{-4}/Hz}$ (Ingate and Berger, 2004).

Long-term drift

Long-term drift is a characteristic of all relative gravimeters and imposes constraints on the way in which instruments are used. Values range from a few μGal/year for instruments with superconducting sensors (e.g., Crossley et al., 2003) to several hundred μGal/day or more for survey gravimeters with quartz sensors (e.g., Bonvalot et al., 1998; Merlet et al., 2008). Drift is modeled as a function of time and removed from measurements as part of data processing. Model parameters for survey instruments are derived from repeat readings on base stations at different times during a survey. Absolute gravity readings obtained at the same site are used as a reference to model the drift of superconducting gravimeters (e.g., Van Camp and Francis, 2007).

External noise sources

The effect of external inputs on the sensor can be reduced to acceptable levels by a number of approaches which are described below. These include shielding the sensitive elements, compensation, and filtering and reducing the input at its source through setup and choice of location and time.

Sensor motion and orientation

Motion of the gravity sensor can produce a response which is only distinguishable from the gravity signal by its frequency content. If the disturbing motion is outside of the signal bandwidth, then its effects can be removed by filtering, otherwise independent means must be used to determine the sensor motion and compensate for its effects. High frequency motion and vibration can excite higher order sensor modes which can couple into the fundamental mode (e.g., Block and Moore, 1970).

The gravity sensor measures the components of gravity and acceleration that are parallel to its sensitive axis. To accurately measure gravity the sensor is vertically aligned. Vertical misalignment of the sensor results in a component of gravity being measured which varies as the cosine of the angle.

Horizontal acceleration can produce "cross coupling" in sensors that have the proof mass in the form of a horizontal pendulum and feedback control which doesn't keep the mass at the zero position with sufficient accuracy (e.g., Torge, 1989). Horizontal velocity causes centrifugal and Coriolis accelerations which combine to produce the Eötvös-effect (Torge, 1989).

The relative magnitude of each of these error sources depends on the mode in which the instrument is operated:

Airborne gravity precision and spatial resolution are ultimately limited by the recovery of aircraft vertical position with GPS (e.g., Bell et al., 1999). This limit has been estimated to be around 1 mGal rms for an 80 s filter (van Kann, 2004), an estimate which is confirmed by actual measurements of aircraft acceleration with GPS (Wei and Schwarz, 1995). Performance at this level has been reported for at least two airborne gravity systems (Studinger et al., 2008). One of these, the Sander Geophysics system, which is based on inertial-grade accelerometer sensors, is unaffected by turbulence up to 0.6 $\mathrm{m/s^2}$ rms (Argyle et al., 2000).

In marine gravity, the ship moves relatively slowly on an approximately equipotential surface (Torge, 1989) so there is no significant vertical acceleration in the signal bandwidth and the relatively high frequency acceleration caused by wave motion is removed by low pass filtering. Horizontal position and velocity, derived from GPS measurements, are used to compensate Eötvös-effects with sufficient accuracy to enable gravity anomalies of 0.2 mGal to be detected with a 0.5 km spatial resolution (LaFehr et al., 1992).

Survey gravimeters are affected by ground motion in the form of microseismic noise, cultural and wind noise, and earthquakes. Microseismic noise peaks in the range between 3 and 8 s (e.g., Ingate and Berger, 2004) and has amplitudes as high as 1 mGal pp (Laswell et al., 2010). Distant earthquakes produce signals with a peak

amplitude at a period of around 20 s (Ingate and Berger, 2004). Cultural and wind noise is generally above 1 Hz (e.g., Wilson, 1953) and depends on local conditions. The effect of these noise sources can be reduced with low pass filtering. Astatised (long-period) sensors have naturally lower sensitivity to this noise; however, proportional (short-period) sensors with digital filtering can produce readings with an rms noise level of a few μGal in most conditions (Scintrex, 2010).

Stationary gravimeters avoid cultural and wind noise through the choice of a quiet reading site (e.g., Torge, 1989). Earthquakes often generate the signal that is being measured. Filtering is used to suppress signals outside the band of interest (e.g., Agnew et al., 1976).

Temperature, pressure, and magnetic field

Sensitivity of spring-mass sensors to these inputs is largely determined by the material that the gravity sensor is constructed from. For example, the elastic suspension of metal spring instruments is made from a material with an elastic modulus that has low temperature sensitivity, but relatively high magnetic field sensitivity. Fused quartz is not sensitive to changes in magnetic field, but its elastic modulus has high temperature sensitivity.

The amplitudes of temperature, pressure, and magnetic field changes that reach the gravity sensor are reduced by control, sealing, and shielding, respectively. In addition, stationary gravimeters are often placed in environments where the temperature is extremely stable (e.g., Van Camp and Francis, 2007). If residual changes reaching the sensor are significant, they are compensated mechanically or electronically. In the Scintrex CG-3 and CG-5 quartz instruments, the temperature is controlled to the m°C level and further compensated to the μ°C level based on a temperature measurement close to the elastic suspension (Hugill, 1990). In the metal spring LaCoste and Romberg instruments residual pressure changes are mechanically compensated (Ander et al., 1999).

Superconducting penetration depth determines the temperature sensitivity of superconducting gravity sensors (Goodkind, 1999), and the levitation force is directly proportional to magnetic field changes. The amplitude of temperature and magnetic field changes at the sensor are reduced to acceptable levels by controlling temperature to a few μ°C and surrounding the instrument with a double μ metal/Niobium magnetic shield.

Transportation effects

Changes in readings, often called tares, changes in drift rate, and elastic relaxation can all contribute to errors as a result of transportation. However, these errors can generally be avoided through good operating procedures. Careful handling and appropriate protection from shock and vibration (e.g., Hamilton and Brule, 1967) can reduce or eliminate tares. Variable drift can be recovered by frequent base station reoccupation (e.g., Webring et al., 2004) and elastic relaxation is minimized in quartz instruments by maintaining vertical orientation during transport (e.g., Zumberge et al., 2008) and by waiting a fixed time after unclamping in metal instruments (e.g., Torge, 1989).

Gravimeters

Some gravimeters in use today are listed in Table 3.

Summary

Relative gravimeters are based on one of four core technologies: spring-mass systems made from quartz or metal, superconducting sensors, or inertial-grade accelerometers. Performance requirements depend on whether instruments are to be used for surveys or in stationary or dynamic

Gravimeters, Table 3 Some relative gravimeters currently in use

Gravimeter	Type	Sensor technology	Period	Displacement sensor	Feedback actuator	Reference
LaCoste and Romberg g meter	Survey	Metal spring-mass	Long	Optical, capacitive	Mechanical (electrostatic)	Webring et al. (2004)
Scintrex CG-3/5	Survey	Quartz spring-mass	Short	Capacitive	Electrostatic	Hugill (1990), Scintrex (2010)
Bodenseewerke GSS30	Dynamic	Metal spring-mass	Short	Capacitive	Electromagnetic	Torge (1989)
GWR	Stationary	Superconducting	Long	Capacitive	Electromagnetic	Goodkind (1999)
ZLS Burris gravimeter	Survey/ Stationary	Metal spring-mass	Long	Capacitive	Electrostatic (mechanical)	ZLS (2010)
Micro-g LaCoste gPhone	Stationary	Metal spring-mass	Long	Capacitive	Electrostatic	Micro-g LaCoste (2010)
Sander AIRGrav	Dynamic	Inertial-grade accelerometer	Short	Capacitive	Electromagnetic	Argyle et al. (2000)
GT-1A	Dynamic	Inertial-grade accelerometer	Short	Optical	Electromagnetic	Canadian Micro Gravity (2010)
ZLS Dynamic Gravity Meter	Dynamic	Metal spring-mass	Long	Capacitive	Mechanical	ZLS (2010)
Micro-g LaCoste Air-Sea II	Dynamic	Metal spring-mass	Long	Capacitive	Mechanical	Micro-g LaCoste (2010)

modes of operation. Useable sensitivity varies from $3 \times 10^{-2}\mu\text{Gal}/\sqrt{\text{Hz}}$ for stationary instruments operated in observatories to 0.5 mGal for dynamic gravimeters operated in aircraft. The defining characteristic of survey instruments is the capability to withstand acceleration and shocks during transport without sustaining changes in the reading value. Dynamic instruments must be able to produce data while moving and subject to large accelerations.

Bibliography

Agnew, D., Berger, J., Buland, R., Farrell, W., and Gilbert, J. F., 1976. International deployment of accelerometers: a network of very long period seismology. *EOS. Transactions of the American Geophysical Union*, **57**, 181–188.

Agnew, D., Berger, J., Buland, R., Farrell, W., and Gilbert, J. F., 1986. Project IDA: a decade in review. *EOS. Transactions of the American Geophysical Union*, **67**(10), 203–212.

Ander, M. E., Summers, T., and Gruchalla, M. E., 1999. LaCoste and Romberg gravity meter: system analysis and instrumental errors. *Geophysics*, **64**, 1708–1719.

Argyle, M., Ferguson, S., Sander, L., and Sander, S., 2000. AIRGrav results: a comparison of airborne gravity data with GSC test site data. *The Leading Edge*, **19**, 1134–1138.

Banka, D., and Crossley, D., 1999. Noise levels of superconducting gravimeters at seismic frequencies. *Geophysical Journal International*, **139**, 87–97.

Bell, R. E., Childers, V. A., and Arko, R. A., 1999. Airborne gravity and precise positioning forgeologic applications. *Journal of Geophysical Research*, **104**(B7), 15281–15292.

Block, B., and Moore, R. D., 1966. Measurements in the earth mode frequency range by an electrostatic sensing and feedback gravimeter. *Journal of Geophysical Research*, **71**, 4361–4375.

Block, B., and Moore, R. D., 1970. Tidal to seismic frequency investigations with a quartz accelerometer of new geometry. *Journal of Geophysical Research*, **75**, 1493–1505.

Bonvalot, S., Diament, M., and Gabalda, G., 1998. Continuous gravity recording with Scintrex CG-3 M meters: a promising tool for monitoring active zones. *Geophysical Journal International*, **135**, 470–494.

Brady, J. L., Hare, J. L., Ferguson, J. F., Seibert, J. E., Klopping, F. J., Chen, T., and Niebauer, T., 2008. Results of the world's first 4D microgravity surveillance of a waterflood–Prudhoe Bay, Alaska. *SPE Reservoir Evaluation & Engineering*, **11**(5), 824–831.

Canadian Micro Gravity, 2010. http://www.canadianmicrogravity.com/.

Carbone, D., and Rymer, H., 1999. Calibration shifts in a LaCoste-and-Romberg gravimeter: comparison with a Scintrex CG-3 M. *Geophysical Prospecting*, **47**(1), 73–83.

Crossley, D., et al., 1999. Network of superconducting gravimeters benefits a number of disciplines. *EOS. Transactions of the American Geophysical Union*, **80**(11), 121, 125, 126.

Crossley, D., Hinderer, J., Llubes, M., and Florsch, N., 2003. Potential of ground gravity measurements to validate GRACE data. *Advances in Geosciences*, **1**, 65–71.

Dehlinger, P., 1978. *Marine Gravity*. Amsterdam: Elsevier.

Foote, S. A., and Grindeland, D. B., 1992. Model QA3000 Q-Flex accelerometer high performance test results. *IEEE Aerospace and Electronic Systems Magazine*, **7**(6), 59–67.

Goodkind, J. M., 1999. The superconducting gravimeter. *The Review of Scientific Instruments*, **70**(11), 4131–4152.

Hamilton, A. C., and Brule, B. G., 1967. Vibration-induced drift in LaCoste and ROmberg geodetic gravimeters. *Journal of Geophysical Research*, **72**(8), 2187–2197.

Hugill, A. L., 1984. *The Design of a Gravimeter with Automatic Readout*. PhD Thesis. Bedford Park, South Australia, Flinders University.

Hugill, A., 1990. Scintrex CG-3 automated gravity meter: description and field results. *SEG Expanded Abstracts*. Society of Exploration Geophysicists, Vol. 9, pp. 601–604.

Ingate, S., and Berger, J., 2004. *Prospects for low-frequency seismometry*. Granlibakken: IRIS Broadband Seismometer Workshop.

LaCoste, L. J. B., 1934. A new type long period vertical seismograph. *Physics*, **5**(7), 178–180.

LaCoste, L. J. B., and Romberg, A., 1945. US Patent 2377889.

LaFehr, T. R., Valliant, H. D., and MacQueen, H. D., 1992. High-resolution marine gravity by digital control. *SEG Expanded Abstracts*. Society of Exploration Geophysicists, Vol. 11, pp. 559–560.

Laswell, S., Niebauer, T. M., Engel, R., Cassidy, J., Courtier, N., and Henton, J., 2010. Recent observations of increased seismic background noise using gPhone gravity meters. *Micro-g LaCoste*. 2010. http://www.microglacoste.com/gPhoneNoise/gPhoneSeismicNoise.pdf.

Mansinha, L., and Kamal, 1992. A test of the superconducting gravimeter as a long-period seismometer. *Physics of the Earth and Planetary Interiors*, **71**(1–2), 52–60.

Melton, B. S., 1976. The sensitivity and dynamic range of inertial seismographs. *Reviews of Geophysics and Space Physics*, **18**(2), 93–116.

Merlet, S., Kopaev, A., Diamen, M., Geneves, G., Landragin, A., and Pereira Dos Santos, F., 2008. Micro-gravity investigations for the LNE watt balance project. *Metrologia*, **45**(3), 265–274.

Micro-g LaCoste, 2010. http://www.microglacoste.com/ (Accessed April 2010).

Richter, B., Wenzel, H.-G., Zürn, W., and Klopping, F., 1995. From Chandler Wobble to free oscillations: comparison of cryogenic gravimeters and other instruments in a wide period range. *Physics of the Earth and Planetary Interiors*, **91**, 131–148.

Scintrex. *CG-5 Scintrex Autograv System Operation Manual*. 2010. http://www.scintrexltd.com/documents/CG-5Rev4.pdf.

Studinger, M., Bell, R., and Frearson, N., 2008. Comparison of AIRGrav and GT-1A airborne gravimeters for research applications. *Geophysics*, **73**(6), 151–161.

Threlfall, R., and Pollock, J. A., 1900. On a quartz thread gravity balance. *Philosophical Transactions of the Royal Society of London. Series A*, **193**, 215–260.

Torge, W., 1989. *Gravimetry*. Berlin: de Gruyter.

Van Camp, M., 1999. Measuring seismic normal modes with the GWR C021 superconducting gravimeter. *Physics of the Earth and Planetary Interiors*, **116**, 81–92.

Van Camp, M., and Francis, O., 2007. Is the instrumental drift of superconducting gravimeters a linear or exponential function of time? *Journal of Geodesy*, **81**, 337–344.

van Kann, F., 2004. Requirements and general principles of airborne gravity gradiometers for mineral exploration." In Lane, R. J. L. (ed.), *Airborne Gravity 2004 – Abstracts from the ASEG-PESA Airborne Gravity 2004 Workshop*, Geoscience Australia Record 2004/18, 1–5.

Webring, M. W., Kucks, R. P., and Abraham, J. D., 2004. Gravity study of the Guernsey Landfill site, Guernsey, Wyoming. U. S. Geological Survey, Reston, VA.

Wei, M., and Schwarz, K. P., 1995. Analysis of GPS derived acceleration from airborne tests. In *Proceedings on the IAG symposium on Airborne Gravity Field Determination at the IUGG XXI General Assembly,* July 2–14, 1995. Alberta, Canada: University of Calgary, Department of Geomatics Engineering, pp. 175–188.

Williams, S., and MacQueen, J. D., 2001. Development of a versatile, commercially proven, and cost-effective airborne gravity system. *The Leading Edge*, **20**(6), 651–654.

Wilson, C. D. V., 1953. An analysis of the vibrations emitted by some man-made sources of microseisms. *Proceedings of the Royal Society of London. Series A*, **217**, 188–202.
ZLS, 2010. http://www.zlscorp.com/products.htm.
Zumberge, M., Alnes, H., Eiken, O., Sasagawa, G., and Stenvold, T., 2008. Precision of seafloor gravity and pressure measurements for reservoir monitoring. *Geophysics*, **73**(6), 133–141.

Cross-references

Earth Tides
Geodesy, Networks and Reference Systems
Gravity Anomalies, Interpretation
Gravity Data, Advanced Processing
Gravity Data, Regional – Residual Separation
Gravity Field of the Earth
Gravity Field, Temporal Variations from Space Techniques
Gravity Measurements, Absolute
Gravity Method, Principles
Gravity Method, Satellite
Gravity Method, Surface
Gravity, Data to Anomalies
Gravity, Global Models
Gravity, Gradiometry

GRAVITY ANOMALIES, INTERPRETATION

Mikhail K. Kaban
Department 1: Geodesy and Remote Sensing, Sektion 1.3: Earth-System Modelling, Deutsches GeoForschungsZentrum Potsdam - GFZ, Potsdam, Germany

Definition

In geophysics *gravity anomalies* are generally defined as the difference between observed gravity field and the field of a reference model. Depending on the reference gravity model, two different types of anomaly variations are considered: *gravity anomalies* and *gravity disturbances*. The geodetic *gravity anomaly* is defined as the difference between gravity on the geoid and normal gravity on the reference ellipsoid (Heiskanen and Moritz, 1967). On the other hand, the *gravity disturbance* is defined as the difference of the fields at the same point on the reference ellipsoid. It has been demonstrated that the gravity disturbances are more appropriate for geophysical purposes (e.g., Hackney and Featherstone, 2003). In any case, it is necessary to take into account this difference in the interpretation. For more details see *Gravity, Data to Anomalies*. Here we use the general term "anomalies" for both types as traditionally accepted in geophysics.

The observed gravity anomalies reflect the effect of density variations relative to the homogeneous reference model. *Interpretation* of the gravity anomalies implies an estimation of these density heterogeneities. The density model should reproduce the observed gravity field, taking into account that the observations may be affected by measurement errors. A separate type of modeling, which may be also related to the interpretation, is an *enhancement* of the gravity anomalies. The observed field combines effects of various mass anomalies integrated over the whole Earth. Therefore, it is often difficult to select the anomaly, which relates to the object under study. The usual way to cope with the problem is to remove the effects, which mask the anomalies of interest (see also *Gravity Data, Regional – Residual Separation*).

Introduction

Density heterogeneity of the Earth, associated with thermal and compositional variations or with deflected boundaries separating layers of different density, is one of the main factors that control dynamic processes and deformations at both shallow and deep levels. Therefore, interpretation of the gravity anomalies or *gravity modeling* is one of the principal methods, which help to understand nature and origin of the tectonic processes and the Earth's dynamics. On the other hand, gravity field is very sensitive to shallow heterogeneities, in particular related to mineral deposits. For this reason, gravity modeling is an important tool for both fundamental and practical purposes.

Formulation of the problem

The relationship between the density distribution ρ and the gravity field g may be written as an operator equation:

$$A\rho = g, \tag{1}$$

where A is the integral operator. In Cartesian coordinates for a 3D body with density $\rho(x',y',z')$ this equation may be written as (Blakely, 1995):

$$g(x,y,z) = \frac{\partial U}{\partial z} = -\gamma \int_{z'} \int_{y'} \int_{x'} \rho(x',y',z') \frac{(z-z')}{r^3} \mathrm{d}x'\mathrm{d}y'\mathrm{d}z', \tag{2}$$

where (x, y, z) are the observation points, U is the gravitational potential, γ is the gravitational constant, and

$$r = \sqrt{(x-x')^2 + (y-y')^2 + (z-z')^2}.$$

Therefore, the problem of determination of the density distribution ρ may be formulated as finding:

$$\min\{||A\rho - g_{\mathrm{obs}}||^2\}, \tag{3}$$

where g_{obs} is the observed gravity anomaly.

Unfortunately, the problem of gravity data inversion is generally *ill posed*. The operator A has no continuous inverse A^{-1}, thus the solution is neither unique nor stable. To overcome this principal weakness Equation 3 should be modified by adding a so-called regularization term Ω (Tikhonov and Arsenin, 1977):

$$\min\{||A\rho - g_{obs}||^2 + \alpha\Omega(\rho)\}, \quad (4)$$

where α defines strength of regularization. The regularization implies some a priori conditions imposed on the density distribution in order to limit a possible range of the solutions. If these conditions are sufficient, the problem becomes *well posed*.

There are basically two ways to obtain a unique and stable solution with the means of interpretation of the gravity anomalies. The first option implies considering simplified models, the parameters of which should be determined. An inclusion of a certain form in a homogeneous or stratified material, a thin layer with variable density or an interface between two layers with different density contrasts, are methods with which the number of unknowns are reduced and the problem is generalized. These models are adequate for modeling strong density contrasts, like the ocean bottom, the basement, the crust–mantle interface, and are optionally used jointly with the concept of *isostasy*. The methods assuming a spatially restricted source are suited in the exploration of superficial bodies as ore bodies, diamond bearing Kimberlyte cylinders, and salt domes. The methods using a density interface are used to define the ocean bathymetry from the gravity field or to determine the Moho, especially in remote areas where seismic studies are scarce or insufficient (Silva et al., 2006; Smith and Sandwell, 1997; Braitenberg et al., 2002). The other way to cope with the problem implies joint analysis of the gravity anomalies together with supplemental geophysical information, in particularly with seismic data that identify the basic structure of the Earth interiors. Using this method, it is possible to construct advanced models, which describe the structure of the Earth more reliably. Two basic interpretation techniques are used within this approach: forward method and inverse method (Blakely, 1995) (see also *Gravity Method, Principles*).

Gravity anomalies: conditions and enhancement

Spatial *resolution* of the gravity anomalies should be adapted to the scale and purpose of the study; it should be equal or better than the size of the smallest object of interest. *Local studies* range from meters to several tens of kilometers. They are chiefly intended to geophysical prospecting, archaeological and construction purposes. It is also a useful tool to investigate local tectonics, for example, related to volcanic activity. If possible, gravity anomalies defined at the original observation points should be used for these investigations. It is desirable to remove a regional trend from the observed anomalies. It is clear from Equation 2 that the gravity effect of any source rapidly decreases when we move this source to a greater depth. A useful approximation may be obtained from a spectral representation of the gravity field. A gravity effect of density variations at depth Z having the wavelength L is reduced by the factor:

$$\exp(-2\pi Z/L), \quad (5)$$

compared to the effect of near-surface variations.

Regional studies range from tens to first thousands of kilometers. They include various tectonic applications. The required resolution of the gravity anomalies is within approximately 5′–30′. Since the original field contains sharp variation, a special attention should be paid to the data preparation. Normally, the observed values are averaged over the compartments that correspond to the grid used in the study. *Global studies* include the whole Earth and are intended to investigate structure of the planet down to the core–mantle boundary as well as global mantle convection. The required data resolution normally ranges from 1° to 5°.

Enhancement of the gravity anomalies. A straightforward way to reduce the ambiguity of the gravity interpretation is to calculate the gravity effect of those structures, which are known from other geophysical methods, and to remove it from the observed anomalies. Then, the residual will highlight the gravity effect of "unknown" heterogeneities. *Bouguer gravity anomalies* represent the first step of the gravity "*stripping*." However, this procedure may be extended using the data on the crustal structure obtained from seismic studies. Attempts to calculate residual "crust-free" gravity anomalies have been made since the first seismic profiles were measured but a reliable 3D model could be constructed only after sufficient initial data on the crustal structure had been accumulated. One of the first maps of the residual gravity field covering a considerable part of North Eurasia was calculated by Artemjev et al. (1994).

After construction of 3D models of the crust based on seismic data, this method of the gravity interpretation has been extensively used over the last decade (e.g., Yegorova and Starostenko, 2002; Kaban et al., 2003, 2010). Detailed maps of sedimentary basins and their density structure are available from numerous prospecting surveys. A number of seismic data showing crustal structure down to the crust–mantle boundary (Mohorovičić discontinuity) has also significantly increased. Based on these data, a number of 3D crustal models with regional and global coverage have been compiled (e.g., Mooney et al., 1998; Tesauro et al., 2008). Density distribution within the crystalline crust is determined based on the relationships between seismic velocity and density (e.g., Christensen and Mooney, 1995). Then, it is possible to calculate gravity effect of the crustal layers and to remove it from the observed field. The modern calculation methods employ integral equations similar to Equation 2 (also in a spherical case) and can provide any desirable accuracy (e.g., IGMAS, http://www.gravity.uni-kiel.de/igmas/). While doing these calculations, it is important to include the regions surrounding a study area (Kaban et al., 2004). The distance, on which this effect might be important, may be estimated from Equation 5. For the continental regions with deep crust–mantle boundary this distance should be at least 500 km. One of the most recent

global maps of the residual mantle anomalies is shown in Figure 1 (Kaban et al., 2003). This map is extensively used in studies of the mantle structure and dynamics.

Forward method

A gravity model is constructed using an iterative approach. This type of interpretation is used when velocity models obtained from seismic profiles are used to constrain the density model. An initial model is constructed based on a priori data and modeling assumptions. The initial structure is based on seismic boundaries, and seismic velocities are converted to densities. The model's gravity anomaly is calculated and compared with the observed field. Then, the model parameters are adjusted using an iterative process to improve the fit between the modeled and observed anomalies. At the first stage, density of the seismic blocks is modified. If the fit is not sufficient, seismic boundaries might be changed as well. One example of such interpretation along the URSEIS profile crossing the southern Urals is shown in Figure 2 (Kimbell et al., 2002).

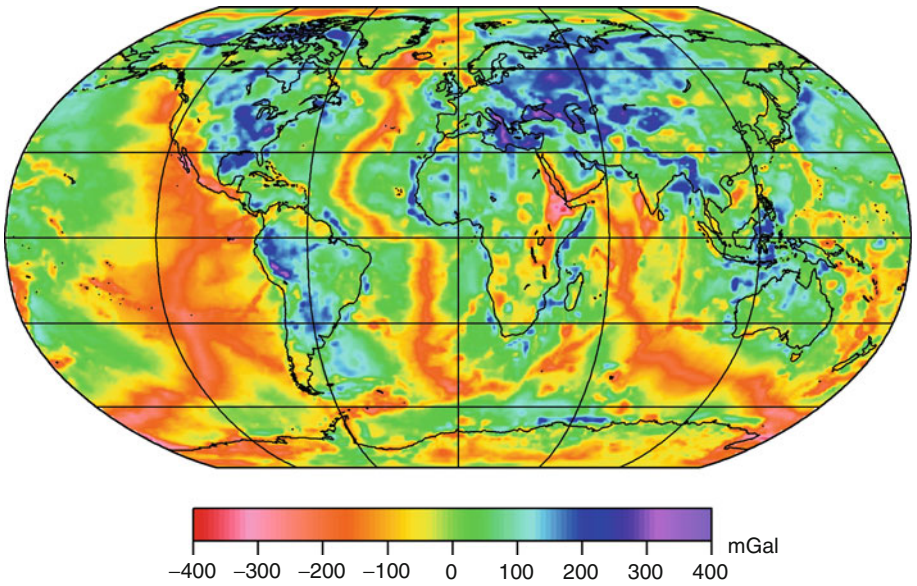
Inverse method

In the inverse method, model parameters are automatically determined by solving Equation 4. This technique is also used for joint interpretation of the gravity and seismic data. For the regional seismic profiles it was developed by Strakhov and Romanyuk (1985), whereas for global studies based on tomography data, by Hager and O'Connel (1981). The idea of the method implies use of seismic velocities to stabilize the solution of Equation 4. Instead of determining density variations in each grid point of the 3D model, a velocity–density relationship for selected blocks and layers should be introduced. This relationship may vary significantly; however, it is rather stable for similar compositions and physical conditions. Therefore, after selecting several blocks and layers, which are supposed to obey the same conditions, it is possible to determine a very limited number of the model parameters (velocity-to-density scaling factors for the specified blocks and layers). This provides the possibility to obtain a unique and stable solution of the inverse problem. A clear advantage of the inverse method with respect to the forward modeling is that it provides a possibility to quantitatively control sensitivity of the gravity anomalies to variations of the model parameters and their uncertainties.

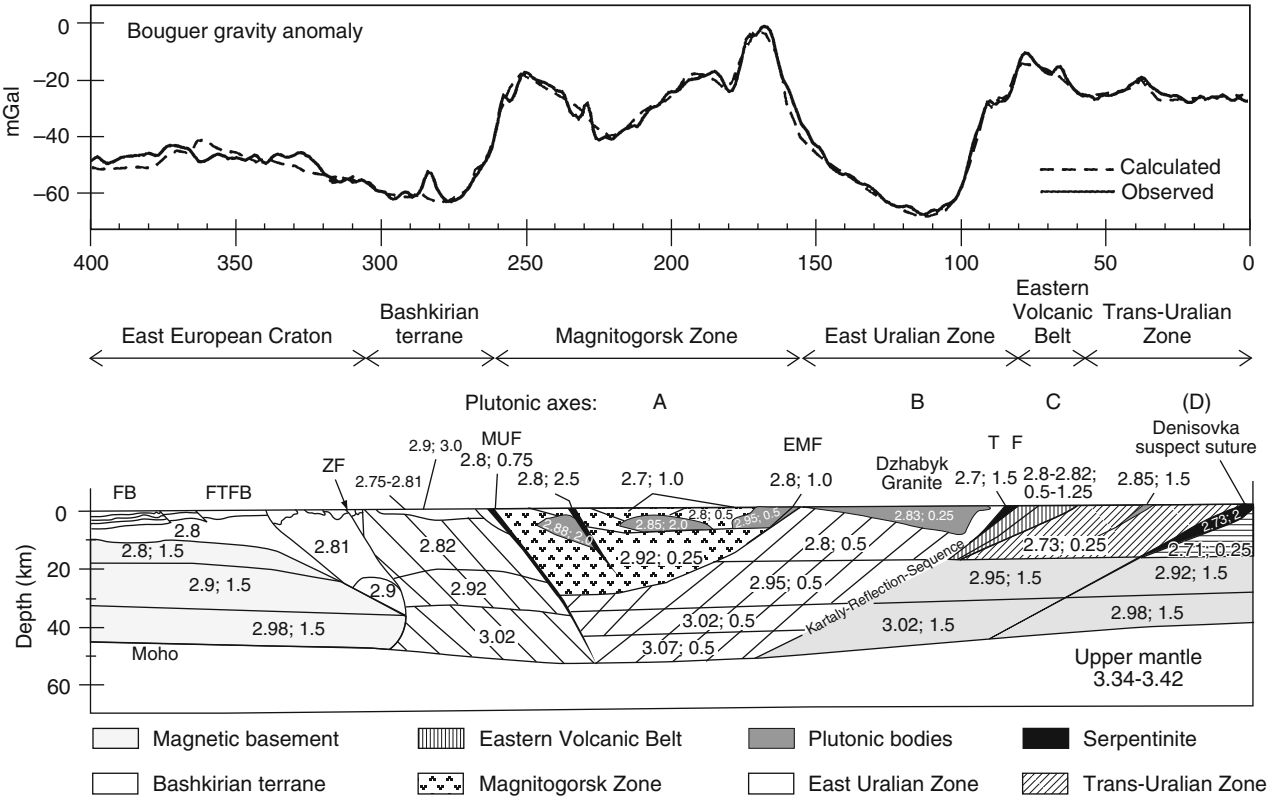
The initial density model is usually produced by a simple linear conversion of the velocity disturbances into density variations. The initial scaling factors (c_j^0) are based on mineral physics and petrological data (j is the number of the selected layer or block). In the inversion they should be rescaled in a least squares adjustment. The inversion can be performed in both space and spectral (or spherical harmonic) domains. The scaling coefficients a_j (to rescale the initially adopted values) are estimated by minimizing the function (Kaban and Schwintzer, 2001):

$$\chi^2 = \sum_i \cos\phi^i \left(\mathrm{dg}^i_{\mathrm{obs}} - \sum_j \hat{a}_j \mathrm{dg}^i_j \right)^2 + \sum_j \alpha_j (\hat{a}_j - 1)^2, \tag{6}$$

where dg^i_j is the initial gravity effect of layer or block j at the surface grid point i characterized by the gravity



Gravity Anomalies, Interpretation, Figure 1 Residual "mantle" anomalies obtained after removing the gravity effect of the crust (relative to a horizontally homogeneous reference model) from the observed field (Kaban et al., 2003).



Gravity Anomalies, Interpretation, Figure 2 2D gravity model for the URSEIS profile (Kimbell et al., 2002). The first number indicates density (Mg/m^3).

anomaly dg^i_{obs}. $\cos\varphi^i$ is a formal weight factor to account for the convergence of the meridians in the equal-angular grid. It is not required in case of the equal-space grid or spherical harmonic inversion. The last term is used to stabilize the solution. In this case it requires that the estimated scaling factors should not be significantly different from the initial ones. α controls strength of regularization. In favorable cases it could be even reduced to zero. Equation 6 leads to the normal equation system with a dimension equal to the number of the unknown scaling factors

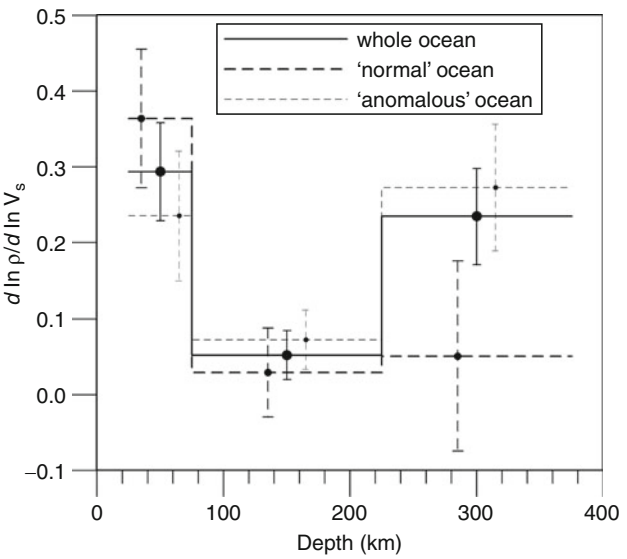
$$\mathrm{B}\hat{\mathbf{a}} = c \tag{7}$$

System (7) is solved for the vector $\hat{\mathbf{a}}$ by inversion of the normal matrix B. The solved-for parameters $\hat{a}_j$ are used to rescale the initial values. The coefficients' standard deviations $s_{\hat{c}_j}$ resulting from the fit in the least squares adjustment then are

$$s_{\hat{c}_j} = c_j^0 s_{\hat{a}_j}, \quad s_{\hat{a}_j} = \sqrt{\frac{1}{f}\chi^2 q_{jj}}\ , \tag{8}$$

where q_{jj} is the diagonal element of B^{-1}, f is the degree of freedom (Kaban and Schwintzer, 2001).

In global studies this technique is used for interpretation of both geoid and gravity anomalies. Geoid is more sensitive to the effect of deep interiors, whereas a use of



Gravity Anomalies, Interpretation, Figure 3 The velocity-to-density scaling coefficients adjusted for "normal" and "anomalous" oceanic areas, respectively, compared with the result for the whole oceanic area. Only the scaling coefficients and standard deviations for the upper mantle are shown, but the inversion was performed for the whole mantle (Kaban and Schwintzer, 2001).

the gravity anomalies provides better resolution, especially for the upper mantle (e.g., Simmons et al., 2009). An example of such study is shown in Figure 3. The velocity-to-density scaling factor is determined for the oceanic upper mantle in the joint inversion of the residual mantle gravity anomalies and seismic tomography (Kaban and Schwintzer, 2001). The obtained density–velocity ratios for the subcrustal layer (down to 75 km) and for the deep part of the oceanic upper mantle (below 225 km) are very close to the ones predicted from the mineral physics analysis. However, for the oceanic layer involving the depths 75–225 km a significantly lower density–velocity ratio is obtained. This may be explained by a compositional layering: dry, partially depleted peridotite residuum underlain by "damp" fertile peridotite and separated by the G discontinuity at a depth of about 70 km. This result is important for geodynamic modeling. It evidences that the base of an oceanic lithosphere plate is defined by a compositional rather than thermal boundary layer or, at least, the base is strongly influenced by a compositional boundary.

Summary

Due to new interpretation techniques and increased computer power, most of the computational problems related to the forward and inverse gravity modeling, which caused big problems even recently, are solved. For example, the direct computations using Equation 2 may be performed for on large grids taking into account 3D density variations. Also a principal difference from the early-stage period is that the interpretation of the gravity anomalies is performed now in combination with other methods. The most profitable results are obtained from the integrative interpretation of the gravity and seismic data. Further combination with geothermal data provides the possibility to separate the effects of temperature and composition variations within the Earth. The obtained density models are used in various applications. This information is essential for a geodynamic modeling that results in determination of the mantle flows, tectonic forces moving the lithosphere blocks, and stress distribution in the crust and mantle at both regional and global levels. Interpretation of the gravity anomalies is also widely used in geophysical prospecting. In some situations the seismic investigations fail to give results, for example, when a high velocity layer masks the underlying low-velocity layers. Here the gravity and magnetic investigations are the principal means to obtain information on the underlying structure.

Two principal problems should be a key issue for the gravity modeling in future. As it was shown, reliability of the gravity modeling depends on additional data used in the interpretation. Therefore, it is necessary to bring together data from various disciplines, which might be used in construction of integrative interpretation models. Geophysicists just started approaching this ultimate goal. The second principal issue is improvement of the technique for integrative interpretation. In particularly, special attention should be paid to solution of nonlinear inverse problems and to ultrahigh dimension problems.

Bibliography

Artemjev, M. E., Kaban, M. K., Kucherinenko, V. A., Demjanov, G. V., and Taranov, V. A., 1994. Subcrustal density inhomogeneities of Northern Eurasia as derived from the gravity data and isostatic models of the lithosphere. *Tectonophysics*, **240**, 249–280.

Blakely, R. J., 1995. *Potential Theory in Gravity and Magnetic Applications*. New York: Cambridge University Press. 441 p.

Braitenberg, C., Ebbing, J., and Götze, H.-J., 2002. Inverse modeling of elastic thickness by convolution method – The Eastern Alps as a case example. *Earth and Planetary Science Letters*, **202**, 387–404.

Christensen, N. I., and Mooney, W. D., 1995. Seismic velocity structure and composition of the continental crust: a global view. *Journal of Geophysical Research*, **100**, 9761–9788.

Hackney, R. I., and Featherstone, W. E., 2003. *Geophysical Journal International*, **154**, 35–43.

Hager, B. H., and O'Connel, R. J., 1981. A simple global model of plate dynamics and mantle convection. *Journal of Geophysical Research*, **86**, 4843–4867.

Heiskanen, W., and Moritz, H., 1967. *Physical Geodesy*. San Francisco: W.H. Freeman.

Kaban, M. K., and Schwintzer, P., 2001. Oceanic upper mantle structure from experimental scaling of V_S and density at different depths. *Geophysical Journal International*, **147**, 199–214.

Kaban, M. K., Schwintzer, P., Artemieva, I. M., and Mooney, W. D., 2003. Density of the continental roots: compositional and thermal contributions. *Earth and Planetary Science Letters*, **209**, 53–69.

Kaban, M. K., Schwintzer, P., and Reigber, Ch., 2004. A new isostatic model of the lithosphere and gravity field. *Journal of Geodesy*, **78**, 368–385.

Kaban, M. K., Tesauro, M., and Cloetingh, S., 2010. An integrated gravity model for Europe's crust and upper mantle. *Earth and Planetary Science Letters*, **296**(3–4), 195–209.

Kimbell G. S., Ayala, C., Gerdes, A., Kaban, M. K., Shapiro, V. A., and Menshikov, Y. P., 2002. Insights into the architecture and evolution of the southern and middle Urals from gravity and magnetic data. In Brown, D., et al. (eds.), *Mountain Building in the Uralides: Pangea to the Present*. Washington: American Geophysical Union, Geophysical Monograph, Vol. 13, pp. 49–65.

Mooney, W. D., Laske, G., and Masters, T. G., 1998. CRUST 5.1: a global crustal model at 5° × 5°. *Journal of Geophysical Research*, **103**, 727–747.

Silva, J. B. C., Costa, D. C. L., and Barbosa, V. C. F., 2006. Gravity inversion of basement relief and estimation of density contrast variation with depth. *Geophysics*, **71**(5), J51–J58.

Simmons, N. A., Forte, A. M., and Grand, S. P., 2009. Joint seismic, geodynamic and mineral physical constraints on three-dimensional mantle heterogeneity; implications for the relative importance of thermal versus compositional heterogeneity. *Geophysical Journal International*, **177**, 1284–1304.

Smith, W. H. F., and Sandwell, D. T., 1997. Global seafloor topography from satellite altimetry and ship depth soundings. *Science*, **277**, 1956–1961.

Strakhov, V. N., and Romanyuk, T. V., 1985. Reconstruction of the density of the Earth's crust and upper mantle from deep seismic sounding and gravimetric data; I. *Izvestiya – Physics of the Solid Earth*, **20**(6), 438–449.

Tesauro, M., Kaban, M. K., and Cloetingh, S., 2008. EuCRUST-07: a new reference model for the European crust. *Geophysical Research Letters*, **35**, doi:10.1029/2007GL032244.

Tikhonov, A. N., and Arsenin, V. Y., 1977. *Solutions of Ill Posed Problems*. New York: V. H. Winston. 258.
Yegorova, T. P., and Starostenko, V. I., 2002. Lithosphere structure of Europe and Northern Atlantic from regional three-dimensional gravity modelling. *Geophysical Journal International*, **151**, 11–31.

Cross-references

Gravity Data, Regional – Residual Separation
Gravity Data, Advanced Processing
Gravity Method, Principles
Gravity, Data to Anomalies
Isostasy

GRAVITY DATA, ADVANCED PROCESSING

Christopher J. Swain, Jonathan F. Kirby
Department of Spatial Sciences, Curtin University of Technology, Perth, WA, Australia

Synonyms

Gravity: spectral methods; Potential field transforms

Definition

Spectral analysis. Estimation and analysis of global frequency content of a signal, assumed stationary, most often using the Fast Fourier Transform (FFT).
Potential field transformations. Process of converting gravity (or magnetic) survey data into a new and physically meaningful form to facilitate its geological interpretation.
Euler and Werner deconvolution. Methods for automatically estimating depths to sources from gravity (or magnetic) survey data.
Wavelet analysis. Estimation and analysis of local frequency content of a non-stationary signal, using a wavelet transform (WT).

Introduction

The use of spectral analysis to interpret gravity anomalies goes back to the 1930s, but the modern approach in terms of the Fourier transform (FT) was developed in the 1960s and 1970s. Most of the methods are equally applicable to magnetic data, which is their more common area of use because of the huge volume of aeromagnetic data now available. Here we will not generally distinguish between the two, though the magnetics case is complicated by the fact that the field and magnetization directions are not in general the same or vertical. Blakely (1995) provides an excellent introduction to the subject, which he divides into two basic aspects: *spectral modeling* and *potential field transformations.*

In spectral modeling, the major advantage of working in the wave-number ($k = 2\pi/\lambda$) domain, after Fourier transforming the data, is that it offers simplifications and insights into the relation between anomalies and their sources. For example, the FT of the gravity anomaly over a thin horizontal layer of variable surface density $\sigma(x,y)$, at depth d, is simply the product of the FT of σ with a constant times $e^{-|k|d}$. Other models similarly separate into products of factors.

Potential field transformations can be carried out in the space domain by either numerical integration or convolution, but in the wave-number domain, the simple forms of the transformations offer considerable insight. They include upward and downward continuation, and vertical and horizontal derivatives, all of which have their uses in enhancing resolution and separating anomalies.

Both aspects were made practical by the development of the FFT, though it should be noted that there are important differences between the theoretical (continuous, infinite) FT and the (discrete, finite) FFT – see below.

The FT assumes that the statistical properties of the data do not vary spatially (i.e., the data are *stationary*), which may not be the case with real data. The wavelet transform allows spectral analysis without this assumption and therefore promises improvements over the FT.

Spectral modeling

Many applications of spectral modeling have been devised: for example, Blakely (1995) gives expressions for the FT of the anomalies for a number of simple models. Here just two of the more commonly used methods will be described.

Parker (1972) extended the simple horizontal layer model (mentioned above) to the case of a layer bounded by irregular top and bottom surfaces. This allows the very rapid calculation (via the FFT) of the gravity anomaly over a square grid for such common cases as sedimentary basins, marine Bouguer corrections from bathymetry, and Airy isostatic corrections from topography.

The method of *statistical models* was pioneered by Spector and Grant (1970) for interpreting magnetic data. In the gravity case, the method assumes that the anomalous field results from an *ensemble* of dense vertical prisms where each of the model parameters is randomly distributed with uniform probability about a mean value. As mentioned above, the FT of the anomaly of each prism is the product of several factors, notably a depth factor ($e^{-|k|d}$), a depth extent (t) factor ($1 - e^{-|k|t}$), and a size factor (a function of the width and breadth of the prism). The method employs the radial *power spectrum* of the observed anomaly, i.e., averaged over concentric annular rings in the (k_x, k_y) plane, implying that isotropy is assumed. The theoretical power spectrum for the ensemble of prisms is the product of the *ensemble averages* of the square of each of the above factors. If the ranges of the parameters are not too large, the ensemble averages will approximate the factors for their mean values. At wavelengths much larger than the body dimensions, the size factor tends to a constant. This is equivalent to assuming a dipole, rather than prism, model. Then the mean depth and depth extents can be estimated by plotting the

logarithm of the power spectrum. The depth factor gives a straight line of slope $-2d$ at medium to large k, and the depth extent factor causes this slope to decrease as k decreases, giving rise to a maximum. Thus in principle the mean values of both d and t can be estimated. The basic problem with the method lies in knowing to what extent the assumptions are justified. If the body dimensions are not small compared to the wavelengths, then it may be possible to estimate their average values, calculate a size factor, and hence correct the power spectrum. But if parameter values are not completely random and uncorrelated, but are self-similar or fractal (for which there is some evidence) then the power spectrum will be proportional to a power of k, invalidating depth estimates. See Blakely (1995) for references.

Potential field transformations

Transformations of gravity and magnetic data have a long history in applied geophysics (e.g., Grant and West, 1965), since their utility was realized long before they could be easily and accurately applied. Three examples of transformations that are commonly used in gravity interpretation are listed below with their uses.

(a) Vertical derivative: largely eliminates the regional and improves resolution.
(b) Upward continuation: very efficiently attenuates near-surface effects; used as a standard "separation filter" (Jacobsen, 1987).
(c) Downward continuation: enhances resolution, i.e., it allows the separation of closely spaced or overlapping anomalies; approximate inversion for a uniform density layer.

These operations depend on a theorem from potential theory stating that the gravity potential caused by any 3D density distribution is identical to that of a thin layer of mass spread over one of its equipotential surfaces. This is *Green's equivalent layer* (Blakely, 1995). For example in the case of upward continuation of gravity data measured on a level surface at height z_0 to a second level surface Δz above it, the equivalent layer is taken to lie in the measurement plane. The theoretical formula for the upward continued field at a point $P(x, y, z_0 + \Delta z)$ is an integral over the complete (and infinite) plane measurement surface since each point in that plane acts as an element of mass. This integral can then be written as a convolution of the data with $\Delta z/r^3$, where r is the distance from point Q in the z_0 plane to point P. Fourier transforming this expression reveals that the FT of the upward continued data is the product of the FT of the measured data and $e^{-|k|\Delta z}$, where $k = \sqrt{(k_x^2 + k_y^2)} = 2\pi/\lambda$ (radians/m). This simple view of upward continuation as a filter with wave-number response $e^{-|k|\Delta z}$ offers physical insight that explains immediately why the operation is effective at attenuating (exponentially) near-surface density anomalies.

The responses of some other common transforms are: downward continuation through height Δz, $e^{|k|\Delta z}$; vertical derivative, $|k|$; vertical integral (potential), $1/|k|$; x-derivative, ik_x. One transform that is particularly useful in mapping is the horizontal gradient amplitude (HGA): $\sqrt{[(\partial g/\partial x)^2 + (\partial g/\partial y)^2]}$ whose maxima occur vertical contacts. Another is the analytic signal amplitude (ASA): $\sqrt{[(\partial g/\partial x)^2 + (\partial g/\partial y)^2 + (\partial g/\partial z)^2]}$.

Using the FFT is much faster than either numerical integration or convolution, except for very small arrays of filter coefficients such as the centered difference approximation generally used for horizontal derivatives. It also implies certain requirements of the data (usually met by *preprocessing*) because the FFT differs from the continuous FT in important ways (e.g., Blakely, 1995). First, it is discrete, with (1D) data spacing Δx, which means that the FFT is periodic, repeating itself every $k = 2\pi/\Delta x$. This leads to aliasing of wave numbers outside the band defined by the *Nyquist* wave number $k = \pm \pi/\Delta x$ into the FFT at smaller wave numbers, unless the data are band limited. The preprocessing can include an antialias filter to ensure this. Second, the FFT is a finite summation over a data length $N\Delta x$, rather than the infinite integral of the FT, and hence implicitly assumes that the data set is periodic, infinitely repeating itself or, equivalently, circular with the last point next to the first. To avoid first order discontinuities at the data edges (which would lead to the Gibbs effect), the preprocessing then includes extending the data in such a way as to produce a smooth transition between neighboring repetitions. With these precautions, the FFT method is capable of accurate results for most transformations, unless they require amplification of wave numbers that are not present in the data, due to it's being either too short (and therefore lacking low wave numbers – below $1/N\Delta x$) or digitized with too large an interval Δx, or otherwise inaccurate.

A modern example of an application that uses a transformation via the FFT occurs in processing data from the Falcon™ airborne gravity gradiometer system, whose sensor measures the horizontal curvature components (Dransfield and Lee, 2004; Nettleton, 1976). The easiest way to convert these components to the more familiar vertical gradient is to grid them and then apply a complex transformation via the FFT.

Euler and Werner deconvolution

These are approximate inversion methods for automatically locating the depths to the tops of multiple sources. The principles of both methods are dealt with by Blakely (1995). The name is by analogy with seismic deconvolution where seismic reflection arrivals are converted into impulses representing reflectors. The methods are similar in many ways, but Werner's method in its original form is only applicable to aeromagnetic profile data, whereas the Euler method can be applied to profiles or grids of magnetic or gravity data. However, generalizations of the methods to different types of gradient data have been made and it has been shown that Werner's method using the analytic signal is in fact

a special case of Euler deconvolution (Nabighian and Hansen, 2001). Werner's method will not be described in detail, but his idea was to linearise the magnetic anomaly on a profile perpendicular to a thin dipping dyke, which contains four unknowns; hence any four measurements along the profile will give four simultaneous equations which can be solved for the position and depth of the top of the dyke.

Euler deconvolution

Euler deconvolution (Thompson, 1982; Reid et al., 1990) is very similar but uses a more general equation than that of a dyke, namely Euler's equation, which can approximately model the behavior of many different simple sources. Euler's equation is satisfied by any homogeneous function $f(x, y, z)$ of degree n (one that is invariant under scaling) and can be written:

$$x\frac{\partial f}{\partial x} + y\frac{\partial f}{\partial y} + z\frac{\partial f}{\partial z} = -nf$$

For example, $f = 1/r$ satisfies this equation with $n = 1$, implying that so does the potential of a point mass. The degree n is called the *structural index*, and defines the attenuation rate of the field f due to the particular source. Examples for gravity field anomalies are (Fitzgerald et al., 2004): point mass, 2; line mass, 1; finite step, 0; contact, −1. Reid et al. (1990) showed that certain extended bodies like thin dipping sheets (e.g., irregular dykes and sills) can also satisfy Euler's equation, but this is not generally true.

In practice, the gradients are generally calculated from a grid of gravity anomaly data as described above. In order to locate the center of a body (x_0, y_0, z_0) from survey data, it is necessary to replace (x, y, z) in Euler's equation by $(x\text{-}x_0, y\text{-}y_0, z\text{-}z_0)$. Implementation of a grid Euler method follows that of a least squares version of Werner's method, where simultaneous equations are formed from a square array (or window) of grid points of between 3 × 3 and 10 × 10 equations and a LS solution found for x_0, y_0, z_0 and a background field B (i.e., f is replace by $f\text{-}B$ on the RHS of the equation). In this case, n cannot be solved for, but must be assumed and the normal procedure is to try several values and select the one giving the most consistent results. The procedure is repeated for all points of the grid, and may generate large numbers of solutions, many of them unrealistic because the assumption of a simple source with constant n for each window is invalid. In practice, most of these solutions can be easily eliminated by selecting only those that (a) have standard deviations for x_0, y_0, z_0 (obtained from the LS solution) of <5–10% of the depth and (b) fall within the window. Another useful way to restrict the number of solutions is to apply the method only close to localized anomalies as defined, for example, by areas of positive curvature of the horizontal gradient (HGA).

Extended Euler deconvolution

For the 2D case, Mushayandebvu et al. (2001) discovered a second equation, similar to Euler's, for homogeneous functions that are also invariant under rotation. They rewrote the two equations for the particular cases of a magnetic contact and a thin dyke so that their right hand sides include terms involving the susceptibility and dip. Their *extended Euler* method finds solutions satisfying both equations that yield susceptibility and dip estimates as well as x_0, y_0, and z_0. Nabighian and Hansen (2001) then showed that the additional equation could be written in terms of the *Hilbert transform* of f (e.g., Li, 2006) which shows that it is just Euler's equation applied to the Hilbert transform. They further extended the new equation to 3D, in terms of generalized x- and y- Hilbert transforms (which can be represented and calculated as potential field transformations with wave-number domain responses $H_x = -ik_x/|k|$ and $H_y = -ik_y/|k|$, as shown by Nabighian, 1984). It has been found that, assuming a value for the structural index and using all three equations for the anomaly data and their x- and y- Hilbert transforms gives better-constrained solutions than using just one equation for the anomaly data. Moreover, the new equations allow a number of other possibilities, including solving for the structural index (Fitzgerald et al., 2004).

Tensor deconvolution

Although gradiometer data can be used with advantage in Euler deconvolution, Mikhailov et al. (2007) show that with full tensor ("FTG") data it is possible to devise a method that dispenses with data windows and locates a source beneath each measurement point based on invariants calculated from the five independent tensor components. They suggest that since this method appears to locate the centers of bodies, whereas Euler deconvolution locates their edges, the methods should complement each other.

Wavelet analysis, with advantages and limitations

Introduction

The wavelet transform (WT) was developed in the early 1980s, in an attempt to better understand the frequency content of seismic signals (e.g., Grossman and Morlet, 1984). Whereas the Fourier transform (FT) provides the amplitudes and phases of all harmonics present in the entire signal (a *global spectrum*), the WT yields these quantities at each time (or space) location of the signal (*local spectra*). For instance, the WT of Beethoven's Ninth Symphony tells one when to play C major, or when to play D minor chords. The FT, however, merely tells the musician that C major and D minor chords occur somewhere in the symphony (e.g., Antoine et al., 2004).

For this reason, the WT is superior in analyzing the harmonic constituents of *non-stationary* signals, that is, signals whose harmonics vary throughout the signal. The application to the Earth's potential fields is evident. The depth, spatial extent, and density/magnetization contrast

of sources vary widely over a given region, leading to anomaly maps with clear differences in spectral characteristics from place to place. The wavelet power spectrum, or "*scalogram*," of such a map thus quantifies the relative power of each harmonic at each spatial location; in contrast, the Fourier power spectrum shows only the total harmonic composition of the map.

The WT has two distinct forms, termed "continuous" and "discrete." Unfortunately, there is no analogy between this use of the terms and their use with the FT: as mentioned, the continuous Fourier transform (CFT) is an integral equation applied to continuous analytic formulae; the discrete Fourier transform (DFT) is its specialization to discretely-sampled signals. Both CFT and DFT, however, use the same *basis functions*, the complex exponentials (sines and cosines). However, with the discrete and continuous wavelet transforms (DWT and CWT, respectively), the names "discrete" and "continuous" apply not to the signal, but to the basis functions (the wavelets). So both CWT and DWT can be applied to discretely sampled signals, and the difference between them lies (mainly) in the wavelets.

Wavelets, whether discrete or continuous, differ from the basis functions of the FT (infinitely or quasi-infinitely extending sinusoids and cosinusoids) in that they are *localized* in space (or time). That is, while they are oscillating functions (with zero mean value), their amplitude decays rapidly with distance (or time) from their center. It is this feature which localizes the wavelet spectra to a region surrounding the point of interest. Contrast this with the DFT, where the complex exponentials maintain their amplitude over the whole study area, rendering localization impossible (though see below). However, a high degree of localization in the space (or time) domain results in a poorer degree of localization in the wave-number (or frequency) domain. This property is referred to as the *uncertainty relation* of spectral analysis, by analogy with Heisenberg's uncertainty principle of quantum mechanics. In the spectral case, the more localized the position (small Δx), the higher the uncertainty in the wave-number content of the signal at that location (large Δk); and vice versa, according to:

$$\Delta x \, \Delta k \geq 2\pi$$

For the DFT, since the Fourier transform of a sine wave is a delta-function (i.e., a spike in the frequency domain), Δk is very (infinitesimally) small and precise, whereas Δx is very (infinitely) imprecise. For the WT, there exists a trade-off between Δx and Δk according to the *scaling* of the wavelet.

It is this scaling that allows harmonic analysis. First, the WT of a given signal is a space (or time) domain convolution of the signal with scaled versions of a mother wavelet, giving *wavelet coefficients* at that scale. When the *wavelet scale* is large, the wavelets are stretched (in space or time), and the convolution yields the low-wave-number (or frequency) components of the signal; when the scale is small the wavelets are squeezed, and the convolution returns the high-wave-number (or frequency) components. However, the uncertainty relation means that large-scale wavelet coefficients tend to be spatially (or temporally) imprecise but spectrally precise, while small-scale coefficients will have the opposite properties.

The difference between the CWT and DWT lies in the choice of scales. In the DWT, the scales are chosen so that the wavelets are orthogonal. This naturally limits the range of scales to the minimum possible without loss of information; a property which is obviously preferable for image compression (e.g., Antoine et al., 2004), and means that the signal can be reconstructed very rapidly from the wavelet coefficients. The scales in the CWT, on the other hand, are not constrained, and the wavelets are hence non-orthogonal. This, though, offers redundancy in the wavelet coefficients which can be put to good use in analysis and interpretation, though reconstruction of the signal is slower.

The CWT and windowed Fourier transform

There are many kinds of discrete and continuous wavelet, each having certain properties that prove useful in a variety of applications. For instance, Morlet wavelets yield global spectra that best approximate those from the Fourier transform (Kirby, 2005), while Poisson wavelets have a close relationship with the Laplacian operator, lending themselves to potential field applications (e.g., Hornby et al., 1999). Furthermore, wavelets may be defined in N dimensions for analysis of N-dimensional signals; in geophysics, the most commonly used are 1D and 2D wavelets. Wavelets may also be complex or real-valued: 2D complex wavelets such as the Morlet wavelet are anisotropic, while 2D real wavelets, such as the Derivative of Gaussian wavelets, are isotropic. Torrence and Compo (1998) provide a description of several 1D continuous wavelets, while Kirby (2005) discusses various 2D continuous wavelets.

The windowed Fourier transform (WFT) is also often used to recover local spectra from a signal. In this, the basis functions are windowed complex exponentials; e.g., the Gabor transform uses a Gaussian window which slides over the signal. In the WFT, while the frequency of the complex exponential can increase or decrease, the window width (Δx) is fixed, meaning a fixed uncertainty in frequency, by the uncertainty relation. With the CWT, however, the window width (and hence frequency uncertainty) is variable (and directly related to wavelet scale). Hence, while the WFT has its uses, it is in general inferior to the CWT for several reasons, two of which are: (1) the WFT is not as good at analyzing signals with a large bandwidth (Foufoula-Georgiou and Kumar, 1994); and, (2) it is not as good as resolving closely spaced spikes or discontinuities (Addison, 2002).

Applications

The advantage of being able to estimate local, as opposed to global, power or phase spectra has seen the WT applied

to many disciplines, not just the Earth sciences. Indeed, a 10-s Scopus search yields over 58,000 articles with "wavelet" in the abstract, with over 4,000 published in 2009 alone. Foufoula-Georgiou and Kumar (1994), Kumar and Foufoula-Georgiou (1997), Addison (2002), and Antoine et al. (2004) discuss many. For example, Moreau et al. (1997) and Hornby et al. (1999) show the usefulness of the CWT for location and edge-detection of sources within potential field data, and subsequent inversion of the wavelet coefficients, by considering the vertical variation of the field and its horizontal derivatives. Ridsdill-Smith and Dentith (1999) also tackle derivatives, but they use the DWT which enables localized frequency-filtering. In another application of the CWT, Panet et al. (2006) reveal density anomalies below Pacific island volcanoes from the new satellite gravity data. And Kirby and Swain (2009) use the CWT to estimate spatial variations in lithospheric elastic thickness via a coherency analysis in the wavelet domain. Finally, there are several types of spherical, rather than planar, wavelet transform, which lend themselves to studies in geodesy (e.g., Klees and Haagmans, 2000; Keller, 2004; Chambodut et al., 2005).

Summary

This contribution outlines the theory and practice of a number of methods for rapidly processing gravity data that are in the form of a grid. The term "processing" here includes transformations, spectral estimation and modeling, and "deconvolution" to produce depth estimates. These are all widely used methods that are equally applicable to grids of aeromagnetic data. We also introduce the continuous wavelet transform which has many potential applications in gravity processing, few of which appear to have been widely used as yet.

A more extensive review of the gravity method, with comprehensive reference list, can be found in Nabighian et al. (2005).

Bibliography

Addison, P. S., 2002. *The Illustrated Wavelet Transform Handbook*. Bristol: Institute of Physics.

Antoine, J.-P., Murenzi, R., Vandergheynst, P., and Ali, S. T., 2004. *Two-dimensional Wavelets and Their Relatives*. Cambridge: Cambridge University Press.

Blakely, R. J., 1995. *Potential Theory in Gravity and Magnetic Applications*. Cambridge: Cambridge University Press.

Chambodut, A., Panet, I., Mandea, M., Diament, M., Holschneider, M., and Jamet, O., 2005. Wavelet frames: an alternative to spherical harmonic representation of potential fields. *Geophysical Journal International*, **163**, 875–899.

Dransfield, M. H., and Lee, J. B., 2004. The Falcon® airborne gravity gradiometer survey systems, in R.J.L. Lane, ed. Airborne Gravity 2004 - Abstracts from the ASEG-PESA Airborne Gravity 2004 Workshop: *Geoscience Australia Record* 2004/18, 15–19.

FitzGerald, D., Reid, A., and McInerney, P., 2004. New discrimination techniques for Euler deconvolution. *Computers and Geosciences*, **30**, 461–469.

Foufoula-Georgiou, E., and Kumar, P. (eds.), 1994. *Wavelets in Geophysics*. San Diego: Academic.

Grant, F. S., and West, G. F., 1965. *Interpretation Theory in Applied Geophysics*. New York: McGraw Hill.

Grossman, A., and Morlet, J., 1984. Decomposition of Hardy functions into square integrable wavelets of constant shape. *SIAM Journal on Mathematical Analysis*, **15**, 723–736.

Hornby, P., Boschetti, F., and Horowitz, F. G., 1999. Analysis of potential field data in the wavelet domain. *Geophysical Journal International*, **137**, 175–196.

Jacobsen, B. H., 1987. A case for upward continuation as a standard separation filter for potential field maps. *Geophysics*, **52**, 1138–1148.

Keller, W., 2004. *Wavelets in Geodesy and Geodynamics*. Berlin: de Gruyter.

Kirby, J. F., 2005. Which wavelet best reproduces the Fourier power spectrum? *Computers and Geosciences*, **31**, 846–864.

Kirby, J. F., and Swain, C. J., 2009. A reassessment of spectral T_e estimation in continental interiors: the case of North America. *Journal of Geophysical Research*, **114**, B08401, doi:10.1029/2009JB006356.

Klees, R., and Haagmans, R. (eds.), 2000. *Wavelets in the Geosciences*. Berlin: Springer.

Kumar, P., and Foufoula-Georgiou, E., 1997. Wavelet analysis for geophysical applications. *Reviews of Geophysics*, **35**, 385–412.

Li, X., 2006. Understanding 3D analytic signal amplitude. *Geophysics*, **71**, L13–L16, doi:10.1190/1.2184367.

Mikhailov, V., Pajot, G., Diament, M., and Price, A., 2007. Tensor deconvolution: a method to locate equivalent sources from full tensor gravity data. *Geophysics*, **72**, 161–169, doi:10.1190/1.2749317.

Moreau, F., Gibert, D., Holschneider, M., and Saracco, G., 1997. Wavelet analysis of potential fields. *Inverse Problems*, **13**, 165–178.

Mushayandebvu, M. F., Lesur, V., Reid, A. B., and Fairhead, J. D., 2001. Magnetic source parameters of two-dimensional structures using extended Euler deconvolution. *Geophysics*, **66**, 814–823.

Nabighian, M. N., 1984. Toward a three-dimensional automatic interpretation of potential field data via generalized Hilbert transforms: Fundamental relations. *Geophysics*, **49**, 780–786.

Nabighian, M. N., 2005. Historical development of the gravity method in exploration. *Geophysics*, **70**, 63ND–89ND, doi:10.1190/1.2133785.

Nabighian, M. N., and Hansen, R. O., 2001. Unification of Euler and Werner deconvolution in three dimensions via the generalised Hilbert transform. *Geophysics*, **66**, 1805–1810.

Nettleton, L. L., 1976. *Gravity and Magnetics in Oil Prospecting*. New York: McGraw-Hill.

Panet, I., Chambodut, A., Diament, M., Holschneider, M., and Jamet, O., 2006. New insights on intraplate volcanism in French Polynesia from wavelet analysis of GRACE, CHAMP, and sea surface data. *Journal of Geophysical Research*, **111**, B09403, doi:10.1029/2005JB004141.

Parker, R. L., 1972. The rapid calculation of potential anomalies. *Geophysical Journal of the Royal Astronomical Society*, **31**, 447–455.

Reid, A. B., Allsop, J. M., and Granser, H., 1990. Magnetic interpretation in three dimensions using Euler deconvolution. *Geophysics*, **55**, 80–91.

Ridsdill-Smith, T. A., and Dentith, M. C., 1999. The wavelet transform in aeromagnetic processing. *Geophysics*, **64**, 1003–1013.

Spector, A., and Grant, F. S., 1970. Statistical models for interpreting aeromagnetic data. *Geophysics*, **35**, 293–302.

Thompson, D. T., 1982. EULDPH: a new technique for making computer-assisted depth estimates from magnetic data. *Geophysics*, **47**, 31–37.

Torrence, C., and Compo, G. P., 1998. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, **779**, 61–78.

Cross-references

Gravity Anomalies, Interpretation
Gravity Data, Regional – Residual Separation
Magnetic Data Enhancements and Depth Estimation

GRAVITY DATA, REGIONAL – RESIDUAL SEPARATION

Kumarendra Mallick[1], Anthwar Vasanthi[1],
Krishna Kant Sharma[2]
[1]National Geophysical Research Institute, Hyderabad, Andhra Pradesh, India
[2]Department of Applied Geology, University of Madras, Chennai, Tamilnadu, India

Synonyms

Gravity effect of shallower and deeper bodies; Isostatic anomaly

Definition

Free-air correction. Accounts for the effect of elevation between the observation point and the datum.
Bouguer correction. Accounts for the effect of material between the observation point and the datum.
Observed Bouguer gravity anomaly. Gravity field after corrections for instrumental drift, free-air, Bouguer, terrain, and latitude.
Regional Bouguer gravity anomaly. Gravity anomaly due to deeper and broader structures.
Residual Bouguer gravity anomaly. Bouguer gravity anomaly due to shallower bodies.
Isostatic anomaly. Bouguer gravity anomaly corrected for isostatic crustal root.
Interpolating or shape function. The weighting functions (coefficients) used to compute the regional Bouguer gravity anomaly.

Introduction

The observed Bouguer gravity anomaly field (g_{obs}) consists of two components – a regional (g_{reg}) and a residual (g_{res}) that can be expressed by a simple relation:

$$g_{obs} = g_{reg} + g_{res}$$

Commonly, the regional Bouguer gravity anomaly is the longer wavelength field due to deep sources, whereas the residual Bouguer gravity anomaly corresponds to short wavelength fields of shallower bodies. The short wavelength fields die off with depth.

However, in practice, the terminology of a regional gravity anomaly varies according to the target of the investigation, and therefore does not necessarily refer to the fields generated at crustal or lower crustal depths. If the target is a mineral deposit located at the depths of 100–200 m, the basement rocks will form the regional structure. In case of hydrocarbon accumulation, the target sources are a few kilometers deep, and the regional field is generated by the rocks at the base of the sedimentary columns. In the studies involving large-scale structures like subduction zones, large sedimentary basins, regions of meteorite impact, etc., the regional fields originate at lower crustal depths or still greater depths at Moho.

The regional and residual, therefore, are relative terms. These vary from problem to problem, and from place to place. The early workers like Griffin (1949), Elkins (1951), Nettleton (1954), Roy (1961), Steenland (1962), and Skeels (1967) considered regional and residual from mineral and hydrocarbon exploration viewpoint. For a regional gravity survey on a continental scale, the concept of isostasy is evoked (Jachens and Griscom, 1985; Blakely, 1995; Chapin, 1996) to compute regional and isostatic anomalies.

In problems where the target bodies occur at shallower depths, the deeper regional field masks the response of the shallower targets and effectively acts as a noise. On the other hand, when the target body lies at greater depths, the response of shallower bodies becomes noise.

The above discussion points to the fact that it is important to separate regional and residual gravity anomalies. In effect, regional and residual gravity separation is a filtering process, and it enhances the signal-to-noise ratio. This step is essential for a precise quantitative interpretation.

It may be noted that the gravity method was first used in geodesy. The geodesists assumed the earth to be an ellipsoid for the processing of the gravity observations. Later on, the geophysicists used geoid as the reference surface for processing of gravity data. Therefore, gravity anomaly and the Bouguer field used by geophysicists is different from that used by geodesists. The former uses geoid whereas the latter the ellipsoid as the reference surface. Of late, there is an attempt to define gravity anomaly that can be used by both the groups (Hackney and Featherstone, 2003; Hinze et al., 2005). The processing techniques are valid for both the datums. Here, the Bouguer gravity anomaly is considered from the viewpoint of the geophysicists.

Regional–residual separation

The regional–residual separation is essentially a filtering process and is carried out broadly by two techniques: graphical and analytical. The graphical method is not used these days anymore but it has historical importance. A variety of techniques are used in the analytical method.

Analytical methods

The analytical techniques are developed both in space and frequency domains. Starting from simple four-point and nine-point averaging, Griffin's center-point-and-one-ring (1949), Agocs' least-squares (1951), Grant's statistical (1954) to polynomial fitting (Coons et al., 1967; Rao et al., 1975; El-Batroukh and Zentani, 1980; Belthrao

et al., 1991; Agarwal and Shivaji, 1992; Leao et al., 1996) and second-derivative techniques are in space domain. Mickus et al. (1991) have developed an algorithm using the minimum curvature technique to separate the regional and residual components.

Another approach is to build a density model of the deep-seated structures with parameters constrained by independent methods like seismics, aeromagnetics, fractals, spectrals, etc., and to compute their gravity effects. The fields generated by the deeper sources are subtracted from the observed Bouguer gravity field to arrive at the residual gravity anomalies. This approach developed at USGS employs the theory of isostasy, and is frequently used to compute isostatic residual anomalies (Jachens and Griscom, 1985; Blakely, 1995; Lowrie, 1997). A new deterministic method based on Airy-Heiskanen isostatic model had been developed by Chapin (1996). This technique assumes that the topography is fractal. The algorithm used in this technique inverts the present-day elevations to determine the depth to the base of the crust. This depth is then used as input to a three-dimensional gravity model that computes the gravity response (regional) of the crust. The regional is then subtracted from the Bouguer gravity to yield the final isostatic residual.

In the frequency domain, the high-pass, low-pass, and band-pass filtering techniques (Naidu, 1968; Ulrich, 1968; Cordell, 1985; Pawlowski, 1994) and upward and downward continuations (Roy, 1966) have been developed for regional and residual separation. Green's concept of an equivalent-source layer is invoked by Pawlowski (1994) to construct a data-adaptive, zero phase, Wiener band-pass filter for regional–residual separation (Pawlowski, 1994). The observed gravity field's Fourier power spectrum is modeled with two Green's equivalent source layers, one equivalent layer for the shallower (residual field) geological sources and another for the deeper (regional field) geological features.

Possible drawbacks in regional computation

In all the above-mentioned analytical techniques, the Bouguer gravity values of the target source enter into the computations. Many investigators (Skeels, 1967; Zurfluch, 1967; Ulrich, 1968; Leao et al., 1996; Ojo and Kangkolo, 1997) have discovered that these techniques are not free from inaccuracies. In Griffin's (1949) center-point-and-one-ring method, the center point of the computation moves over the entire map space, including the anomalous target zone. When the gravity values are picked up from one such anomalous target zone for averaging, a small residual still remains mixed with the regional. Besides this, the optimum choice of radius is crucial to arrive at proper residual. Leao et al. (1996) pointed out that "any attempt to model a complex regional by a higher-order polynomial will force the polynomial to represent part of the residual field: thereby distorting the estimated regional." In case of two most used techniques, both least-squares (Agocs, 1951; Coons et al., 1967; El-Batroukh and Zentani, 1980; Leao et al., 1996) and spectral methods (Naidu, 1968; Ulrich, 1968; Cordell, 1985; Klasner et al., 1985; Pawlowski, 1994), the residual is assumed to be zero-mean. Ulrich (1968) and Leao et al. (1996) observed that a departure from this assumption may contaminate the residual by pseudo anomalies.

New approach based on finite element analysis (FEA)

In order to overcome the inaccuracies mentioned above, a new space-domain operator, based on finite element concept, has recently been developed by Mallick and Sharma (1997, 1999) and applied to a variety of research problems (Sharma et al., 1999; Mallick et al., 1999; Mallick and Vasanthi, 2001; Vasanthi and Mallick, 2001; Kannan and Mallick, 2003; Vasanthi and Mallick, 2005; Vasanthi et al., 2006). The method is altogether different from all the existing techniques in the sense that the observed gravity values over the target zones in the map space do not enter into the regional computations. This technique is robust, well performing, and not site specific and overcomes most of the drawbacks encountered in the existing analytical methods.

Figure 1a shows an eight-node element in Cartesian coordinates. The center is located at x_c, y_c with sides 2a and 2b parallel to x- and y- axes, respectively. The nodes 1 through 8 are shown by filled circles. When a gravity survey is superimposed by one such element (or interconnected elements), the regional anomaly can conveniently be approximated at any point (x, y) inside the element by

$$g_d(x,\ y) = \sum_{i=1}^{8} N_i(x,\ y) g_i \tag{1}$$

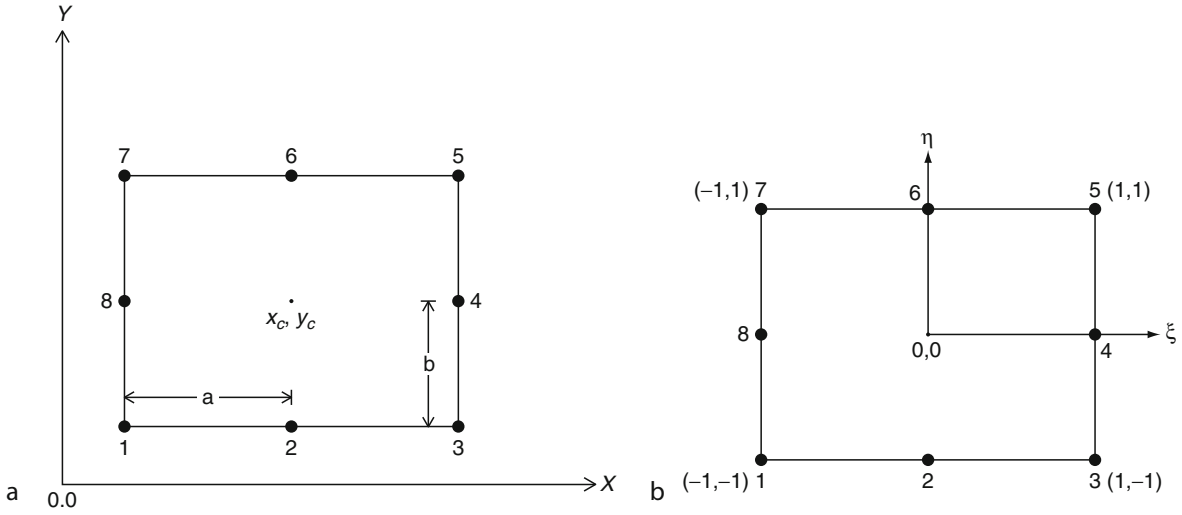
where d stands for the deeper structures, N_i (x, y) are the weights or the shape functions (Cheng and Yeo, 1979) of the element, and g_i represent the nodal gravity values. Although here an eight-node quadratic element is described, a linear element with four nodes or a cubic element with 12 nodes may also be considered. Instead of working in x–y Cartesian coordinate system, it is simpler to carry out the computations in a reference element defined by a nondimensional ξ–η space. This is achieved by the following simple substitution,

$$\xi = \frac{(x - x_c)}{a},\ \eta = \frac{(y - y_c)}{b} \tag{2}$$

These nondimensional coordinates ξ and η vary between -1 and 1. Figure 1b shows the reference element with eight nodes. The regional at any point (ξ, η) inside and on the sides of the reference element is rewritten as

$$g_d(\xi, \eta) = \sum_{i=1}^{8} N_i(\xi, \eta) g_i \tag{3}$$

The nodal gravity values g_i in Figure 1a are reassigned to the corresponding nodes in the reference element. The



Gravity Data, Regional – Residual Separation, Figure 1 (**a**) Eight-node element in *x-y* real space. The center is located at x_c, y_c with sides 2a and 2b parallel to *x*- and *y*- axes, respectively, (**b**) Eight-node element in ξ–η reference space.

8 nodal values in the periphery of the gravity survey represent the regional field.

Equation 3 is an important step in the development of the computational scheme for the following reasons. First, since ξ and η are nondimensional, the reference element can represent an observed gravity map of any size and any orientation. Second, only the nodal gravity values, g_i are used to compute the regional. No other gravity values from the map space enter into the computation, thereby overcoming the effects of the target zones. It may be noted here that all other techniques, both in space and frequency domains, developed so far over the past six decades for the regional computation do use gravity values lying inside the survey space. Some of these internal gravity values are, in fact, due to the target bodies. Therefore, there is a possibility of the regional anomaly getting contaminated by a portion of residual anomaly. Third, the weighting functions N_i (ξ, η) are defined without ambiguity. The strength of this approach lies in its simplicity.

Application to gravity surveys

The efficacy of the new technique was put to test in processing two potential fields – the gravity map of proven oil field structures in western Texas (Coffin, 1950) chosen from the distant past and the other from the large-scale structure covering parts of Klamath Mountains and Cascade Range in north-central California (Blakely, 1995).

Hydrocarbon prospect in western Texas

Figure 2a shows a gravity survey in western Texas (Coffin, 1950). By direct drilling the regions marked A and B have been proved to be oil-bearing. However, over both A and B, there are no strong anomalies. The deflections in gravity contours over regions A and B give slight indications of the presence of the structures. It is evident that a strong NW–SE regional field masked the gravity effects of the shallower structures.

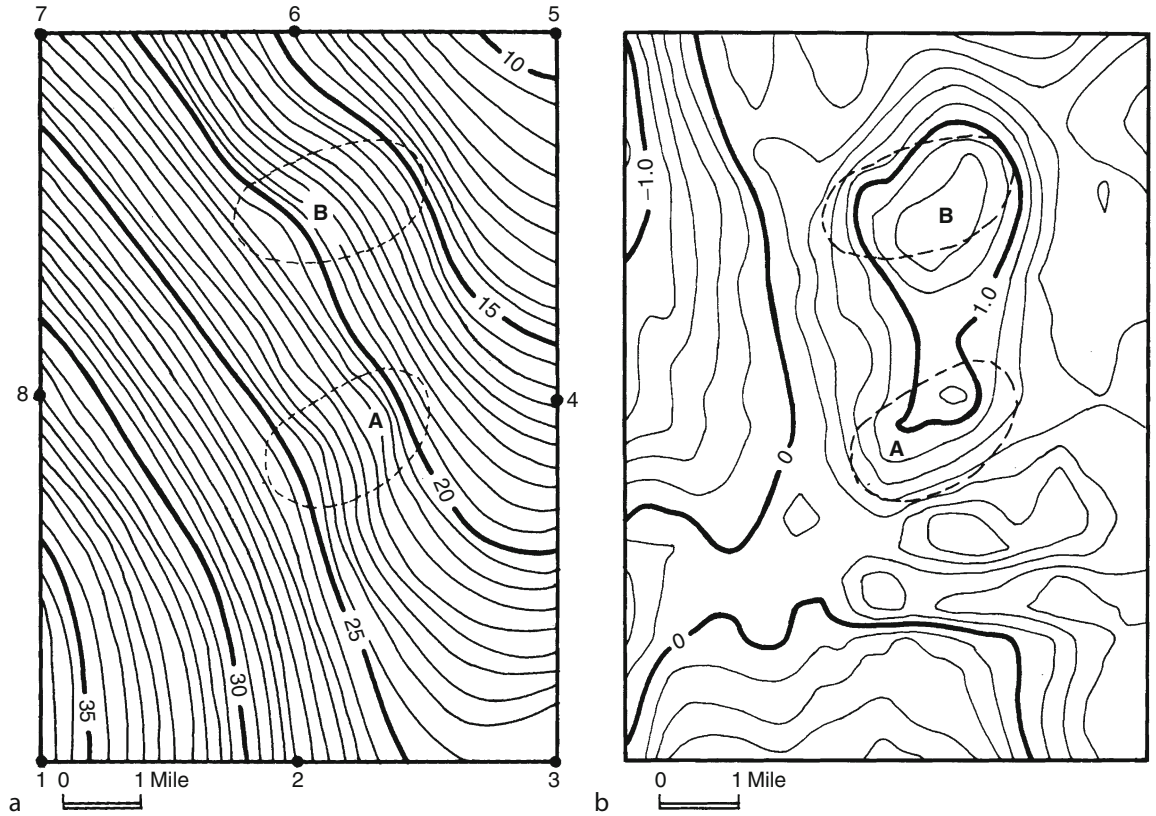
In the next step, the regional trend has been removed from the observed Bouguer gravity field by the present procedure to arrive at a residual gravity field in Figure 2b. In order to achieve this, the observed gravity values at nodes 1, 2, 3, . . ., 8 were used to compute the regional field and then the residual gravity anomalies. The residual map clearly brings out two contour closures over the target responses at A and B. This example makes it clear that the regional trend does mask the target response. The contours at B are more spread out since the target here lies deeper than that at region A.

Klamath mountains and cascade range (model-based space domain)

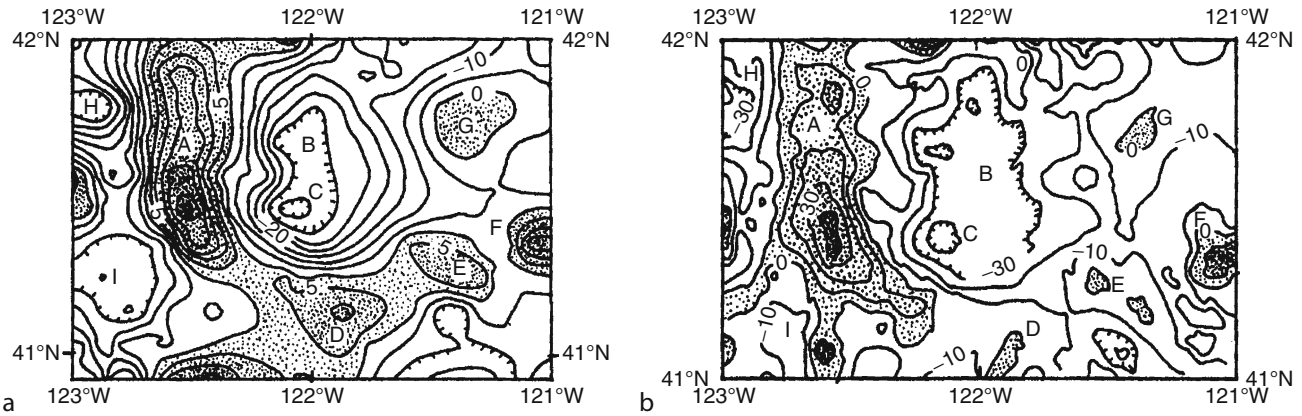
A large-scale gravity survey covering parts of Klamath Mountains and Cascade Range in north-central California lying between lat. 40° 52′ 30″ and 42°N, and long. 121° and 123°W has been taken up.

In order to compute the isostatic residual anomaly, Jachens and Griscom (1985) had adopted Airy-Heiskanen isostatic model and made the following assumptions: (1) A three-layer crust with uniform densities, (2) The density of the top layer, $\rho = 2{,}670$ kg/m^3, (3) Sea-level crustal thickness, $T = 25$ km, obtained by seismic refraction surveys, (4) The crust–mantle density contrast, $\rho = 400$ kg/m^3, close to the density contrast, $\rho = 385$ kg/m^3 (Woollard, 1966) obtained by extensive seismic studies, and more than 15,000 density measurements.

In order to cross-check their results, Jachens and Griscom (1985) had also examined the regional for this region based on Pratt-Hayford model response with a depth of compensation of 113.7 km. This regional did



Gravity Data, Regional – Residual Separation, Figure 2 (**a**) Bouguer gravity map of western Texas (After Coffin, 1950). A and B, shown by *broken line*, are two oil fields proved by direct drilling. Contour interval: 1 mGal. (**b**) Finite element residual gravity map of western Texas.



Gravity Data, Regional – Residual Separation, Figure 3 (**a**) Finite element residual gravity map of the Klamath Mountains and Cascade Range lying between lat. 40°52′30″ and 42°N, and long. 121° and 123°W. Contour interval: 5 mGal. (**b**) Isostatic residual map of the Klamath Mountains and Cascade Range lying between lat. 41° and 42° N, and long. 121° and 123° W (After Blakely, 1995). Contour interval: 10 mGal.

not differ appreciably from that obtained by Airy-Heiskanan isostatic model.

Over the same region, the finite element scheme was employed to compute the regional and the residual gravity anomalies. In the present computational scheme, there were no such constraints with regard to density and crustal thickness. This makes the new technique unique.

Figure 3a shows the finite element residual anomalies. The Klamath Mountains, Cascade Range, and all other important geological features are evident in this figure. For a comparison, the model-based isostatic residual map (Blakely, 1995, Figure 7.21f) is shown in Figure 3b. The FEA residual anomalies in Figure 3a, A through I appear to be better than or comparable to the corresponding model-based anomalies in Figure 3b.

Summary

The separation of regional and residual components from the observed Bouguer gravity (and magnetic) anomaly has remained a subject of debate over the past six to seven decades. Starting from simple graphical to model-based estimation of regional, both in space and frequency domain, the expected uniqueness has not been achieved. This is due to human judgment and mathematical as well as geological assumptions. With this backdrop, a new concept based on finite element theory is described that performs as good as or often better than the existing techniques both in frequency and space domains. This has been demonstrated with two examples. FEA technique is not site specific and can be universally used.

Bibliography

Agarwal, B. N. P., and Shivaji, C. H., 1992. Separation of regional and residual anomalies by least-squares orthogonal polynomial and relaxation techniques: a performance evaluation. *Geophysical Prospecting*, **40**, 143.

Agocs, W. B., 1951. Least-squares residual anomaly determination. *Geophysics*, **16**, 686.

Belthrao, J. F., Silva, J. B. C., and Costa, J. C., 1991. Robust polynomial fitting method for regional gravity estimation. *Geophysics*, **56**, 80.

Blakely, R. J., 1995. *Potential Theory in Gravity and Magnetic Applications*. Cambridge: Cambridge University Press.

Chapin, A. D., 1996. A deterministic approach toward isostatic gravity residuals – A case study from South America. *Geophysics*, **61**, 1022.

Cheng, Y. K., and Yeo, M. F., 1979. *A Practical Introduction to Finite Element Analysis*. London: Pitman.

Coffin, R. C., 1950. Geological imagination in the interpretation of geological data. *Quarterly of the Colorado School of Mines*, **45**, 11.

Coons, R. L., Woollard, G. P., and Hershey, G., 1967. Structural significance and analysis of Mid-Continent gravity high. *American Association of Petroleum Geologists Bulletin*, **51**, 2381.

Cordell, L., 1985. A stripping filter for potential data, *55th Annual International Meeting of Society of Exploration Geophysicists*, SEG, Extended Abstracts: pp. 217–218.

El-Batroukh, S. I., and Zentani, A. S., 1980. Gravity interpretation of Raguba field, Sirte Basin, Libya. *Geophysics*, **45**, 1153.

Elkins, T. A., 1951. The second derivative method of gravity interpretation. *Geophysics*, **16**, 29.

Grant, F. S., 1954. A theory for the regional correction of potential data. *Geophysics*, **19**, 23.

Griffin, W. P., 1949. Residual gravity in theory and practice. *Geophysics*, **15**, 39.

Gupta, V. K., and Ramani, N., 1980. Some aspects of regional-residual separation of gravity anomalies in Precambrian terrain. *Geophysics*, **45**, 1412.

Hackney, R. I., and Featherstone, W. E., 2003. Geodetic versus geophysical prospectives of the 'gravity anomaly'. *Geophysical Journal International*, **154**, 35.

Hinze, W. J., Aiken, C., Brozena, J., Coakley, B., Dater, D., Flanagan, G., Forsberg, R., Hildenbrand, T., Kellar, R., Kellogg, J., Kucks, R., Li, X., Mainville, A., Morin, R., Pilkington, M., Plouff, D., Ravat, D., Roman, D., Jamie, U. F., Veronneau, M., Webring, M., and Winester, D., 2005. New standards for reducing gravity data: the North American gravity database. *Geophysics*, **70**, J125.

Jachens, R. C., and Griscom, A., 1985. An isostatic anomaly maps of California – a residual map for interpretation of anomalies from infrastructural sources. In Hinze, W. J. (ed.), *The Utility of Regional Gravity and Magnetic Anomaly Maps*. Tuisa: Society of Exploration Geophysicists, pp. 347–360.

Kannan, S., and Mallick, K., 2003. Accurate regional–residual separation by finite element approach, Bouguer gravity of a Precambrian mineral prospect in northwestern Ontario. *First Break*, **21**, 39.

Klasner, J. S., King, E. R., and Jones, W. J., 1985. Geologic interpretation of gravity and magnetic data for northern Michigan and Wisconsin. In Hinze, W. J. (ed.), *The Utility of Regional Gravity and Magnetic Anomaly Maps*. Tuisa: Society of Exploration Geophysicists, pp. 267–286.

Leao, J. W. D., Menezes, P. T. L., Beltrao, J. F., and Silva, J. B. C., 1996. Gravity inversion of basement relief constrained by the knowledge of depth at isolated points. *Geophysics*, **61**, 1702.

Lowrie, W., 1997. *Fundamentals of Geophysics*. Cambridge: Cambridge University Press.

Mallick, K., and Sharma, K. K., 1997. Computation of regional gravity anomaly – a novel approach. *Proceedings of Indian Academic Sciences (Earth Planetary Sciences)*, **106**, 55.

Mallick, K., and Sharma, K. K., 1999. A finite element method for computation of the regional gravity anomaly. *Geophysics*, **64**, 461.

Mallick, K., and Vasanthi, A., 2001. Delineation of subsurface extent of Perinthatta anorthosite, Kerala, India. *Journal of the Geological Society of India*, **57**, 321.

Mallick, K., Sharma, K. K., and Rao, V. K., 1999. A new gravity interpretation: a case study from Pahute Mesa, Nevada test site. *Current Science*, **76**, 1495.

Mickus, K. L., Aiken, C. L. V., and Kennedy, W. D., 1991. Regional–residual gravity anomaly separation using the minimum curvature technique. *Geophysics*, **56**, 279.

Naidu, P. S., 1968. Spectrum of the potential field due to randomly distributed source. *Geophysics*, **33**, 337.

Nettleton, L. L., 1954. Regionals, residuals and structures. *Geophysics*, **19**, 1.

Ojo, S. B., and Kangkolo, R., 1997. Shortcomings in the determination of regional fields by polynomial fitting: a simple solution. *Journal of Applied Geophysics*, **36**, 205.

Pawlowski, R. S., 1994. Green's equivalent-layer concept in gravity band-pass filter design. *Geophysics*, **59**, 69.

Rao, B. S. R., Murthy, I. V. R., and Rao, V. C., 1975. A successive approximation method of deriving residual gravity. *Geoexploration*, **13**, 129.

Roy, A., 1961. On some properties of residual and derivatives. *Journal of Geophysical Research*, **66**, 543.

Roy, A., 1966. The method of continuation in mining geophysical interpretation. *Geoexploration*, **3**, 65.

Sharma, K. K., Rao, V. K., and Mallick, K., 1999. Finite element gravity regional and residual anomalies and structural fabrics of northwest Ganga basin. *Journal of the Geological Society of India*, **54**, 169.

Skeels, D. C., 1967. What is residual gravity? *Geophysics*, **32**, 872.

Steenland, N. C., 1962. Gravity and aeromagnetic explorationin the Paradox basin. *Geophysics*, **27**, 73.

Ulrich, T. J., 1968. Effect of wavelength filtering on the shape of the residual anomaly. *Geophysics*, **33**, 1015.
Vasanthi, A., and Mallick, K., 2001. Occurrence patterns of Kimberlite pipes based on gravity and magnetic anomalies in Wajrakarur region, A.P. *Journal of the Geological Society of India*, **58**, 251.
Vasanthi, A., and Mallick, K., 2005. Bouguer gravity anomalies and occurrence patterns of Kimberlite pipes in Narayanpet-Maddur regions, Andhra Pradesh, India. *Geophysics*, **70**, 13.
Vasanthi, A., Sharma, K. K., and Mallick, K., 2006. Discussion and reply on "New standards for reducing gravity data: the North American gravity database (Hinze, W. J. et al., 2005, *Geophysics*, **70**, J25–J32). *Geophysics*, **71**, 31.
Woollard, G. P., 1966. Regional isostatic relations in the United States. In Steinhart, J. S., and Smith, T. J. (eds.), The Earth beneath the continents. Geophysical Monograph, Vol. 10, p. 557.
Zurfluch, E. G., 1967. Application of two-dimensional linear wavelength filtering. *Geophysics*, **32**, 1015.

Cross-references

Magnetic Anomalies, Interpretation

GRAVITY FIELD OF THE EARTH

Christopher Jekeli
Division of Geodetic Science, School of Earth Sciences, Ohio State University, Columbus, OH, USA

Definition and scope

The gravitational field of the Earth is generally understood as the field generated by the masses of the terrestrial body. However, actual measurements made on the surface include tidal components due to the sun and moon (and theoretically other planets) and the atmosphere, as well as the centrifugal acceleration due to Earth's rotation. Therefore, geodesists distinguish between terrestrial gravitation (mass attraction of the Earth) and gravity (gravitation plus centrifugal acceleration), where tidal and atmospheric effects are treated as corrections. This chapter develops the basic physical concepts of terrestrial gravitation assumed to be generated solely by the subsurface mass density distribution. After an introduction to the Newtonian gravitational potential and its properties, the subsequent sections of the chapter discuss the spatial variation of the field on and above the Earth's surface, the geophysical interpretation of the low-degree harmonics of the field, and briefly the variation of the field inside to the Earth.

Introduction

Historical notes

A study of Earth's gravitational field is a study of Earth's mass structure, its dynamical influence on satellites, and lately its mass redistribution and transport in time. It is also fundamentally and historically a geodetic study of Earth's shape, described largely (70%) by the surface of the oceans (and their conceptual continuation under the continents). Gravitation is a physical phenomenon so pervasive and incidental that humankind generally has taken it for granted with scarcely a second thought. The Greek philosopher Aristotle (384–322 BC) allowed no more than to assert that gravitation is a natural property of material things that causes them to fall (or rise, in the case of some gases), and the more material the greater the tendency to do so. It was enough of a self-evident explanation that it was not yet to receive the scrutiny of the scientific method. Almost 2,000 years later Galileo Galilei (1564–1642) finally took up the challenge to understand gravitation through observation and scientific investigation. His experimentally derived law of falling bodies corrected the Aristotelian view and divorced the effect of gravitation from the mass of the falling object – all bodies fall with the same acceleration. This truly monumental contribution to physics was, however, only a local explanation of how bodies behaved under gravitational influence. Johannes Kepler's (1571–1630) observations of planetary orbits pointed to other types of laws, principally an inverse-square law according to which bodies are attracted by forces that vary with the inverse square of distance. The genius of Issac Newton (1642–1727) brought it all together in his *Philosophiae Naturalis Principia Mathematica* of 1687 with a single and simple all-embracing law that in one bold stroke explained the dynamics of the entire universe (today there is more to understanding the dynamics of the cosmos, but Newton's law remarkably holds its own). The mass of a body was again an essential aspect, not as a self-attribute as Aristotle had implied, but as the source of attraction for other bodies: each material body attracts every other material body according to a very specific rule (Newton's law of gravitation). Newton regretted that he could not explain exactly why mass has this property (as one still yearns to know today within the standard models of particle and quantum theories). Even Albert Einstein (1879–1955) in developing his theory of general relativity (i.e., the theory of gravitation) could only improve on Newton's theory by incorporating and explaining action at a distance (gravitational force acts with the speed of light as a fundamental tenet of the theory). Through the principle of equivalence of inertial and gravitational mass, Einstein explained gravitation as the effect that mass has on the fabric of space-time. But what actually mediates the gravitational effect (or, attraction) still intensely occupies modern physicists and cosmologists.

Gravitation since its early scientific formulation initially belonged to the domain of astronomers, at least as far as the observable universe was concerned. Theory successfully predicted the observed perturbations of planetary orbits and even the location of previously unknown new planets (Neptune's discovery in 1846 based on calculations motivated by observed perturbations in Uranus's orbit was a major triumph for Newton's law). However, it was also discovered that gravitational acceleration varies on Earth's surface, with respect to altitude and latitude. Newton's law of gravitation again provided the backdrop for the variations observed with pendulums.

An early achievement for his theory came when he successfully predicted the polar flattening in Earth's shape on the basis of hydrostatic equilibrium, which was confirmed finally (after some controversy) with geodetic measurements of long triangulated arcs in 1736–1737 by Pierre de Maupertuis and Alexis Clairaut. Gravitation thus also played a dominant role in geodesy, the science of determining the size and shape of the Earth, promulgated in large part by the father of modern geodesy, Friedrich R. Helmert (1843–1917) (Helmert, 1884).

Knowledge of terrestrial gravitation through the twentieth century was crucial for geodetic determination of a well-defined reference surface (a level surface) for heights and to correct astronomic observations of latitude and longitude for the irregular direction of gravitation (Heiskanen and Veing Meinesz, 1958; Molodensky et al., 1962; Heiskanen and Moritz, 1967; Torge, 1989; Sansò and Rummel, 1997). Also, geophysicists realized the potential of using gravity measurements to infer subsurface structures that could trap oil and gas deposits (e.g., Heiland, 1940; Nettleton, 1976; Telford et al., 1990). The Earth sciences increasingly rely on accurate gravimetry and gravitational models to aid in the interpretation of other sensor data. For example, oceanographers need an accurate level surface to infer surface currents from sea surface heights measured by satellite altimetry. Geophysicists analyzing crust deformation from GPS measurements of surface point coordinates must know whether the uplift or subsidence is accompanied by a change in subsurface mass – is it an elastic or a viscoelastic response? Ice sheet mapping (e.g., in Antarctica) using airborne laser altimetry must be accompanied with other sensors, such as gravity meters, that can infer the ice-bedrock interface. Hydrologists and glaciologists can now monitor the regional seasonal and secular changes in water/ice mass as seen by gravitation-mapping satellites as they continually orbit the Earth. In general, planetary probes glean significant geophysical information about their target from the gravitational effect on their orbits as tracked from Earth.

Other perhaps more utilitarian applications also rely on gravitational information. For example, longer-term precision navigation and guidance in the absence of external references, such as GPS, rely on inertial instruments. These autonomous navigation systems sense only inertial acceleration and require gravitational compensation, since a vehicle follows a path governed by both inertial and gravitational acceleration (Einstein's equivalence principle) (Jekeli, 2000).

Although not as prominent nor as prestigious in modern scientific endeavors, gravimetry at all scales, from point measurements to global surveys, continues to hold a critical, at least desirable place in the quest to answer a plethora of macroscopic inquiries into Earth's systems. For a modern view of the role that knowledge of the Earth's gravity field plays in the Earth sciences see Plag and Pearlman (2009) and references therein.

Newton's law of gravitation

In its original form Newton's law of gravitation applies only to idealized point masses. It describes the force of attraction, $\boldsymbol{F}$, experienced by two such solitary masses as being proportional to the product of the masses, m_1 and m_2; inversely proportional to the square of the distance, ℓ, between them; and directed along the line joining them, as expressed by the unit vector, $\boldsymbol{n}$:

$$\boldsymbol{F} = G\frac{m_1 m_2}{\ell^2}\boldsymbol{n}. \tag{1}$$

G is Newton's gravitational constant that takes care of the units between the left- and right-hand sides of the equation; it can be determined by experiment and the current accepted value is (Mohr et al., 2008):

$$G = (6.67428 \pm 0.00067) \times 10^{-11}\mathrm{m}^3/\left(\mathrm{kgs}^2\right). \tag{2}$$

The unit vector in Equation 1 is directed from either point mass to the other, and thus the gravitational force is attractive and applies equally to one mass as the other. Newton's law of gravitation is universal as far as we know, requiring reformulation only in Einstein's more comprehensive theory of general relativity, which describes gravitation as a characteristic curvature of the space-time continuum (Newton's formulation assumes instantaneous action and differs significantly from the general relativistic concept only when very large velocities or masses are involved, or, conversely, in applications of extremely high precision).

We can ascribe a gravitational acceleration to the gravitational force, which represents the acceleration that one mass undergoes due to the gravitational attraction of the other. Specifically, from the law of gravitation, we have (for point masses) the gravitational acceleration of m_1 due to the gravitational attraction of m_2:

$$\boldsymbol{g} = G\frac{m_2}{\ell^2}\boldsymbol{n}. \tag{3}$$

The vector $\boldsymbol{g}$ is independent of the mass, m_1, of the body being accelerated (which Galileo found by experiment).

By the law of superposition, the gravitational force, or the gravitational acceleration, due to many point masses is the vector sum of the forces or accelerations generated by the individual point masses. Manipulating vectors in this way is certainly feasible, but fortunately a more appropriate concept of gravitation as a scalar field simplifies the treatment of arbitrary mass distributions.

This more modern view of gravitation (adopted by eighteenth- and nineteenth-century mathematicians, such as Lagrange (1736–1813), Gauss (1777–1855), and Green (1793–1841)) holds that it is a *field* having a gravitational potential, V, defined in terms of the gravitational acceleration, $\boldsymbol{g}$, that a test particle would undergo in the field according to the equation

$$\nabla V = g, \tag{4}$$

where ∇ is the gradient operator (a vector). We use here the convention that the potential is always positive. Further elucidation of gravitation as a field grew from Einstein's attempt to incorporate gravitation into his theory of special relativity where no reference frame has particular significance above all others. It was necessary to consider that gravitational force is not a real force (i.e., it is not an applied force, like friction or propulsion) – rather it is known as a kinematic force, that is, one whose action is proportional to the mass on which it acts (like the centrifugal force); see Martin (1988). Under this precept, the geometry of space is defined intrinsically by the gravitational fields contained therein. We continue with the classical Newtonian potential, but interpret gravitation as an acceleration different from the acceleration induced by real, applied forces. This becomes especially important when considering the measurement of gravitation (Jekeli, 2007).

The gravitational potential, V, is a *scalar* function, and it is derived directly on the basis of Newton's law of gravitation. To make it completely consistent with this law and thus declare it a Newtonian potential, we must impose the following conditions of regularity at infinity:

$$\lim_{\ell\to\infty} \ell V = Gm \quad \text{and} \quad \lim_{\ell\to\infty} V = 0, \tag{5}$$

where m is the attracting mass. It easily follows from Equations 3–5 that the gravitational potential at any point in space due to a point mass is

$$V(\boldsymbol{x}) = \frac{Gm(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}, \tag{6}$$

where $\ell = |\boldsymbol{x}-\boldsymbol{x}'|$ is the distance from the mass point, $\boldsymbol{x}'$, to the evaluation point, $\boldsymbol{x}$, of the potential. Note that the field of a point mass has a singularity at $\boldsymbol{x} = \boldsymbol{x}'$.

Applying the law of superposition, the gravitational potential of many point masses is the sum of the potentials of the individual points. And, for infinitely many points in a closed, bounded region with infinitesimally small masses, dm, the summation changes to an integration:

$$V(\boldsymbol{x}) = G \int_{\text{mass}} \frac{dm(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|}. \tag{7}$$

Changing variables, $dm(\boldsymbol{x}') = \rho(\boldsymbol{x}')dv$, where ρ represents density (mass per volume) and dv is a volume element, we have:

$$V(\boldsymbol{x}) = G \iiint_v \frac{\rho(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|} dv. \tag{8}$$

In spherical polar coordinates the evaluation and integration points are (θ, λ, r) and (θ', λ', r'), respectively. The volume element in this case is given by $dv = r'^2 \sin\theta'\, d\lambda'\, d\theta'\, dr'$. It is readily shown that if the coordinate origin is placed at $\boldsymbol{x}$, then $dv = |\boldsymbol{x}'|^2 \sin\theta'\, d\lambda'\, d\theta'\, dr'$, and the apparent singularity disappears – the integral is said to be weakly singular. Indeed, the potential of a volume of mass density is continuous and differentiable everywhere.

Suppose the density distribution over the volume depends only on radial distance (from the center of mass, assumed to be at the origin): $\rho = \rho(r')$, and that the evaluation is at an exterior point. The surface bounding the masses necessarily is a sphere (say, of radius R) and with an appropriate choice of coordinates, one easily shows (e.g., Jekeli, 2007) that the integral (Equation 8) becomes:

$$V(\theta, \lambda, r) = \frac{GM}{r}, \quad r \geq R, \tag{9}$$

where M is the total mass bounded by the sphere. To a reasonably good approximation the Earth's density structure is spherically symmetric (see Chap. 4), and the external gravitational potential is that of a point mass (the approximation is valid to one part in two thousand).

Besides volumetric mass (density) distributions, it is of interest to consider surface distributions. Imagine an infinitesimally thin layer of mass on a surface, s, where the units of density in this case are those of mass per area. Then, analogous to Equation 8, the potential is:

$$V(\boldsymbol{x}) = G \iint_s \frac{\kappa(\boldsymbol{x}')}{|\boldsymbol{x}-\boldsymbol{x}'|} ds, \tag{10}$$

where κ is the surface mass density. In this case, V is a continuous function everywhere, but its first derivatives are discontinuous at the surface.

Now, imagine two infinitesimally close density layers (double layer, or layer of mass dipoles), where the units of density are those of mass per area times length. The potential in this case is given by Heiskanen and Moritz (1967, p. 8):

$$V(\boldsymbol{x}) = G \iint_s \kappa(\boldsymbol{x}') \frac{\partial}{\partial n}\left(\frac{1}{|\boldsymbol{x}-\boldsymbol{x}'|}\right) ds, \tag{11}$$

where $\partial/\partial n$ is the directional derivative along the perpendicular to the surface. Now, V itself is discontinuous at the surface, as are all its derivatives. In all cases, V is a Newtonian potential, having been derived from the basic formula (Equation 6) for a point mass that follows from Newton's law of gravitation, Equation 1.

The following properties of the gravitational potential are useful and foundational for many applications. First, consider Stokes's theorem for a vector function, $\boldsymbol{f}$, defined on a surface, s:

$$\iint_s (\nabla \times \boldsymbol{f}) \cdot \boldsymbol{n}\, ds = \oint_p \boldsymbol{f} \cdot d\boldsymbol{r}, \tag{12}$$

where p is any closed path in the surface, $\boldsymbol{n}$ is the unit vector perpendicular to the surface, and $d\boldsymbol{r}$ is a differential displacement along the path. From Equation 4, we find

$$\nabla \times \boldsymbol{g} = \boldsymbol{0}, \tag{13}$$

since $\nabla \times \nabla = \boldsymbol{0}$; and hence, applying Stokes's theorem, we find with $f = m\boldsymbol{g}$ that

$$w = \oint_p m\boldsymbol{g} \cdot d\boldsymbol{r} = 0. \tag{14}$$

That is, the gravitational field is conservative: the work, w, expended in moving a mass around a closed path in this field vanishes. In contrast, dissipating forces (real forces!), like friction, expend work or energy, which shows again the special nature of the gravitational force.

It can be shown (Kellogg, 1953, p. 156) that the second partial derivatives of a Newtonian potential, V, satisfy the following differential equation, known as Poisson's equation:

$$\nabla^2 V = -4\pi G\rho, \tag{15}$$

where $\nabla^2 = \nabla \cdot \nabla$ formally is the scalar product of two gradient operators and is called the Laplacian operator. In Cartesian coordinates, it is given by:

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}. \tag{16}$$

Equation 15 is a local characterization of the potential field, as opposed to the global characterization given by Equation 8. Poisson's equation holds wherever the mass density, ρ, satisfies certain conditions similar to continuity (Hölder conditions, see Kellogg, 1953, pp. 152–3). A special case of Equation 15 applies for those points where the density vanishes (i.e., in free space); then Poisson's equation turns into Laplace's equation:

$$\nabla^2 V = 0. \tag{17}$$

It is easily verified that the point mass potential satisfies Equation 17; that is,

$$\nabla^2 \left(\frac{1}{|\boldsymbol{x} - \boldsymbol{x}'|} \right) = 0, \quad \boldsymbol{x} \neq \boldsymbol{x}'. \tag{18}$$

The solutions to Laplace's Equation 17 are known as harmonic functions (here, we also impose the conditions (Equation 5) on the solution, if it is a Newtonian potential and if the mass-free region includes infinity). Hence, every Newtonian potential is a harmonic function in free space. The converse is also true: every harmonic function can be represented as a Newtonian potential of a mass distribution.

Several structures of the Earth's crust may be approximated by forms characterized essentially by a two-dimensional cross section, such as sheets, slabs, dikes, and faults that extend indefinitely in one dimension. The gravitational potential associated with such density contrasts is known as a logarithmic potential, since the potential (up to an arbitrary constant) of an infinite, straight line (that is, a differential element of the cross section) is given by (Telford et al., 1990; Kellogg, 1953)

$$\delta V(\boldsymbol{x}) = 2G\mu \ln \frac{1}{r}, \tag{19}$$

where r is the distance of the evaluation point, $\boldsymbol{x}$, from the line and μ is the constant linear mass density of the line. If the line is indefinitely extended in the y-direction, then the potential of any form with cross A in the x-z plane is

$$V(\boldsymbol{x}) = 2G\mu \iint_A \ln \frac{1}{\sqrt{(x - x')^2 + (z - z')^2}} dx' dz'. \tag{20}$$

Strictly speaking, the logarithmic potential is not a Newtonian potential since it does not vanish when $r \to \infty$ (on account of the arbitrary constant); however, it is a useful approximation in many geophysical applications.

Spatial and temporal variation above the surface

This section develops two types of solutions to standard boundary-value problems when the boundary is a sphere: the spherical harmonic series and an integral with a Green's function. Attention is given to the Dirichlet boundary-value problem for the exterior region (that includes infinity). That is, the gravitational potential, V, satisfies Laplace's Equation 17 everywhere outside the sphere, and the solution to this partial differential equation is sought based on values of V given on the sphere. Other types of boundary-value problems presume given linear combinations of zero- and first-order derivatives of the potential on the boundary. Also, other boundaries that better approximate the Earth yield more difficult solutions – the ellipsoid is relatively easy (Hotine, 1969; Jekeli, 1988); whereas, the Earth's surface is challenging (Molodensky et al., 1962; Moritz, 1980). With the spherical boundary, one obtains a particularly straightforward spectral representation that allows analysis of the spatial variation of the field at different wavelengths.

Spherical harmonics

For simple boundaries, Laplace's Equation 17 is relatively easy to solve provided there is an appropriate coordinate system. For the Earth, the solutions commonly rely on approximating the boundary by a sphere of radius, R. In spherical polar coordinates, (θ, λ, r), the Laplacian operator is given by

$$\nabla^2 = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial}{\partial r}\right) + \frac{1}{r^2}\frac{1}{\sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial}{\partial \theta}\right) + \frac{1}{r^2 \sin^2\theta}\frac{\partial^2}{\partial \lambda^2}, \tag{21}$$

and solutions to $\nabla^2 V = 0$ in the space outside a sphere of radius R, with the center at the coordinate origin, are as

follows (Kellogg, 1953, p. 143; Morse and Feshbach, 1953, p. 1264; Hobson, 1965, p. 10):

$$V(\theta,\lambda,r) = P_{nm}(\cos\theta)\sin m\lambda\frac{1}{r^{n+1}} \text{ or}$$
$$V(\theta,\lambda,r) = P_{nm}(\cos\theta)\cos m\lambda\frac{1}{r^{n+1}}, \tag{22}$$

where $P_{nm}(t)$ is the associated Legendre function of the first kind and n, m are integers such that $0 \le m \le n$, $n \ge 0$. Other solutions exist, but only Equation 22 is consistent with the problem of finding a real-valued Newtonian potential for the exterior space of the Earth (regular at infinity and 2π-periodic in longitude).

The general solution is a linear combination of solutions (Equation 22) for all possible integers, n and m,

$$V(\theta,\lambda,r) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n}\left(\frac{R}{r}\right)^{n+1} v_{nm}\bar{Y}_{nm}(\theta,\lambda), \tag{23}$$

where $\bar{Y}_{nm}$ are surface spherical harmonic functions defined as

$$\bar{Y}_{nm}(\theta,\lambda) = \bar{P}_{n|m|}(\cos\theta)\begin{cases}\cos m\lambda, & m \ge 0\\ \sin|m|\lambda, & m < 0\end{cases} \tag{24}$$

and $\bar{P}_{nm}$ is a normalization of P_{nm} so that the orthogonality of the spherical harmonics is

$$\frac{1}{4\pi}\iint_{\sigma}\bar{Y}_{nm}(\theta,\lambda)\bar{Y}_{n'm'}(\theta,\lambda)d\sigma = \begin{cases}1, & n = n' \text{ and } m = m'\\ 0, & n \ne n' \text{ or } m \ne m'\end{cases} \tag{25}$$

and where $\sigma = \{(\theta,\lambda)|0 \le \theta \le \pi, 0 \le \lambda \le 2\pi\}$ represents the unit sphere, with $d\sigma = \sin\theta d\theta d\lambda$. For a complete mathematical treatment of spherical harmonics, see Müller (1966). The functions $\bar{Y}_{nm}(\theta,\lambda)/r^{n+1}$, are also known as solid spherical harmonics.

The constant coefficients, v_{nm}, known as Stokes's constants, have identical units of measure. Applying the orthogonality relation, Equation 25, to the general solution (Equation 23), these coefficients can be determined if the function, V, is known on the bounding sphere (Dirichlet boundary condition):

$$v_{nm} = \frac{1}{4\pi}\iint_{\sigma}V(\theta,\lambda,R)\bar{Y}_{nm}(\theta,\lambda)d\sigma. \tag{26}$$

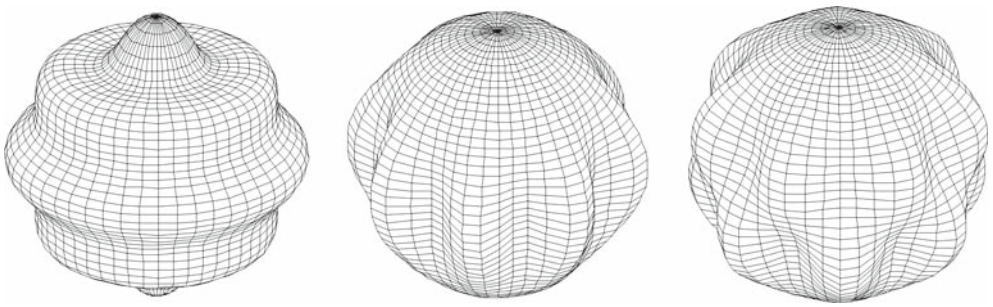
The spherical harmonic expansion of V, Equation 23, with Equation 26 for the coefficients is a solution to the Dirichlet boundary-value problem if the boundary is a sphere. The solution thus exists and is unique in the sense that these boundary values generate no other potential.

In a more formal mathematical setting, the solution (Equation 23) is an infinite linear combination of orthogonal basis functions (eigenfunctions) and the coefficients, v_{nm}, are the corresponding eigenvalues. One may also interpret the set of coefficients as the spectrum (Legendre spectrum) of the potential on the sphere of radius, R (analogous to the Fourier spectrum of a function on the plane or line). The integers, n, m, correspond to wave numbers, and are called degree (n) and order (m), respectively.

The spherical harmonics are further classified as zonal ($m = 0$), meaning that the zeros of $\bar{Y}_{n0}$ divide the sphere into latitudinal zones; sectorial ($m = n$), where the zeros of $\bar{Y}_{nn}$ divide the sphere into longitudinal sectors; and tesseral, where the zeros of $\bar{Y}_{nm}$ tessellate the sphere (Figure 1).

Recent global mappings of the Earth's gravitational field using satellite observations, specifically the GRACE mission (Gravity Recovery and Climate Experiment, Tapley et al., 2004) have emphasized the temporal dependence of the potential due to mass transport on and within the Earth (hydrologic flux, glacial isostatic adjustment, ice mass loss). In terms of the global spherical harmonic expansion this implies a temporal dependence in the Stokes's coefficients:

$$V(\theta,\lambda,r;t) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n}\left(\frac{R}{r}\right)^{n+1} v_{nm}(t)\bar{Y}_{nm}(\theta,\lambda). \tag{27}$$



Gravity Field of the Earth, Figure 1 Spherical harmonics of the types, zonal (*left*), sectorial (*middle*), and tesseral (*right*).

While the spherical harmonic series has its advantages in global representations and spectral interpretations of the field, the support region for each of its basis functions is global, which can be inefficient from the standpoint of data analysis. As more data are incorporated, or improved, each coefficient, v_{nm}, must be reestimated. Representations using local-support basis functions have recently been developed (e.g., Freeden et al., 1998) and are topics of current research. Such constructions also permit multi-resolution decompositions analogous to wavelet analysis.

Green's functions

The Green's function representation of the field also offers a more regional characterization of the field. Although the Green's function strictly has global support, it attenuates rapidly from the evaluation point of the potential and reflects that a remote change in boundary value generally does not appreciably affect the local determination of the field.

When the boundary is a sphere, the solution to the boundary-value problem using a Green's function is easily derived for the potential from its spherical harmonic series representation. Additional integral relationships (with appropriate Green's functions) among all the derivatives of the potential can be derived similarly (Jekeli, 2007). Using the decomposition formula for the nth-degree Legendre polynomial,

$$P_n(\cos\psi) = \frac{1}{2n+1}\sum_{m=-n}^{n}\bar{Y}_{nm}(\theta,\lambda)\bar{Y}_{nm}(\theta',\lambda'), \tag{28}$$

where

$$\cos\psi = \cos\theta\cos\theta' + \sin\theta\sin\theta'\cos(\lambda-\lambda'), \tag{29}$$

and, substituting the harmonic coefficients, Equation 26, into the series, Equation 23, we find

$$V(\theta,\lambda,r) = \frac{1}{4\pi}\iint_{\sigma} V(\theta',\lambda',R)U(\psi,r)d\sigma, \tag{30}$$

where with $\ell = \sqrt{r^2+R^2-2rR\cos\psi}$

$$\begin{aligned} U(\psi,r) &= \sum_{n=0}^{\infty}(2n+1)\left(\frac{R}{r}\right)^{n+1}P_n(\cos\psi) \\ &= \frac{R(r^2-R^2)}{\ell^3}. \end{aligned} \tag{31}$$

The closed formula (right side of Equation 31) comes from an appropriate manipulation of the Coulomb expansion (Cushing, 1975, p. 155) for the reciprocal distance,

$$\frac{1}{\ell} = \frac{1}{R}\sum_{n=0}^{\infty}\left(\frac{R}{r}\right)^{n+1}P_n(\cos\psi). \tag{32}$$

Solutions (Equations 23 and 30) to the Dirichlet boundary-value problem for a spherical boundary are identical. The integral (Equation 30) is known as the Poisson integral and the function U is also known as Poisson's kernel.

The disturbing potential

For convenience, one separates Earth's gravitational potential into a reference potential ("Normal Reference Field") and the disturbing potential, T. The disturbing potential is harmonic in free space, satisfies Poisson's integral if the boundary is a sphere, and is expandable in a series of solid spherical harmonics:

$$T(\theta,\lambda,r) = \frac{GM}{R}\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\left(\frac{R}{r}\right)^{n+1}\delta C_{nm}\bar{Y}_{nm}(\theta,\lambda) \tag{33}$$

where M is the total mass (including the atmosphere) of the Earth and δC_{nm} are unitless harmonic coefficients, being also the difference between coefficients for the total and reference gravitational potentials. The coefficient, δC_{00}, is zero under the assumption that the reference field accounts completely for the central part of the total field. Note that these coefficients also specifically refer to the sphere of radius, R.

The gravity anomaly (in spherical approximation) is defined by

$$\Delta g(\theta,\lambda,r) = \left(-\frac{\partial}{\partial r}-\frac{2}{r}\right)T(\theta,\lambda,r), \tag{34}$$

or, using Equation 33,

$$\Delta g(\theta,\lambda,r) = \frac{GM}{R^2}\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\left(\frac{R}{r}\right)^{n+2}(n-1)\delta C_{nm}\bar{Y}_{nm}(\theta,\lambda). \tag{35}$$

The spectral relationship between gravity anomalies and the disturbing potential, obtained by comparing Equations 33 and 35, leads to a corresponding Green's function that transforms one to the other. This function in geodesy is called more commonly the Stokes function:

$$\begin{aligned} S(\psi,r) &= \sum_{n=2}^{\infty}\frac{2n+1}{n-1}\left(\frac{R}{r}\right)^{n+1}P_n(\cos\psi) \\ &= 2\frac{R}{\ell}+\frac{R}{r}-3\frac{R\ell}{r^2}-5\frac{R^2}{r^2}\cos\psi \\ &\quad -3\frac{R^2}{r^2}\cos\psi\ln\frac{\ell+r-R\cos\psi}{2r} \end{aligned} \tag{36}$$

and thus (Hofmann-Wellenhof and Moritz, 2005; Jekeli, 2007)

$$T(\theta,\lambda,r) = \frac{R}{4\pi}\iint_{\sigma}\Delta g(\theta',\lambda',R)S(\psi,r)d\sigma, \tag{37}$$

which is known as Stokes's integral. This famous equation in geodesy solves the Robin boundary-value problem for the sphere: the potential and its normal derivative are given on the boundary; see Equation 34.

The solution is not unique since, as Equation 35 shows, the gravity anomaly has no first-degree harmonics, which, therefore, cannot be determined from the boundary values. Conventionally, as shown in Equation 36, the Stokes kernel also excludes the zero-degree harmonic and thus the complete solution for the disturbing potential is given by

$$T(\theta,\lambda,r) = \frac{GM}{r}\delta C_{00} + \frac{GM}{R}\sum_{m=-1}^{1}\left(\frac{R}{r}\right)^2 \delta C_{1m}\bar{Y}_{1m}(\theta,\lambda) + \frac{R}{4\pi}\iint_{\sigma} \Delta g(\theta',\lambda',R)S(\psi,r)d\sigma. \quad (38)$$

The central term, δC_{00}, is proportional to the difference in GM of the Earth and reference ellipsoid and is zero to high accuracy. The first-degree harmonic coefficients, δC_{1m}, are proportional to the center-of-mass coordinates and can also be set to zero with appropriate definition of the coordinate system (see Low-Degree Harmonics as Density Moments). Thus, the Stokes integral (Equation 37) is the more common expression for the disturbing potential, but it embodies hidden constraints.

Spectral contributions

The correlation function of the disturbing potential on a sphere is defined as (Moritz, 1980; Jekeli, 2010)

$$\phi_T(\psi) = \frac{1}{8\pi^2}\int_0^{2\pi}\iint_{\sigma} T(\theta,\lambda,R)T(\theta',\lambda',R)\sin\theta d\theta d\lambda d\alpha, \quad (39)$$

where the points (θ,λ,R) and (θ',λ',R) are connected by Equation 29 and their relative azimuth, α,

$$\tan\alpha = \frac{-\sin\theta'\sin(\lambda-\lambda')}{\sin\theta\cos\theta' - \cos\theta\sin\theta'\cos(\lambda-\lambda')}, \quad (40)$$

and where the integration is performed over all possible pairs of these points, with fixed spherical distance, ψ, $0 \le \psi \le \pi$.

Because of its sole dependence on ψ, ϕ_T can be expressed as an infinite series of Legendre polynomials:

$$\phi_T(\psi) = \sum_{n=0}^{\infty}(2n+1)(\Phi_T)_n P_n(\cos\psi) \quad (41)$$

where the coefficients, $(\Phi_T)_n$ constitute the Legendre transform of ϕ_T:

$$(\Phi_T)_n = \frac{1}{2}\int_0^{\pi}\phi_T(\psi)P_n(\cos\psi)\sin\psi d\psi. \quad (42)$$

Substituting Equations 28 and 39 into Equation 42, then simplifying using the orthogonality of spherical harmonics, (Equation 25), and the definition of the Legendre spectrum, Equation 26, we find:

$$(\Phi_T)_n = \frac{1}{2n+1}\sum_{m=-n}^{n} v_{nm}^2. \quad (43)$$

The quantities, $(\Phi_T)_n$, constitute the power spectral density (psd) of T; and,

$$(c_T)_n = (2n+1)(\Phi_T)_n \quad (44)$$

is known as the degree variance or variance per degree, since the total variance, from Equation 41, is

$$\phi_T(0) = \sum_{n=0}^{\infty}(c_T)_n. \quad (45)$$

For the first and second radial derivatives of the disturbing potential, called the gravity disturbance ($\delta g = -\partial T/\partial r$) and disturbing vertical-vertical gradient ($\delta\Gamma = \partial^2 T/\partial r^2$), one has analogous psd's that are related to the psd of T as follows:

$$(\Phi_{\delta\Gamma})_n = \left(\frac{n+2}{R}\right)^2(\Phi_{\delta g})_n = \left(\frac{n+2}{R}\right)^2\left(\frac{n+1}{R}\right)^2(\Phi_T)_n. \quad (46)$$

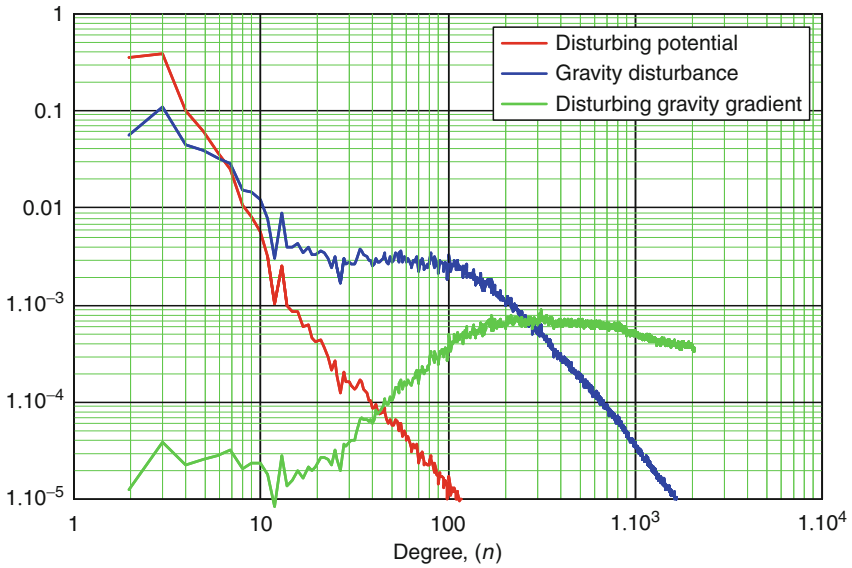
Using the most recent high-degree Earth Gravitational Model, EGM08 (Pavlis et al., 2008), Figure 2 illustrates the relative spectral content of the derivatives of the (disturbing) gravitational potential. Clearly, the higher derivatives emphasize the shorter wavelengths (or higher degrees, higher frequencies). The EGM08 model has maximum degree 2,160, corresponding to about 9.3 km resolution on the Earth's surface ($180^\circ/2{,}160 = 5$ arcmin).

Low-degree harmonics

The low-degree spherical harmonics of the Earth's gravitational potential lend themselves to interpretation with respect to the most elemental distribution of the Earth's mass density, which also leads to fundamental geometric characterizations, particularly for the second-degree harmonics. Let

$$C_{nm}^{(a)} = \frac{a}{GM}\left(\frac{R}{a}\right)^{n+1} v_{nm} \quad (47)$$

be unitless coefficients that refer to a sphere of radius, a, which is the equatorial radius of the Earth ellipsoid. (Recall that coefficients, v_{nm}, Equation 50, refer to a sphere



Gravity Field of the Earth, Figure 2 Degree variances of radial derivatives of the Earth's disturbing potential. All values are normalized by the respective total variance.

of radius, R.) Relative to the central harmonic coefficient, $C_{00}^{(a)} = 1$, the next significant harmonic, $C_{20}^{(a)}$, is more than three orders of magnitude smaller; and, the remaining harmonic coefficients are at least two to three orders of magnitude smaller than that. The attenuation of the coefficients after degree two is much more gradual (Table 1), indicating that the bulk of the potential can be described by an ellipsoidal field. The normal gravitational field is such a field, but it also adheres to a geodetic definition that requires the underlying ellipsoid to be an equipotential surface in the corresponding normal gravity field. The following sections examine the low-degree harmonics from the two perspectives of interpretation and reference.

Low-degree harmonics as density moments

Returning to the general expression for the gravitational potential in terms of the Newtonian density integral (Equation 8), and substituting the spherical harmonic series for the reciprocal distance (Equation 32 with 29):

$$V(\theta,\lambda,r) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n}\left(\frac{R}{r}\right)^{n+1}\left(\frac{G}{R^{n+1}(2n+1)}\iiint_{\text{volume}}(r')^n \bar{Y}_{nm}(\theta',\lambda')\rho(\theta',\lambda',r')dv\right)\bar{Y}_{nm}(\theta,\lambda) \tag{48}$$

yields a multipole expansion (so-called from electrostatics) of the potential. The spherical harmonic (Stokes) coefficients are multipoles of the density distribution:

$$v_{nm} = \frac{G}{R^{n+1}(2n+1)}\iiint_{v}(r')^n \bar{Y}_{nm}(\theta',\lambda')\rho(\theta',\lambda',r')dv. \tag{49}$$

One may also consider the nth-order moments of the mass density (from the statistics of distributions) defined by

$$\mu_{\alpha\beta\gamma}^{(n)} = \iiint_{v}(x')^{\alpha}(y')^{\beta}(z')^{\gamma}\rho dv, \qquad n = \alpha+\beta+\gamma. \tag{50}$$

The multipoles of degree n and the moments of order n are related, though not all $(n+1)(n+2)/2$ moments of order n can be determined from the $2n+1$ multipoles of degree n, when $n \geq 2$ (clearly risking confusion, we defer to the common nomenclature of "order" for moments and "degree" for spherical harmonics). This indeterminacy is directly connected to the inability to determine the density distribution uniquely from external measurements of the potential (Chao, 2005), which is the classic geophysical inverse problem.

The zero-degree Stokes coefficient is coordinate invariant and is proportional to the total mass of the Earth:

$$v_{00} = \frac{G}{R}\iiint_{v}\rho dv = \frac{GM}{R}. \tag{51}$$

It also represents a mass monopole, and it is proportional to the zeroth moment of the density.

Gravity Field of the Earth, Table 1 Spherical harmonic coefficients of the total gravitational potential[a]

Degree, (n)	Order, (m)	$C_{nm}^{(a)}$	$C_{n,-m}^{(a)}$
2	0	−4.84170E-04	0.0
2	1	−2.39832E-10	1.42489E-09
2	2	2.43932E-06	−1.40028E-06
3	0	9.57189E-07	0.0
3	1	2.03048E-06	2.48172E-07
3	2	9.04802E-07	−6.19006E-07
3	3	7.21294E-07	1.41437E-06
4	0	5.39992E-07	0.0
4	1	−5.36167E-07	−4.73573E-07
4	2	3.50512E-07	6.62445E-07
4	3	9.90868E-07	−2.00976E-07
4	4	−1.88472E-07	3.08827E-07

[a]GRACE model GGM02S (Tapley et al., 2005)

The first-degree harmonic coefficients (representing dipoles) are proportional to the coordinates of the center of mass, (x_{cm}, y_{cm}, z_{cm}), which are proportional to the first-order moments of the density, as verified by recalling the definition of the first-degree spherical harmonics:

$$v_{1m} = \frac{G}{\sqrt{3}R^2} \iiint_v \rho \begin{Bmatrix} r' \sin\theta' \sin\lambda', & m=-1 \\ r' \cos\theta', & m=0 \\ r' \sin\theta' \cos\lambda', & m=1 \end{Bmatrix} dv \tag{52}$$
$$= \frac{GM}{\sqrt{3}R^2} \begin{Bmatrix} y_{cm}, & m=-1 \\ z_{cm}, & m=0 \\ x_m, & m=1 \end{Bmatrix}$$

Nowadays, by tracking satellites, we have access to the center of mass of the Earth (including its atmosphere) since it defines the center of their orbits. Ignoring the small motion of the center of mass (annual amplitude of several mm) due to the temporal variations in the mass distribution, we may choose the coordinate origin for the gravitational potential to coincide with the center of mass, thus annihilating the first-degree coefficients.

The second-order density moments likewise are related to the second-degree harmonic coefficients (quadrupoles). They also define the inertia tensor of the body. The inertia tensor is the proportionality factor in the equation that relates the angular momentum vector, $\boldsymbol{H}$, and the angular velocity, $\boldsymbol{\omega}$, of a body, like the Earth:

$$\boldsymbol{H} = I\boldsymbol{\omega}, \tag{53}$$

and is given by

$$I = \begin{pmatrix} I_{xx} & I_{xy} & I_{xz} \\ I_{yx} & I_{yy} & I_{yz} \\ I_{zx} & I_{zy} & I_{zz} \end{pmatrix}. \tag{54}$$

It comprises the moments of inertia on the diagonal

$$I_{xx} = \iiint_{\text{volume}} \left(y'^2 + z'^2\right)\rho dv,$$
$$I_{yy} = \iiint_{\text{volume}} \left(z'^2 + x'^2\right)\rho dv, \tag{55}$$
$$I_{zz} = \iiint_{\text{volume}} \left(x'^2 + y'^2\right)\rho dv,$$

and the products of inertia off the diagonal

$$I_{xy} = I_{yx} = -\iiint_{\text{volume}} x'y'\rho dv,$$
$$I_{xz} = I_{zx} = -\iiint_{\text{volume}} x'z'\rho dv, \tag{56}$$
$$I_{yz} = I_{zy} = -\iiint_{\text{volume}} y'z'\rho dv.$$

Note that

$$I_{xx} = \mu_{020}^{(2)} + \mu_{002}^{(2)}, \quad I_{xy} = -\mu_{110}^{(2)}, \quad I_{xz} = -\mu_{101}^{(2)}$$
$$I_{yy} = \mu_{002}^{(2)} + \mu_{200}^{(2)}, \quad I_{yz} = -\mu_{011}^{(2)} \tag{57}$$
$$I_{zz} = \mu_{200}^{(2)} + \mu_{020}^{(2)}$$

and there as many (six) independent tensor components as second-order density moments. Using the explicit expressions for the second-degree spherical harmonics, we have from Equation 49 with $n = 2$:

$$v_{2,-2} = -\frac{\sqrt{15}\,G}{5R^3} I_{xy}, \quad v_{2,-1} = -\frac{\sqrt{15}\,G}{5R^3} I_{yz},$$
$$v_{2,1} = -\frac{\sqrt{15}\,G}{5R^3} I_{xz}, \quad v_{2,0} = \frac{\sqrt{5}\,G}{10R^3}\left(I_{xx} + I_{yy} - 2I_{zz}\right), \tag{58}$$
$$v_{2,2} = \frac{\sqrt{15}\,G}{10R^3}\left(I_{yy} - I_{xx}\right).$$

These are also known as MacCullagh's formulas. Not all density moments (or, moments of inertia) can be determined from the Stokes coefficients.

If the coordinate axes are chosen so as to diagonalize the inertia tensor (products of inertia are equal to zero), then they are known as principal axes of inertia, or also "figure" axes. For the Earth the z-figure axis is very close to the spin axis (within several meters at the pole); both axes move with respect to each other and with respect to the Earth's crust, with combinations of various periods (daily, monthly, annual, etc.) as well as secularly in a wandering fashion. Because of these polar motions the figure axis is not a useful coordinate axis that defines a frame fixed to the (surface of the) Earth. However,

because of the proximity of the figure axis to the defined reference z-axis, the second-degree, first-order harmonic coefficients of the geopotential, and hence the products of inertia, I_{xz} and I_{yz}, are relatively small (Table 1).

The arbitrary choice of the x-axis of our Earth-fixed reference coordinate system certainly did not attempt to eliminate the product of inertia, I_{xy} (the x-axis is defined by the intersection of the Greenwich meridian with the equator; and the y-axis completes a right-handed mutually orthogonal triad). However, it is possible to determine where the x-figure axis is located by combining values of the second-degree, second-order harmonic coefficients. Let u, v, and w be the axes that define a coordinate system in which the inertia tensor is diagonal; and, assume that the approximation, $I_{ww} = I_{zz}$, holds. A rotation by the angle, $-\lambda_0$, about the w- (also z-) figure axis brings this ideal coordinate system back to the conventional one in which we calculate the harmonic coefficients. Tensors transform under rotation, defined by matrix, $\mathcal{R}$, according to

$$I_{xyz} = \mathcal{R} I_{uvw} \mathcal{R}^{\mathrm{T}}. \tag{59}$$

With the rotation about the w-axis given by the matrix:

$$\mathcal{R} = \begin{pmatrix} \cos\lambda_0 & -\sin\lambda_0 & 0 \\ \sin\lambda_0 & \cos\lambda_0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{60}$$

and with Equation 16 it is straightforward to show that

$$\begin{aligned} v_{2,-2} &= -\frac{\sqrt{15}G}{10R^3}(I_{uu} - I_{vv})\sin 2\lambda_0 \\ v_{2,2} &= -\frac{\sqrt{15}G}{10R^3}(I_{uu} - I_{vv})\cos 2\lambda_0 \end{aligned} \tag{61}$$

Hence, we have

$$\lambda_0 = \frac{1}{2}\tan^{-1}\frac{v_{2,-2}}{v_{2,2}}, \tag{62}$$

where the quadrant is determined by the signs of the harmonic coefficients. From Table 1 and Equation 47, we find that $\lambda_0 = -14.929°$; that is, the u-figure axis is in the mid-Atlantic between South America and Africa.

The second-degree, second-order harmonic coefficient, $v_{2,2}$, indicates the asymmetry of the Earth's mass distribution with respect to the equator. Since $v_{2,2} > 0$ (for the Earth), Equation 61 shows that $I_{vv} > I_{uu}$ and thus the equator "bulges" more in the direction of the u-figure axis; conversely, the equator is flattened in the direction of the v-figure axis. This flattening is relatively small:$1.1 \cdot 10^{-5}$.

Finally, consider the most important second-degree harmonic coefficient, the second zonal harmonic, $v_{2,0}$. Irrespective of the x-axis definition, it is proportional to the difference between the moment of inertia, I_{zz}, and the average of the equatorial moments, $(I_{xx} + I_{yy})/2$. Again, since $v_{2,0} < 0$, the Earth bulges more around the equator and is flattened at the poles. The second zonal harmonic coefficient is roughly 1,000 times larger than the other second-degree coefficients and thus indicates a substantial polar flattening (owing to the Earth's early more fluid state). This flattening is approximately 0.003.

Normal reference field

A model for the Earth's gravitational field that is mathematically simple and yet accurate to better than one part in sixty-three thousand (for the potential) is based on an ellipsoid of revolution rather than the sphere. The precise geodetic definition of the normal (reference) gravitational potential is that it is generated by an ellipsoidal mass that rotates with the Earth such that the gravitational potential plus the potential due to the centrifugal acceleration (as related according to Equation 4) is a constant, U_0, on the ellipsoid. This definition sets up a Dirichlet boundary-value problem, whose solution is an analytic expression in ellipsoidal coordinates (Hofmann-Wellenhof and Moritz, 2005). The boundary is defined by two parameters, the semi-major axis, a, and polar flattening, f, of the (rotationally symmetric) normal ellipsoid; and the boundary values of the potential are defined by the given constant, U_0, and the rotation rate of the Earth, ω_e (which specifies the centrifugal potential at any particular latitude). The normal ellipsoid, by definition, is centered at Earth's center of mass.

The solution for the normal potential is thus based on four fundamental constants, Earth's rotation rate, the size and shape parameters of the ellipsoid, and a gravitational scale parameter. The constant, U_0, is related to several other gravitational quantities, such as the value of normal gravity (mass attraction plus the centrifugal acceleration) at the equator, or the total mass of the ellipsoid, or it could simply be taken as the potential value of the geoid. It can be shown (Hofmann-Wellenhof and Moritz, 2005) that

$$U_0 = \frac{GM}{E}\tan^{-1}\frac{E}{b} + \frac{1}{3}\omega_e^2 a^2, \tag{63}$$

where $b = a(1-f)$ is the semi-minor axis, $E = \sqrt{a^2 - b^2}$ is the linear eccentricity of the ellipsoid, and M is its total mass. The mass density is not constant within the normal ellipsoid and its variation does not approximate Earth's internal structure. Rather the normal gravity potential (the sum of gravitational potential due to mass attraction and the centrifugal potential) is designed to emulate the actual external field, being constant on the ellipsoid, as the actual gravity potential is constant on the geoid.

The normal field is always defined by adopting values of the four parameters that approximately match the corresponding actual parameters for the Earth. Before artificial Earth-orbiting satellites, the gravitational scale could best be accessed by measuring gravity (e.g., a nominal value at the equator served as the physical scale parameter). By tracking satellites, the value of GM for the Earth (total mass including the atmosphere) can be estimated to high accuracy, and is still used today to define the physical scale of the normal potential. The ellipsoid parameters

were traditionally determined from (partially) global geometric surveying networks projected onto a best-fitting ellipsoid. Today, the semi-major axis is well determined from a best fit of an ellipsoid to a global set of ocean heights derived from satellite altimeter data. The polar flattening, according to "Low-Degree Harmonics as Density Moments," is related to the second zonal harmonic of Earth's gravitational potential, which likewise can be estimated very accurately from satellite tracking data. The second zonal harmonic coefficient, the *dynamical form factor*, thus serves as the shape parameter for the normal ellipsoid. Table 2 lists internationally adopted normal ellipsoids over the past 100 years.

The exact analytical expression for the normal gravity potential then yields exact expressions for the normal gravity (and gravitation) and all its gradients. Nevertheless, approximate series expressions are sometimes preferred, especially to conform to a particular coordinate system. For example, the normal gravitational potential (due only to mass attraction) in spherical polar coordinates is given by

$$V(\theta,\lambda,r) = \frac{GM}{r}\left(1 - \sum_{n=2}^{\infty} J_{2n}\left(\frac{a}{r}\right)^{2n} P_{2n}(\cos\theta)\right), \tag{64}$$

where with $e = E/a$ (the first eccentricity),

$$J_{2n} = (-1)^{n+1}\frac{3e^{2n}}{(2n+1)(2n+3)}\left(1 - n + 5n\frac{J_2}{e^2}\right), \tag{65}$$

and where

$$J_2 = \frac{1}{Ma^2}(I_{zz} - I_{xx}) = -\frac{\sqrt{5}R^3}{GMa^2}v_{2,0} = -\sqrt{5}C_{2,0}^{(a)}, \tag{66}$$

making use of Equations 58 and 47. From Equation 65 it is clear that only the first few coefficients (say, up to $n = 4$) are significant for all applications, since $e^2 \approx 0.0066$ (Table 3). Also, the normal potential, comprising only even zonal harmonics, is independent of longitude and symmetric with respect to the equator.

Normal gravity (including the centrifugal acceleration) on the ellipsoid as function of *geodetic* latitude, ϕ (angle of the perpendicular to the ellipsoid with respect to the equator), is given by the Somigliana formula (Moritz, 1980):

$$\gamma(\phi) = \frac{a\gamma_a\cos^2\phi + b\gamma_b\sin^2\phi}{\sqrt{a^2\cos^2\phi + b^2\sin^2\phi}}, \tag{67}$$

where normal gravity on the equator and at the poles is

$$\gamma_a = \frac{GM}{ab}\left(1 - m - \frac{m}{6}\frac{e'q'_0}{q_0}\right), \tag{68}$$

$$\gamma_b = \frac{GM}{a^2}\left(1 + \frac{m}{3}\frac{e'q'_0}{q_0}\right), \tag{69}$$

with $m = \omega_e^2 a^2 b/GM$, $e' = E/b$ (the second eccentricity), and

$$q_0 = \frac{1}{2}\left(1 + \frac{3}{e'^2}\right)\tan^{-1}e' - \frac{3}{2e'^2},$$
$$q'_0 = 3\left(1 + \frac{1}{e'^2}\right)\left(1 - \frac{1}{e'^2}\tan^{-1}e'\right) - 1. \tag{70}$$

At points along the perpendicular to the ellipsoid, the following approximate formula for normal gravity is often used:

$$\gamma(\phi,h) = \gamma(\phi)\left(1 - \frac{2}{a}(1 + f + m - 2f\sin^2\phi)h + \frac{3}{a^2}h^2\right), \tag{71}$$

which is accurate to 1.4×10^{-6} m/s^2, or better, up to altitude, $h = 20{,}000$ m, above the ellipsoid.

Earth's internal gravitational field

While Earth's external gravitational field may be determined entirely from measurements and potential theory,

Gravity Field of the Earth, Table 2 Defining parameters for normal ellipsoids of geodetic reference systems

Reference system[a]	a (m)	J_2	GM (m^{3}s^{-2})	ω_e (rad s^{-2})
International ref system (1930)	6,378,388	1.0920E-03[b]	3.98633E14[b]	7.2921151E-05
GRS67	6,378,160	1.0827E-03	3.98603E14	7.2921151467E-05
GRS80	6,378,137	1.08263E-03	3.986005E14	7.292115E-05
WGS84	6,378,137	1.08262982131E-03	3.986004418E14	7.2921151467E-05
Best current values[c]	6,378,136.7 ± 0.1 (mean-tide system)	(1.0826359 ± 0.0000001)E-03 (zero-tide system)	(3.986004418 ± 0.000000008)E14 (includes atmosphere)	7.292115E-5 (mean value)

[a]GRS = Geodetic Reference System; WGS = World Geodetic System
[b]Derived values from defining parameters, $f = 1/297$ and $\gamma_a = 9.780490$ m/s^2
[c]Groten (2004)

Gravity Field of the Earth, Table 3 Normal gravity harmonic coefficients corresponding to GRS80 (Moritz, 2000)

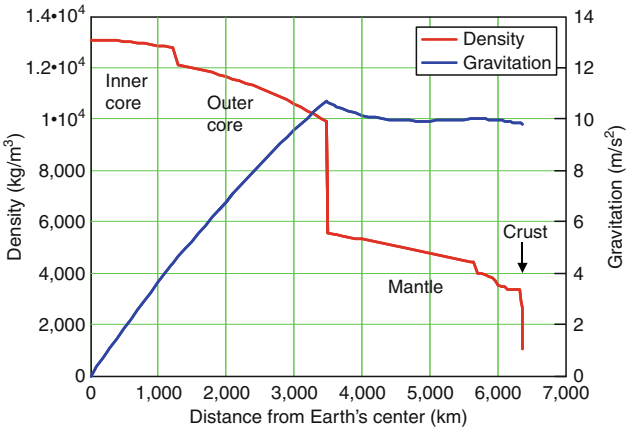
n	J_{2n}
1	1.08263E-03
2	−2.37091222E-06
3	6.08347E-09
4	−1.427E-11
5	1.2 E-14

the internal field is much less accessible for measurement and depends on models of the subsurface structure that can only be inferred from surface measurements and constraints provided by various other geophysical observations. Some have likened the study of the Earth's interior to the study of the cosmos – the investigator is incapable of performing in situ measurements and is left to use various (and sometimes ingenious) proxies to infer the Earth's interior makeup. Of course, there are borehole measurements and these are probing ever deeper into the Earth's crust, but they are isolated and expensive. A complete reconstruction of the Earth's (deeper) interior comes at present primarily from seismological data by analyzing the travel times of hundreds of thousands to millions of compressional (P) and shear (S) waves as they pass through the Earth's body. Near surface reconstruction also relies on seismic sounding (seismograms from intentional explosions) as well gravimetry, magnetometry, and other geophysical sensors, including ground penetrating radar, electromagnetic induction, and electric resistivity methods.

We may argue that knowledge of the interior gravity field is of secondary importance to an understanding of Earth's internal structure inasmuch as this knowledge is derived from an Earth model that is already determined from other data. Only when gravity can be measured within the Earth (e.g., in boreholes) does the interior gravitational field offer a contributory importance to our modeling of the Earth's interior. At present the deepest borehole in the world, at just over 12,260 m, is located in Russia on the Kola Peninsula on the Barents Sea northeast of Finland. Most holes are drilled for geological exploration and reach depths of several hundred meters.

Nevertheless it is of interest to compute gravitational attraction along a radius of the Earth, as modeled by a radially layered composite of mass density. From Equation 8 it is easily shown that the gravitational potential inside a radially symmetric body (having radius, R) has the form

$$V(r) = \frac{4\pi G}{r}\int_0^r \rho(r')r'^2 dr' + 4\pi G\int_r^R \rho(r')r'dr', \quad 0 \le r \le R. \tag{72}$$



Gravity Field of the Earth, Figure 3 Gravitation and mass density inside the Earth according to the radially symmetric PREM.

Thus, the gravitational acceleration is given by

$$g(r) = -\frac{d}{dr}V(r) = \frac{4\pi G}{r^2}\int_0^r \rho(r')r'^2 dr'. \tag{73}$$

For the Preliminary Reference Earth Model (PREM) of 1981 (Dziewonski and Anderson, 1981; Dziewonski, 1989), which consists of approximately 80 layers, the radial variation of gravitation is shown in Figure 3 together with the change in density.

On a more local scale, we may use the attraction of an infinite horizontal plate of constant density, ρ_0, as a model to approximate the value of gravity in the crust (or ocean) near the Earth's surface. The attraction of such a plate of thickness, d, anywhere above or below the plate is

$$B = 2\pi G\rho_0 d, \tag{74}$$

which is also known as the Bouguer plate effect. Suppose that g_P is the value of gravity at a surface point, P, on the Earth. Assuming that the topography is relatively flat in its vicinity, remove a plate of material with density, ρ_0, and thickness, d, and the gravity value at P will diminish by B. Now bring the gravity value to the level d through free space using a normal gradient approximation, which changes its value by $F \approx -(\partial\gamma/\partial h)d$ (also known as the free-air reduction). Finally restore the material plate, which again decreases the value of gravity, now at depth, d, by B.

The resulting gravity as a function of depth thus becomes

$$g(d) \approx g_p - \left(\frac{\partial\gamma}{\partial h} + 4\pi G\rho_0\right)d. \tag{75}$$

Using nominal values for the normal gradient, $\partial\gamma/\partial h \approx -0.3086$ mGal/m (where 1 mGal $= 10^{-5}$ m/s^2), and

the ocean water or crust density, $\rho_w = 1{,}027\ \mathrm{kg/m^3}$ or $\rho_0 = 2{,}670\ \mathrm{kg/m^3}$, respectively, this becomes

$$g(d) \approx g_p + 0.2225\ d\ [\mathrm{mGal}], \quad d \text{ in units of meter below sea surface;} \tag{76}$$

$$g(d) \approx g_p + 0.0848\ d\ [\mathrm{mGal}], \quad d \text{ in units of meter below land surface.} \tag{77}$$

This is also known as the Poincaré–Prey reduction and may be used to model gravity in boreholes and in the oceans, given gravity at the surface. The model of gravity within the Earth's upper crust (or ocean) can be improved with more careful modeling of the effects of variable topography and variable density (especially in the oceans).

Summary

The Earth's gravitational field, defined as the field generated by the subsurface mass density, embodies a wealth of geophysical information at global and local scales. The global field is often described in terms of a spherical harmonic expansion, where the low-degree harmonics have a definite (though incomplete) interpretation with respect to the moments of the density distribution. The attenuation of the potential in the spherical harmonic (spatial frequency) domain is sufficiently gradual that first- and second-order radial derivatives readily characterize the local, near surface structure of the Earth. Whereas the external gravitational field of the Earth can be modeled in detail on the basis of surface measurements and potential theory, the deeper internal gravitational field model is derived secondarily from density distribution models obtained from other sources, such as seismological data.

Bibliography

Chao, B. F., 2005. On inversion for mass distribution from global (time-variable) gravity field. *Journal of Geodynamics*, **29**, 223–230.

Cushing, J. T., 1975. *Applied Analytical Mathematics for Physical Scientists*. New York: Wiley.

Dziewonski, A. M., 1989. Earth structure, global. In James, D. E. (ed.), *The Encyclopedia of Solid Earth Geophysics*. New York: Van Nostrand Reinhold, pp. 331–358.

Dziewonski, A. M., and Anderson, D. L., 1981. Preliminary Reference Earth Model (PREM). *Physics of the Earth and Planetary Interiors*, **25**, 297–356.

Freeden, W., Gervens, T., and Schreiner, M., 1998. *Constructive Approximation on the Sphere, with Applications in Geomathematics*. Oxford: Clarendon.

Groten, E., 2004. Fundamental parameters and current (2004) best estimates of the parameters of common relevance to astronomy, geodesy, and geodynamics. *Journal of Geodesy, The Geodesist's Handbook*, **77**(10–11), 724–731.

Heiskanen, W. A., and Moritz, H., 1967. *Physical Geodesy.* San Francisco: Freeman.

Heiskanen, W. A., and Veing Meinesz, F. A., 1958. *The Earth and its Gravity Field*. New York: McGraw-Hill.

Heiland, C. A., 1940. *Geophysical Exploration*. New York: Prentics-Hall.

Helmert, F. R., 1884. *Die Mathematischen und Physikalischen Theorien der Höheren Geodäsie*. Leibzig: B.G. Teubner, Vol. 2. reprinted in 1962 by Minerva GMBH, Frankfurt/Main.

Hobson, E. W., 1965. *The Theory of Spherical and Ellipsoidal Harmonics*. New York: Chelsea.

Hofmann-Wellenhof, B., and Moritz, H., 2005. *Physical Geodesy.* Berlin: Springer Verlag.

Hotine, M., 1969. *Mathematical Geodesy.* Washington: U.S Department of Commerce.

Jekeli, C., 1988. The exact transformation between ellipsoidal and spherical harmonic expansions. *Manuscripta Geodaetica*, **14**, 106–113.

Jekeli, C., 2000. *Inertial Navigation Systems with Geodetic Applications*. Berlin: Walter deGruyter.

Jekeli, C., 2007. Potential theory and static gravity field of the earth. In Schubert, G. (ed.), *Treatise on Geophysics*. Oxford: Elsevier, Vol. 3, pp. 11–42.

Jekeli, C., 2010. Correlation Modeling of the Geopotential Field in Classical Geodesy. In Freeden, W., et al. (eds.), *Handbook of Geomathematics*, Berlin: Springer-Verlag, pp. 834–863

Kellogg, O. D., 1953. *Foundations of Potential Theory.* New York: Dover.

Martin, J. L., 1988. *General Relativity: A Guide to its Consequences for Gravity and Cosmology.* New York: Wiley.

Mohr, P. J., Taylor, B. N., and Newell, D. B., 2008. CODATA Recommended Values of the Fundamental Physical Constants: 2006. *Reviews of Modern Physics*, **80**, 633–730.

Molodensky, M. S., Eremeev, V. F., and Yurkina, M. I., 1962. *Methods for the Study of the External Gravity Field and Figure of the Earth*. Jerusalem: Israel Program of Scientific Translations. Russian original, 1960.

Moritz, H., 1980. Advanced Physical Geodesy. Wichmann, Karlsruhe (reprint 2008 by School of Earth Sciences, Ohio State University, Columbus, Ohio).

Moritz, H., 2000. Geodetic Reference System 1980. *Journal of Geodesy*, **74**(1), 128–133.

Morse, P. M., and Feshbach, H., 1953. *Methods of Theoretical Physics, Parts I and II*. New York: McGraw-Hill.

Müller, C., 1966. *Spherical Harmonics. Lecture Notes in Mathematics*. Berlin: Springer-Verlag.

Nettleton, L. L., 1976. *Gravity and Magnetics in Oil Prospecting*. New York: McGraw-Hill.

Pavlis, N. K., Holmes S. A., Kenyon, S. C., Factor, J. K., 2008. An Earth gravitational model to degree 2160: EGM2008. Presented at the General Assembly of the European Geosciences Union, Vienna, Austria, April 13–18, 2008.

Plag, H. P., and Pearlman, M. (eds.), 2009. *Global Geodetic Observing System, Meeting the Requirements of a Global Society on a Changing Planet in 2020*. Berlin: Springer.

Sansò, F., and Rummel, R. (eds.), 1997. *Geodetic Boundary Value Problems in View of the One Centimeter Geoid. Lecture Notes in Earth Sciences 65*. Berlin: Springer-Verlag.

Tapley, B. D., Bettadpur, S., Watkins, M., Reigber, C., (2004). The Gravity Recovery and Climate Experiment, mission overview and early results. *Geophysical Research Letters*, **31**(9), doi:10.1029/2004GL019920.

Tapley, B. D., Ries, J., Bettadpur, S., Chambers, D., Cheng, M., Condi, F., Gunter, B., Kang, Z., Nagel, P., Pastor, R., Pekker, T., Poole, S., and Wang, F., 2005. GGM02 – An improved Earth gravity field model from GRACE. *Journal of Geodesy*, doi:10.1007/s00190-005-0480-z.

Telford, W. M., Geldart, L. P., and Sheriff, R. E., 1990. *Applied Geophysics*, 2nd edn. Cambridge: Cambridge U. Press.

Torge, W., 1989. *Gravimetry.* Berlin: Walter deGruyter.

Cross-references

Earth Tides
Earth, Density Distribution
Geodesy, Figure of the Earth
Geodesy, Ground Positioning and Leveling
Geodesy, Physical
Geodynamics
Geoid Determination, Theory and Principles
Geoid Undulation, Interpretation
Gravimeters
Gravity Anomalies, Interpretation
Gravity Data, Advanced Processing
Gravity Data, Regional – Residual Separation
Gravity Field, Temporal Variations from Space Techniques
Gravity Measurements, Absolute
Gravity Method, Airborne
Gravity Method, Principles
Gravity Method, Satellite
Gravity Method, Surface
Gravity, Data to Anomalies
Gravity, Global Models
Gravity, Gradiometry
International Gravity Formula
Isostasy
Spherical Harmonic Analysis Applied to Potential Fields

GRAVITY FIELD, TEMPORAL VARIATIONS FROM SPACE TECHNIQUES

Anny Cazenave[1], G. Ramillien[2], Richard Biancale[2]
[1]Laboratoire d'Etudes en Géophysique et Océanographie Spatiales (LEGOS), Centre National d'Etudes Spatiales, Toulouse, France
[2]DTP, OMP, Toulouse, France

Definition and introduction

Observing the Earth from space over the past 50 years has resulted in spectacular advances in various areas of the geosciences. This is the case, in particular, for measuring the precise shape of the Earth and its broad scale internal structure from detailed measurements of the planet's gravity field. Since the early 1970s, tracking the orbits of tens of satellites at different altitudes and orbit inclinations has gradually improved knowledge of Earth's average (static) gravity field (see section on *Gravity Method, Satellite*). Besides the precise description of the static component, attempts have been made to detect temporal variations of the gravity field. For example, a secular decrease of Earth's flattening (the degree 2, order 0 term) was discovered by precise analysis of the Lageos laser satellites (Yoder et al., 1983; Cox and Chao, 2002), and attributed to postglacial rebound (PGR) (the viscoelastic response of the crust and mantle to melting of the large ice sheets that covered part of the northern hemisphere during the last ice age). Seasonal variations of the low degree terms of the Earth's gravity field expansion (up to $\sim$ degree 6) have been detected from analysis of satellite orbital perturbations (Cheng and Tapley, 2004; Chen et al., 1999). These variations were attributed to seasonal transfer of air mass inside the atmosphere as well as to annual fluctuations of snow and land water storage (e.g., Chen et al., 1999; Cazenave et al., 1999). However until recently, temporal changes of only the longest-wavelength (lowest degree) harmonic coefficients had been monitored from orbital perturbations. This limited effective spatial resolution to several thousand kilometers, because large uncertainty in higher degree coefficients prevented detection of any time-varying signal.

In 2002, the US–German GRACE (Gravity Recovery and Climate Experiment) mission has been providing a precise survey of Earth's time-variable gravity field, with unprecedented temporal and spatial sampling (Tapley et al., 2004; Wahr et al., 2004). GRACE time-variable gravity fields provide a means of measuring temporal and spatial variations of mass redistribution within the Earth system on time scales varying from season to multiyears. The GRACE mission has launched a new era in studying a series of geophysical problems ranging from deep Earth structure to tracking mass redistribution on and near the surface of the Earth. GRACE has greatly improved understanding of mass redistribution in various compartments of the climate system (atmosphere, oceans, terrestrial water, stores and cryosphere). GRACE has fundamentally enriched a number of fields, including (but not limited to) the global water cycle and land hydrology, mass balance of polar ice sheets and mountain glaciers, ocean mass and global sea-level change, and solid Earth geophysics.

The GRACE mission

Launched in March 2002, GRACE utilizes a state-of-the-art technique to observe variation of Earth's gravity by tracking the inter-satellite range and the range rate between two coplanar, low altitude satellites via a K-band ranging system. In addition, each satellite is equipped with a SuperSTAR Accelerometer, GPS receiver/antenna, Star Cameras, and Laser Retro Reflector to complement the science instruments. The GRACE Science Data System uses the range and range rate data, along with ancillary data, to estimate a new gravity field every month, in the form of corrections to a well-defined background gravity model used in the data processing procedure (Bettadpur, 2007). The monthly sampling rate has been selected in order to accumulate sufficient observations to ensure spatial resolution of about 400 km. Generally there is a trade-off in selecting the temporal sampling interval. Accumulating data over longer time intervals increases spatial resolution, but decreases temporal resolution and vice versa.

Modeling temporal geoid variations

The static component of the gravity field G_0 corresponds to nearly 99% of the total field, mainly due to solid Earth contributions. This term can be easily evaluated and removed by computing the temporal mean of a long enough series of GRACE monthly geoids, or alternatively

considering a single geoid computed with a long period of time of satellite observations. In this study, the monthly time-variable geoid $\delta G(t)$ is merely computed as the difference between the monthly geoid $G(t)$ measured by GRACE at time t, and the static mean field component:

$$\delta G(t) = G(t) - G_0. \tag{1}$$

Thus, the corresponding geoid differences (also called monthly GRACE geoids in the following) only reflect short-term geoid change associated with mass redistributions inside the Earth system.

Let $\delta C_{nm}(t)$ and $\delta S_{nm}(t)$ be the time-variable "normalized" Stokes coefficients of the spherical harmonic expansion of geoid height, where n and m are the degree and order, respectively. Thus, the time-variable geoid is classically expressed as (Heiskanen and Moritz, 1967):

$$\delta G(t) = \sum_{n=1}^{N}\sum_{m=0}^{n} (\delta C_{nm}(t)\cos(m\lambda) + \delta S_{nm}(t)\sin(m\lambda))P_{nm}(\cos\theta), \tag{2}$$

where N is the maximum degree of the harmonic decomposition, θ is the colatitude, λ is the longitude, and P_{nm} is the associated Legendre polynomial that is dimensionless.

GRACE geoids are generally further expressed in surface mass change (usually in equivalent water height) assuming that mass redistribution occurs in thin surface layers compared to the Earth's dimension (Swenson and Wahr, 2002), so that:

$$\begin{pmatrix}\delta C_{nm}(t)\\ \delta S_{nm}(t)\end{pmatrix} = \frac{(1+z_n)\mathrm{Re}^2}{(2n+1)M} \int\int_S \delta q(\theta,\lambda,t)\begin{pmatrix}\cos\\ \sin\end{pmatrix}(m\lambda)P_{nm}(\cos\theta)\delta S, \tag{3}$$

where $\delta q(\theta, \lambda, t)$ is a surface density of water/ice/air mass, expressed in terms of equivalent water thickness at time t, whose harmonic coefficients, $\delta A_{nm}(t)$ and $\delta B_{nm}(t)$, can be evaluated from the corresponding geoid anomaly coefficients through the linear filtering:

$$\begin{pmatrix}\delta C_{nm}(t)\\ \delta S_{nm}(t)\end{pmatrix} = W_n\begin{pmatrix}\delta A_{nm}(t)\\ \delta B_{nm}(t)\end{pmatrix}. \tag{4}$$

W_n is an isotropic spatial filter that weights the surface density coefficients, and whose analytical expression is (Swenson and Wahr, 2002; Ramillien et al., 2005):

$$W_n = \frac{2\pi G\mathrm{Re}\rho_W}{(2n+1)\gamma(\theta)}(1+z_n). \tag{5}$$

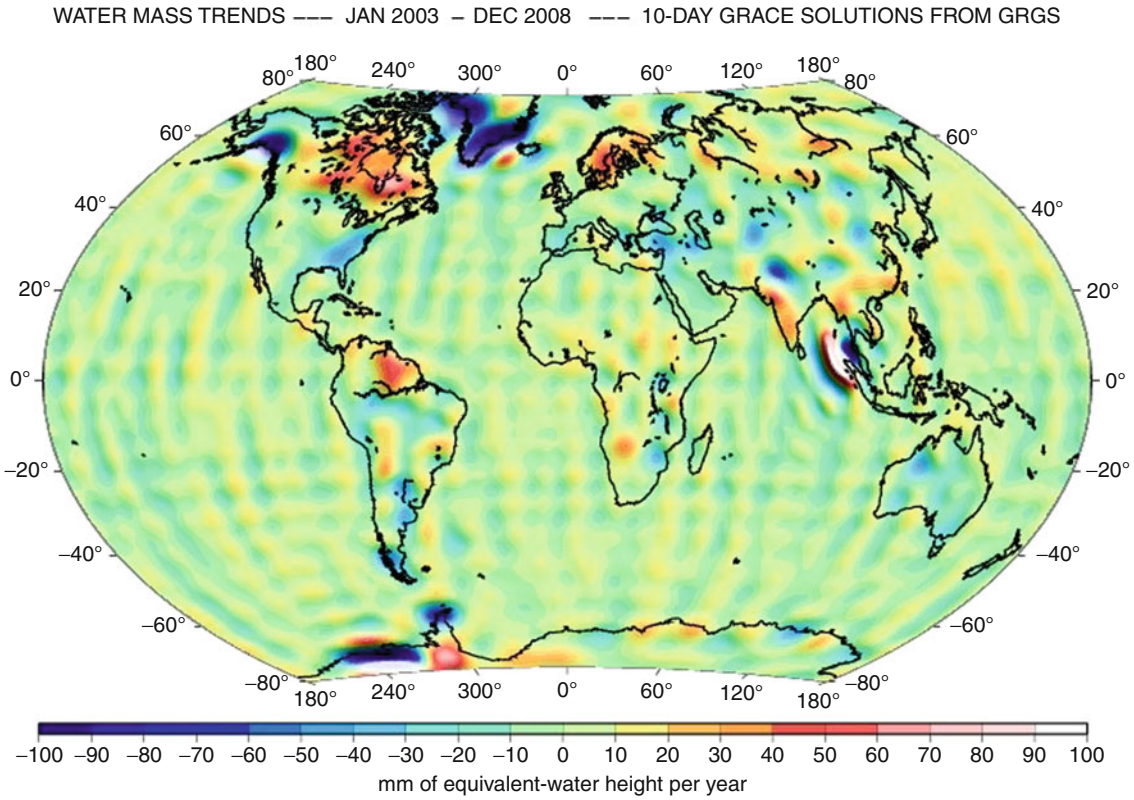
In Equations 3 and 5 above, z_n represents the Love numbers that takes into account instantaneous elastic compensation of the Earth' surface to the load; $\gamma(\theta)$ is the normal gravity on the reference ellipsoid at the colatitude θ. G ($\sim$6,67.10^{-11} $m^3kg^{-1}s^{-2}$) is the gravitational constant, M is the total mass of the Earth ($\sim$5,97602.10^{24} kg), and Re ($\sim$6,378 km) is mean Earth's radius. ρ_W ($\sim$1,000 kgm^{-3}) is the water density.

Equations 4 and 5 are used to compute surface density variations $\delta q(\theta, \lambda, t)$ (or mass redistribution due to air, water, and ice mass change in surface fluid envelopes) from the monthly geoid harmonic coefficients.

GRACE data errors

Assessing the errors of GRACE monthly solutions is a major challenge. Two categories of errors are recognized: (1) errors introduced by data processing (errors of GRACE measurements and models used to remove other geophysical signals) and (2) post-processing errors (e.g., spatial smoothing of the GRACE solutions to remove high-frequency noise and leakage errors due to signal contamination from regions outside the studied area). Among errors in category (1) correlated errors due to un-modeled effects and contamination of other geophysical signals during data processing are highly problematic. Because of the limited spatiotemporal sampling provided by the GRACE satellites, these errors are aliased into the estimated spherical harmonic coefficients, giving rise to north–south stripes evident in GRACE geoid maps. To reduce this organized spatial noise, different smoothing methods have been proposed (category 2 errors), which significantly affects GRACE estimates of surface mass change by reducing signal amplitude at the higher degrees and orders suppressed by these filters. Another post-processing error arises from the waveband-limited spherical harmonics considered in the solution (typically 50 or 60, corresponding to spatial resolution of 350–400 km). The result is contamination from surrounding regions ("leakage"). The GRACE orbital configuration (altitude and inter-satellite distance) and the need for spatial filtering are the main limitations to spatial resolution, estimated to be on the order of 400 km. Current accuracy of monthly GRACE solutions is estimated to $\sim$2 cm equivalent water layer over areas exceeding 1 million km^2 (e.g., Wahr et al., 2006).

Figure 1 shows a map of global mass trends from GRACE, over a full 6-year time span (January 2003 through December 2008), computed from best-fit linear trends to time series at every location (based on GRACE solutions computed by GRGS, Biancale et al., 2006). Units are in equivalent water height per year. On land, this map shows predominantly water mass gain or loss in major river basins and ice mass loss over ice-covered regions (e.g., the Greenland and west Antarctica ice sheets, Alaska glaciers, etc.). The large positive mass trends over Northern Canada and Fennoscandia are due to postglacial rebound (the solid Earth response to last deglaciation). The signature of the 2004 Sumatra-Andaman Earthquake is also well visible.



Gravity Field, Temporal Variations from Space Techniques, Figure 1 Map of surface mass trends (expressed in millimeters of equivalent water height) estimated from GRACE data between January 2003 and December 2008. Blue color corresponds to negative mass trend (mass loss) while red color represents mass increase (Data updated from Biancale et al., 2006).

Applications of GRACE

In this section are presented a few examples of present-day mass variations detected by GRACE over continental river basins, Greenland and Antarctica ice sheets, and the oceans. The capability of GRACE in measuring solid Earth mass changes (e.g., postglacial rebound and co/post-seismic deformations due to large earthquakes) is also briefly discussed.

Hydrology

The global water cycle that exchanges water among oceans, atmosphere, and land plays a key role in the physical and chemical processes that influence Earth's climate and its change over time. On land, terrestrial water storage (TWS) change, as a major component of the global water cycle, reflects changes in water stored in soil, snow over land, and ground-water reservoirs, and is closely connected to accumulated precipitation (P), evapotranspiration (ET), and surface and subsurface runoff (R) within a given area or basin, through the water balance equation relating a change in TWS in unit time Δt:

$$\frac{\Delta(TWS)}{\Delta t} = P - ET - R. \quad (6)$$

For many areas (not covered by snow), TWS is virtually a measure of total water content in surface stores, soil layers, and ground-water reservoirs, and is also a good indicator of abnormal climate conditions, such as droughts and floods. Accurately quantified TWS change provides a key measure of the continental water cycle and available water resources in a given region or river basin. However, estimating TWS change is difficult because limited ground water and soil moisture observations are available, and often are simply nonexistent (e.g., Rodell and Famiglietti, 1999) at basin or smaller scales.

In recent years, GRACE data have been successfully applied to the study of global and basin scale TWS changes (e.g., Ramillien et al., 2008a). Early studies focused on the seasonal cycle and showed good agreement with global hydrological models. Now exceeding 7 years in length, the GRACE time series provides some measure of interannual variability in TWS. For example, several studies have detected drought events in different regions of the world (e.g., Amazonia in 2005, Chen et al., 2009a; East Africa in 2005, Swenson and Wahr, 2009; Australia in 2006, Leblanc et al., 2009). Anthropogenic effects on the terrestrial water cycle, such as ground-water pumping, are also detectable by GRACE (e.g., Rodell et al., 2009;

Tiwari et al., 2009). Other studies have used GRACE TWS in combination with additional information to estimate changes in vertical or horizontal fluxes, for example, P-ET, ET, or R, of the right-hand side of the water balance equation (Equation 6).

Ice sheets mass change

The mass balance of ice sheets is of considerable interest in the context of global warming and present-day sea-level rise. GRACE satellite gravity measurements offer the opportunity to study the mass balance of polar ice sheets from a new perspective, using GRACE gravity change to directly measure mass variation or redistribution. A series of recent studies has created a clear picture of present-day ice loss from the two largest ice sheets on Earth and demonstrated the unique potential of satellite gravimetry for monitoring large-scale ice mass change. Using extended GRACE time series (6–7 years), studies indicate a clear acceleration of ice loss from both ice sheets since about 2006. For example, estimates by Velicogna (2009) over Greenland indicate an ice loss rate of 267–286 Gt/yr during 2006–2009, compared with a loss rate of 137–159 Gt/yr during the period 2002–2005. Over Antarctica, Chen et al. (2009b) find a mass loss of 220–246 Gt/yr during the period 2006–2009, an alarming acceleration of the already large loss rates of 104–144 Gt/yr estimated for 2002–2005. On average over 2002–2009, ice mass loss reported by GRACE over the two ice sheets corresponds to ~1 mm/yr sea-level rise, that is, about 1/3 the observed rate of rise (Cazenave and Llovel, 2010, Cazenave and Chen, 2010).

Ocean mass change

Net water flux into and out of the ocean causes its mass to change. Such changes cause gravitational variations detectable by GRACE (e.g., Chambers et al., 2004). This has allowed, for the first time, direct estimates of the global ocean mass contribution to sea-level change. However, GRACE is also sensing postglacial rebound. On average over the oceanic domain, this effect is large, and of the same order of magnitude as the ocean mass change due to current land ice melt itself. As postglacial rebound models vary significantly from each other with respect to this effect (e.g., Paulson et al., 2007; Peltier, 2009), GRACE ocean mass rates combined with land ice melt rates can place an independent constraint on the postglacial rebound correction.

Satellite altimetry allows precise measurement of present-day global mean sea-level change (e.g., Cazenave and Llovel, 2010). When combined with the GRACE-based ocean mass change contribution, the steric component (i.e., effects of vertically integrated sea-water temperature and salinity changes) to sea level can be deduced (e.g., Chambers, 2006; Lombard et al., 2007). The results appear to be comparable to estimates based on in situ hydrographic (temperature and salinity) data.

Postglacial rebound

Large-scale readjustment of Earth's crust and mantle (postglacial rebound) follows melting of ice loads. Postglacial rebound induces secular changes in the gravity field observable by GRACE. The large positive mass trend centered over the Hudson Bay (Canada) in Figure 1 is one of the most prominent signatures of this process. Comparison between GRACE and other measures of postglacial rebound (e.g., surface gravity change, vertical crustal motions, model estimates of ice-load extent, and melting history) has confirmed that GRACE is detecting Earth's recovery following melting of the Laurentide ice sheet at the end of the last ice age (e.g., Peltier, 2009; Wu et al., 2010).

Co-seismic and post-seismic deformations due to large earthquakes

Besides postglacial rebound (Peltier, 2009) and the refinement of Earth's static gravity field (Pavlis et al., 2008), GRACE has contributed to the understanding of other solid Earth processes. In particular, GRACE appears able to detect the sea-floor deformation, afterslip, and viscoelastic relaxation associated with large earthquakes. This is the case with the Sumatra-Andaman earthquake (magnitude of 9.3) of December 26, 2004, the largest recorded seismic event in about 40 years. GRACE successfully detected the gravity change associated with the corresponding co-seismic subduction and uplift, in agreement with model predictions (e.g., Han et al., 2006; Panet et al., 2007).

Summary

Since 2002, the US–German GRACE (Gravity Recovery and Climate Experiment) mission has been providing a precise survey of Earth's time-variable gravity field, with unprecedented temporal and spatial sampling. GRACE time-variable gravity fields provide a means of measuring temporal and spatial variations of mass redistribution within the Earth system. The GRACE mission has launched a new era in studying a series of geophysical problems ranging from deep Earth structure to tracking mass redistribution on and near the surface of the Earth. GRACE has greatly improved understanding of mass redistribution in various compartments of the climate system (atmosphere, oceans, terrestrial water, and cryosphere) (Ramillien et al., 2008b; Cazenave and Chen, 2010). GRACE has fundamentally enriched a number of fields, including (but not limited to) the global water cycle and land hydrology, mass balance of polar ice sheets and mountain glaciers, ocean mass and global sea-level change, and solid Earth geophysics. While such novel data offer new possibilities for studying regional to global scale mass redistributions occurring inside and at the surface of the Earth, they complement measurements performed by in situ gravity meters (absolute and relative, including super conducting gravity meters) for detecting a variety of local phenomena such as volcano deformations,

seismic activity, land subsidence, water table movement and water storage change, etc. (see examples in Budetta and Carbone, 1998; Mrlina et al., 2003; Zerbini et al., 2007; Richter et al., 2004; Anderson et al., 2005).

Bibliography

Anderson, O. B., Seneriratne, S. I., Hinderer, J., and Viterbo, P., 2005. GRACE (derived terrestrial water storage depletion associated with the 2003 European heat wave. *Geophysical Research Letters*, **32**, L18405, doi:10.1029/2005GL026574.

Bettadpur, S., 2007. CSR level-2 processing standards document for product release 04, GRACE 327-742, The GRACE Project, Center for Space Research, University of Texas at Austin.

Biancale, R., Lemoine, J.-M., Balmino, G., Loyer, S., Bruisma, S., Perosanz, F., Marty, J.-C., and Gégout, P., 2006. 3 years of geoid variations from GRACE and LAGEOS data at 10-day intervals from July 2002 to March 2005. CNES/GRGS product, data available on CD-ROM.

Budetta, G., and Carbone, D., 1998. Temporal variations in gravity at Mt Etna (Italy) associated with the 1989 and 1991 eruptions. *Bulletin of Volcanology*, **59**, 311–326.

Cazenave, A., and Chen, J., 2010. Time-variable gravity from space and present-day mass redistribution in the Earth system. *Earth and Planetary Science Letters*, **298**, 263–274. (published online 24 August 2010).

Cazenave, A., and Llovel, W., 2010. Contemporary sea level rise. *Annual Review of Marine Science*, **2**, 145–173.

Cazenave, A., Mercier, F., Bouille, F., and Lemoine, J. M., 1999. Global-scale interactions between the solid Earth and its fluid envelopes at the seasonal time scale. *Earth and Planetary Science Letters*, **171**, 549–559.

Chambers, D. P., 2006. Observing seasonal steric sea level variations with GRACE and satellite altimetry. *Journal of Geophysical Research*, **111**(C3), C03010, doi:10.1029/2005JC002914.

Chambers, D. P., Wahr, J., and Nerem, R. S., 2004. Preliminary observations of global ocean mass variations with GRACE. *Geophysical Research Letters*, **31**, L13310, doi:10.1029/2004GL020461.

Chen, J. L., Wilson, C. R., Eanes, R. J., and Tapley, B. D., 1999. Geophysical contributions to satellite nodal residual variation. *Journal of Geophysical Research*, **104**(B10), 23,237–23,244.

Chen, J. L., Wilson, C. R., Tapley, B. D., Yang, Z. L., and Niu, G. Y., 2009a. 2005 drought event in the Amazon River basin as measured by GRACE and estimated by climate models. *Journal of Geophysical Research*, **114**, B05404, doi:10.1029/2008JB006056.

Chen, J. L., Wilson, C. R., Blankenship, D. D., Tapley, B. D. 2009. Accelerated Antarctic ice loss from satellite gravity measurements. *Nature Geosciences*, doi:10.1038/NGEO694.

Cheng, M., and Tapley, B. D., 2004. Variations in the Earth's oblateness during the past 28 years. *Journal of Geophysical Research*, **109**, B09402, doi:10.1029/2004JB003028.

Cox, C. M., and Chao, B. F., 2002. Detection of a large-scale mass redistribution in the terrestrial system since 1998. *Science*, **297**, 831–833.

Han, S.-C., Shum, C. K., Bevis, M., Ji, C., and Kuo, C.-Y., 2006. Crustal dilatation observed by GRACE after the 2004 Sumatra-Andaman earthquake. *Science*, **313**, 658–662, doi:10.1126/science.1128661.

Heiskanen, W. A., and Moritz, H., 1967. *Physical Geodesy.* San Francisco: Freeman.

Leblanc, M. L., Tregoning, P., Ramillien, G., Tweed, S. O., and Fakes, A., 2009. Basin-scale, integrated observations of the early 21st century multiyear drought in southeast Australia. *Water Resources Research*, **45**, W04408, doi:10.1029/2008WR007333.

Lombard, A., Garcia, D. E., Cazenave, A., Ramillien, G., and Flechtner, F., 2007. Estimation of steric sea level variations from combined GRACE and Jason-1 data. *Earth and Planetary Science Letters*, **254**, 194–202, doi:10.1016/j.epsl.2006.11.035.

Mrlina, J., Radwan, A. H., Hassan, R. M., Mahmoud, S. M., Tealeb, A. A., and Issawy, E. A., 2003. Temporal variations of gravity in the Aswan region of Egypt. *Journal of Geodynamics*, **35**, 499–509.

Panet, I., Mikahailov, V., Diament, M., Pollitz, F., King, J., de Viron, O., Holschneider, M., Biancale, R., and Lemoine, J. M., 2007. Co-seismic and post seismic signatures of the Sumatra 2004 December and 2005 March earthquakes in GRACE satellite gravity. *Geophysical Journal International*, **171**, 177–190.

Paulson, A., Zhong, S., and Wahr, J., 2007. Limitations on the inversion for mantle viscosity from postglacial rebound. *Geophysical Journal International*, **168**, 1195–1209, doi:10.1111/j.1365-246X.2006.03222.x.

Pavlis, N. K., Holmes, S. A., Kenyon, S. C., and Factor, J. K., 2008. An Earth gravitational model to degree 2160: EGM2008; presented at the 2008 General Assembly of the European Geosciences Union, Vienna, Austria, April 13–18. WYRZYKOWSKI T. 1988, Monograph of precise leveling networks in Poland (in polish). IGiK, Warsaw.

Peltier, W. R., 2009. Closure of the budget of global sea level rise over the GRACE era: the importance and magnitudes of the required corrections for global isostatic adjustment. *Quaterly Science Reviews*, **28**, 1658–1674.

Ramillien, G., Frappart, F., Cazenave, A., and Guentner, A., 2005. Change in land water storage from 2 years of GRACE satellite data. *Earth and Planetary Science Letters*, **235**, 283–301.

Ramillien, G., Bouhours, S., Lombard, A., Cazenave, A., Flechtner, F., and Schmidt, R., 2008a. Land water contributions from GRACE to sea level rise over 2002–2006. *Global and Planetary Change*, **60**, 381–392.

Ramillien, G., Famiglietti, J. S., and Wahr, J., 2008b. Detection of continental hydrology and glaciology signals from GRACE: a review. *Surveys in Geophysics*, **29**, 361–374.

Richter, B., Zerbini, S., Matoni, F., and Simon, D., 2004. Long-term crustal deformation monitored by gravity and space techniques at Medicina, Italy and Wettzell, Germany. *Journal of Geodynamics*, **38**, 281–292.

Rodell, M., and Famiglietti, J. S., 1999. Detectability of variations in continental water storage from satellite observations of the time dependent gravity field. *Water Resources Research*, **35**, 2705–2724, doi:10.1029/1999WR900141.

Rodell, M., Velicogna, I., and Famiglietti, J. S., 2009. Satellite-based estimates of groundwater depletion in India, *Nature*, **460**, doi:10.1038/nature08238.

Swenson, S., and Wahr, J., 2002. Methods for inferring regional surface mass anomalies from Gravity Recovery and Climate Experiments (GRACE) measurements of time-variable gravity. *Journal of Geophysical Research*, **107**, 2193, doi:10.1029/2001JB000576.

Swenson, S., and Wahr, J., 2009. Monitoring the water balance of Lake Victoria, East Africa, from space. *Journal of Hydrology*, **370**, 163–176.

Tapley, B. D., Bettadpur, S., Watkins, M. M., and Reigber, C., 2004. The gravity recovery and climate experiment: mission overview and early results. *Geophysical Research Letters*, **31**(9), L09607, doi:10.1029/2004GL019920.

Tiwari, V. M., Wahr, J., and Swenson, S., 2009. Dwindling groundwater resources in northern India from satellite gravity observations. *Geophysical Research Letters*, **36**, L18401, doi:10.1029/2009GL039401.

Velicogna, I., 2009. Increasing rates of ice mass loss from the Greenland and Antarctic ice sheets revealed by GRACE.

Geophysical Research Letters, **36**, L19503, doi:10.1029/2009GL040222.
Wahr, J., Swenson, S., Zlotnicki, V., and Velicogna, I., 2004. Time-variable gravity from GRACE: first results. *Geophysical Research Letters*, **31**, L11501, doi:10.1029/2004GL019779.
Wahr, J., Swenson, S., and Velicogna, I., 2006. Accuracy of GRACE mass estimates. *Geophysical Research Letters*, **33**, L06401.
Wu, X., et al., 2010. Simultaneous estimation of global present-day water transport and glacial isostatic adjustment, *Nature Geosciences*, doi:10.1038/ngeo938, (published online August 15 2010).
Yoder, C. F., Williams, J. G., Dickey, J. O., Schultz, B. E., Eanes, R. J., and Tapley, B. D., 1983. Secular variation of Earth's gravitational harmonic J2 coefficient from Lageos and nontidal acceleration of Earth rotation. *Nature*, **307**, 57–762, doi:10.1038/303757a0.
Zerbini, S., Richter, B., Rocca, F., van Dam, T., and Matoni, F., 2007. A combination of space and terrestrial geodetic techniques to monitor land subsidence: case study, the southeastern Po Plain, Italy. *Journal of Geophysical Research, Solid Earth*, **112** (B5), B05401.

Cross-references

Geodesy, Physical
Geoid Undulation, Interpretation
Gravity Field of the Earth
Gravity Method, Satellite
Gravity Modeling, Theory and Computation
Gravity, Global Models
Spherical Harmonic Analysis Applied to Potential Fields

GRAVITY FIELD, TIME VARIATIONS FROM SURFACE MEASUREMENTS

Virendra M. Tiwari[1], Jacques Hinderer[2]
[1]Gravity and Magnetic Studies Group, National Geophysical Research Institute (CSIR), Hyderabad, Andhra Pradesh, India
[2]Institut de Physique du Globe de Strasbourg, CNRS, Université de Strasbourg, Strasbourg, France

Synonyms

Earth's dynamic gravity field; Temporal variations of the Earth's gravity

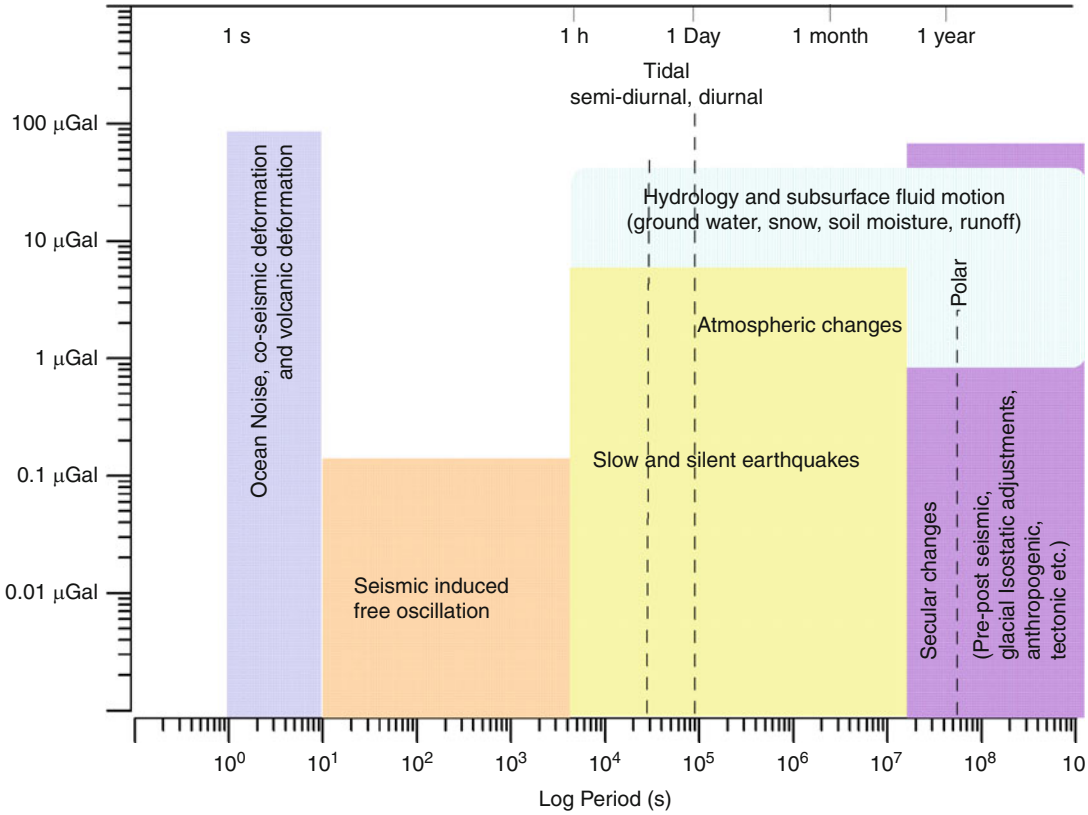
Definition

Temporal gravity changes. Variation of Earth's gravity field with time due to various tidal, oceanic, atmospheric, hydrological, glaciological, and tectonic processes.
Tidal gravity changes. The periodic variation of gravity field due to the attraction of Sun and Moon.
Non-tidal gravity changes. Gravity changes unrelated to the attraction of Sun and Moon.

Introduction

The gravity field of the Earth is an attraction force exerted by the Earth on any unit mass normal to the equipotential surface. It is a function of mass distribution and its spatial location, inside, on, and above the Earth's surface, following the Newton's law of gravitation. Continuous changes in the Earth system and its fluid envelops related to the solid and ocean tides, oceanic circulation, atmospheric changes, hydrological changes, Earth's polar motion, Earth's inner and fluid core motions, redistribute the mass and produce variations in the Earth's gravitational field on a large range of spatial and temporal scales (Figure 1). The variation in the gravity field on timescales of thousands to millions of years caused by redistribution of mass due to the processes that shape the Earth's surface is termed "static gravity field" as this change does not occur over human timescale.

Temporal variation of Earth's gravity field is determined by measurements of the gravitational field at the same location with time. Surface measurements of time-varying gravity field started in the nineteenth century using spring type relative gravity meters and were mainly employed to study tidal related gravity changes. Gravity changes can now be precisely recorded by terrestrial measurements using absolute gravity meters (AG), superconducting gravity meter (SG), relative gravity meter, or a combination of them. AG is the most precise instrument available that can measure the absolute gravity field with an accuracy of 1–3 μGal (1 μGal = 10^{-8} m s^{-2} = 10 nm s^{-2}; Niebauer et al., 1995). SG can measure changes in the Earth's gravity field with 0.1 μGal accuracy (Hinderer et al., 2007). A strong effort to monitor the time-varying gravity field globally has been made recently with the launch of the CHAMP and GRACE satellites (Wahr, 2007). Precise measurements of the changes in the Earth's gravity field either by satellite (NRC, 1997; Wahr, 2007) or from surface measurements (Hinderer and Crossley, 2000) thus offer the possibilities to constrain the underlying causes of dynamic changes of mass distribution in the Earth system processes. Both measuring methods have some advantages and drawbacks. For example, localized gravity changes (i.e., effect due to injection of water in oil field) cannot be detected in satellite measurements and similarly gravity changes occurring over a large region, particularly over remote areas cannot be mapped using surface measurements. An attempt is made to intercompare the satellite-derived gravity fields with surface gravity measurements observed with superconducting gravimeters (SG) in the Global Geodynamics Project (GGP; http://www.eas.slu.edu/GGP/ggphome.html) despite the fact that the ground and space gravity measurements are not sensitive to the same effects. Radial displacement is important for ground observations but has no effect on satellites and ground gravity measurement is a point measurement while satellite measurements are integrated response over a large region (typically a few hundreds of km; Wahr, 2007). Nevertheless, first results of intercomparison of gravity changes observed by the European SG network and GRACE satellite data reported a fair agreement in the time-varying gravity field originated due to hydrological changes (Crossley and Hinderer, 2005; Hinderer et al., 2006).



Gravity Field, Time Variations from Surface Measurements, Figure 1 Schematic diagram showing temporal gravity changes that can be studied from the surface gravity measurements (1 μGal = 10^{-8} m s^{-2} = 10 nm s^{-2}).

Concepts and applications

Primarily there are two categories of time-varying gravity signals; one that varies periodically and another one that is non-periodic. Periodic gravity changes such as tides can be separated from the observations based on their periodicity but non-periodic changes that arise from combinations of different contributing sources are difficult to recognize. However, they are identified using other complementary information. There are about 30 SGs in the world, which are continuously measuring gravity changes and form an international network called GGP (GGP Project; Crossley et al., 1999). GGP stations are sparsely distributed worldwide with a very few stations in the southern hemisphere. These data are shared within network for various applications and are also used for comparison with satellite-derived gravity changes that allows upscaling of terrestrial data and downscaling of satellite data (Hinderer et al., 2007) for broader understanding of causes of gravity changes. AG measurements are also made at these locations for calibrations of SGs and for cross validation of secular changes. Both instruments (AG and SG) are essentially laboratory instruments; however, AG can be transported but SG is unsuitable for relocation. A combination of different surface gravity measurements is used to estimate spatiotemporal changes in Earth's gravity field. Details of surface gravity measurements, their accuracy, and potential application, as suggested by Crossley and Hinderer (2005) with minor modifications, is summarized and presented in Table 1.

The following sections present a brief description of the main sources of gravity changes that are recorded by surface gravity measurements and used to study the sources responsible for these changes. An example of different contributions (solid tides, polar motion and length of the day, atmospheric loading, tidal and non-tidal ocean loading, soil moisture and snow loading) to the time-varying gravity in Strasbourg is presented in Figure 2. The amplitude of the solid tides and the ocean tidal loading would be ten times larger compared to plotted amplitudes in the Figure 2, where only long-period tides are considered.

Tides

Tidal gravity changes are caused by the attraction of other planetary bodies like the Sun and Moon. The attraction of these planetary bodies deforms the Earth's solid and fluid surface and produces changes in the gravity field. These changes are periodic, mainly semi-diurnal and diurnal, but also with long-period components (fortnightly, monthly, half-yearly, and yearly as shown in Figure 2). The semi-diurnal and diurnal tides can be very large

Gravity Field, Time Variations from Surface Measurements, Table 1 Types of surface gravity measurements and their potential applications in studying time variable gravity field

	Absolute gravity meter (AG: FG 5)	Superconducting gravity meter (SG: GWR type)	Relative gravity meters (CG: Scintrex, CG3M, CG5M)
Precision	$\leq$1 μGal	0.001 μGal	1 μGal
Accuracy	1–3 μGal	0.1 μGal	10 μGal
Drift	0	0.01–0.05 μGal/day (variable)	3–10 μGal/day (variable)
Setup time and Operation	AG is a transportable instrument that operates in a controlled temperature environment. It can be set up in 1–3 h. 1–3 days measurements are desirable to get best results	Setup of SG requires a few days and some weeks to stabilize. It is housed at one place and needs calibration and drift monitoring by AG for better accuracy	CG is a field instrument that can be set up in a few minutes. A network of stations is used for measurement loops. Requires repeated measurements at a reference station every few hours
Dominant noise	Microseismic and man-made noise	Environmental effects	Environmental effects, microseismic noise, vehicle shaking, drift
Possible applications	Tidal, hydrological changes, pre-co-post seismic changes, glacial isostatic adjustment, tectonics, oil, gas, and geothermal reservoir monitoring, volcanic mass changes	Tidal, polar motion, hydrological changes, atmospheric and ocean loading's core processes, pre-co-post seismic changes, free oscillation of the glacial isostatic adjustments, volcanic mass changes	Tidal, hydrological changes, co-seismic changes for large quakes, volcanic mass changes

(with peak to peak value of $\sim \pm$ 150 μGal) according to the station location and usually are the largest contribution in any surface gravity measurements. They can be computed accurately and therefore can be removed from observed gravity changes to study non-tidal gravity changes (Wenzel, 1996). The precise tidal gravity changes are measured with SGs, AGs, and other surface gravimeters (Hinderer et al., 2007) on land and are recorded in satellite altimetry data over the oceanic region (Wahr, 2007). Tidal gravity changes are the elastic response of the Earth to the tidal forces and hence, modelling of tidal gravity changes provide information on the internal structure and properties of the Earth as its elastic transfer function varies according to the internal distribution of density, compressibility, and rigidity parameters. Tidal observations also provide independent evidence of general relativity theory of planetary movements (Warburton and Goodkind, 1977).

Sea level variation

Non-steric redistribution of ocean mass causes non-negligible changes in the Earth's gravity field. These are commonly recorded in satellite altimetry data. Movement of ocean waters also affects gravity changes inland; this may be quantified using results from satellite altimetry data (Fukuda and Sato, 1997). It is reported that gravity changes caused by sea level change deducted from satellite altimetry data are of the order of few μGal near coastal regions that can be observed in SG data.

Atmospheric pressure

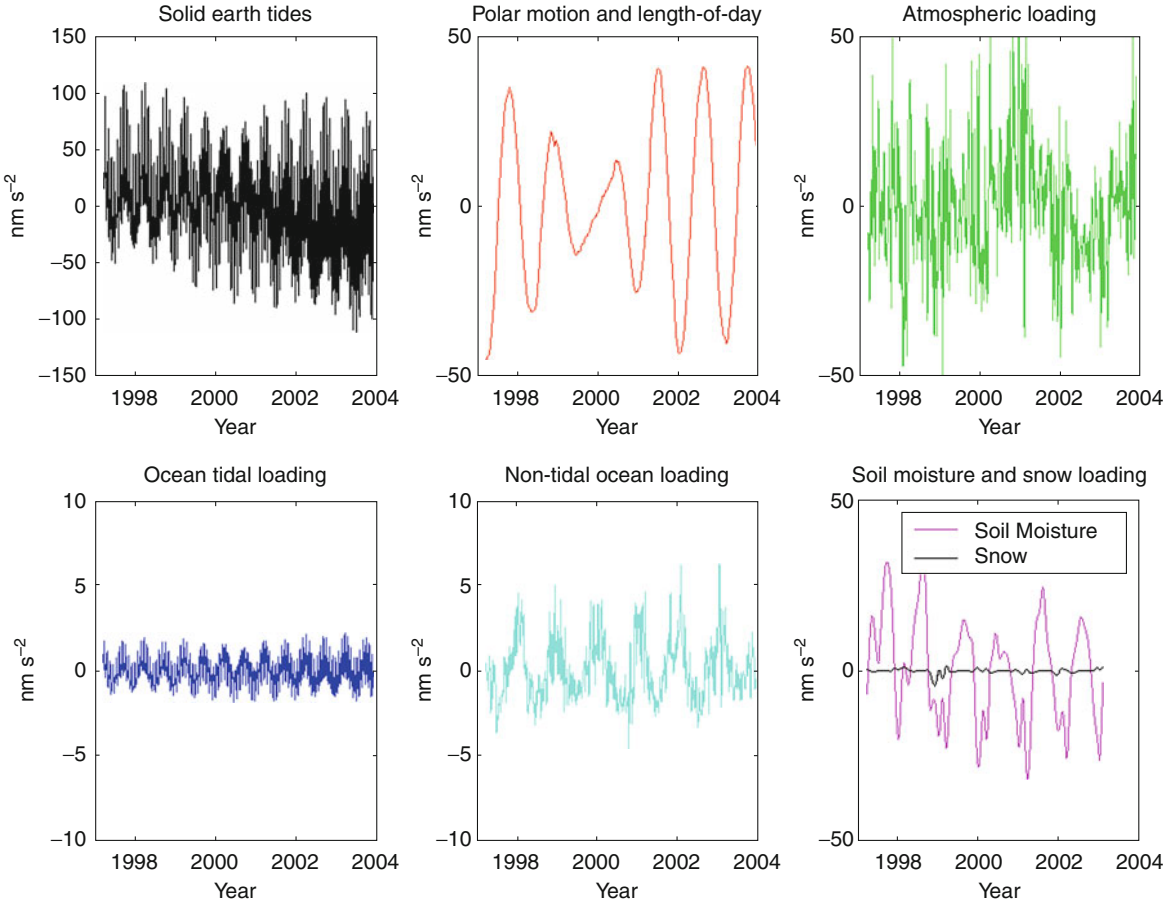
Atmospheric pressure changes alter surface gravity values because of the redistribution of mass in the atmosphere (attraction force) and the surface loading (elastic deformation) due to the change in the surface mass (Kroner and Jentzsch, 1999; Boy et al., 2002). The induced gravity changes are in the order of $\pm$ 5 μGal (Figure 2) and can reach up to 10–15 μGal for large pressure fronts (30–50 hPa). They are often estimated using a simple atmospheric admittance function between gravity and atmospheric pressure – that is 0.3 μGal/hPa (Crossley et al., 1995). These effects are generally considered as noise in gravity data and corrected using local barometric pressure data.

Polar motion

The direction of the Earth's axis of rotation is not steady in space but rather varies with time with two dominant periods (1 year and 435 days); this motion is generally known as polar motion or oscillation of the rotation pole. The 435-day component is referred to as Chandler wobble by astronomers. This signal can be detected by various techniques (spectral analysis, least squares adjustment) in surface gravity using continuous and stable SG observations (Hinderer et al., 2007). The related signals are of several μGal and change according to the station coordinates (Figure 2).

Free oscillation of the Earth

Large quakes are known to excite the free oscillations of the Earth (Hinderer et al., 2007; Arora et al., 2008), which lead to surface gravity changes in the seismic frequency



Gravity Field, Time Variations from Surface Measurements, Figure 2 Various contributions to time-varying gravity in Strasbourg, France (Courtesy J.-P. Boy; (1 μGal = 10 nm s^{-2}). The amplitude of the solid tides and the ocean tidal loading would be ten times larger than plotted amplitudes, which are only long-period tides.

band. These changes accompanying the eigenmodes occur at periods between a few hundreds of seconds and 54 min, which is the period of the free elastic oscillation of the Earth, and can be well investigated using SG records (Rosat et al., 2004). Besides, free oscillations unrelated to quakes but rather due to atmosphere–ocean–solid Earth interactions (termed "hum") were first reported from the SG measurements at Syowa, Antarctica (Nawa et al., 1998).

Glacial isostatic adjustment

On long timescales the Earth behaves like a viscoelastic material and deforms when submitted to a load. Major changes in climatic conditions of the Earth, from its birth to the present, caused large scale changes in strengthening and weakening of the ice sheets of the cold regions of the Earth's surface. The response of the Earth to the glaciations and deglaciations is known as Glacial Isostatic Adjustment (GIA). Land based gravity measurements have provided useful constraints to GIA, for instance, in the North American region (Lambert et al., 2001), Greenland (van Dam et al., 2000), or in Svalbard (Sato et al., 2006). One of the important properties of deep Earth that GIA can provide is information about the viscosity of the Earth materials.

Volcanic mass changes

The growth of the volcanic dome and fluxes of volcanic masses cause local gravity changes. Volcanic mass redistributions produce gravity change in the order of tens to hundreds of μGal and can be observed using any kind of surface measurements. Surface gravity measurements along with other geodetic measurements like GPS are prospective tools for the estimate of mass changes during the volcanic processes (Jousset et al., 2000).

Crustal deformation associated with quakes

Sudden displacements of faults during a quake cause redistribution of mass as well as change in the elevation, both leading to local gravity changes. An example of gravity changes before and after Haicheng quake, 1975, and the Tangshan

quake, 1976 is reported by Chen et al. (1979). Later on there have been several studies that document co-seismic gravity changes in surface measurements with SGs (Imanishi et al., 2004) and also in GRACE satellite observations (Han et al., 2006; Tiwari et al., 2008). Gravity records of silent and slow quakes, for example, from Cascadian subduction zone showing the concordance of gravity changes of about 10 μGal with episodic slip and seismic tremor activity (several times a year) is one of the interesting observations in temporal gravity changes (Hinderer et al., 2007).

Hydrology

Variations in land water storage (surface water, ground water, soil moisture, and snow) induce variations in the gravity field recorded on the surface (Crossley and Hinderer, 2005). The contributions of soil moisture and snow at Strasbourg, France are plotted in Figure 2. The amplitude of gravity changes associated with seasonal hydrological changes are larger than secular changes (Tiwari et al., 2006, Hinderer et al., 2007) and sometimes spread over a large area that can be mapped through satellite measurements (Tiwari et al., 2009) because large scale variations are difficult to be evaluated using ground-based observations due to the difficulty of adequate sampling. Hydrological fluxes at different time and spatial scales can now be mapped by monitoring temporal gravity changes, which holds promise to provide closing water cycle on local and regional scale (Cazenave et al., 2011, this book). Recent studies using the European regional network of SGs demonstrated that surface measurements of the seasonal component of continental hydrology can help in validating the satellite missions (Hinderer et al., 2007) and also suggest the complementary nature of satellite and ground-based observations.

Oil and gas reservoir monitoring

Time lapse gravity measurements are regularly used for monitoring subsurface density changes due to the fluid motions of the injected water and gas in the reservoir for enhancing recovery as well as for sequestration of CO_2 gases (Hare et al., 1999; Alnes et al., 2008). The amplitude of gravity changes is quite appreciable in the region where injections are large and reservoir depth is small. Such changes in gravity field occur over injection time period and are generally in the order of few tens of μGal to hundreds of μGal depending on the volume of injection and depth of injection. An estimate of other sources of the gravity changes (e.g., water table, atmospheric changes) in the area of measurements is required to model injected mass precisely.

Summary

Measurements and modelling of time-varying gravity changes allow understanding of various phenomena, as discussed in the earlier sections. Isolating the individual contributions from the aggregate effect in the temporal series of gravity changes has been a major challenge over a long time. Nevertheless, as number of observations and complementary information of temporal gravity changes are increasing, different components of time series of gravity changes can be better separated. Above mentioned examples and discussion suggest that repeat or continuous surface gravity measurements can be utilized for a wide variety of geophysical processes that deform the Earth's surface and move mass inside, on and above the Earth's surface causing the gravity field to vary with time. Surface measurements are also vital as they could provide a test, or validation, for satellite-derived temporal gravity changes (Crossley and Hinderer, 2005).

Bibliography

Alnes, H., Eiken, O., and Stenvold, T., 2008. Monitoring gas production and CO_2 injection at the Sleipner field using time-lapse gravimetry. *Geophysics*, **73**, WA155.

Arora, B. R., Kamal, Kumar, A., Rawat, G., Kumar, N., and Choubey, V. M., 2008. First observations of free oscillations of the from Indian superconducting gravimeter in Himalaya. *Current Science*, **95**, 1611–1617.

Boy, J.-P., Gegout, P., and Hinderer, J., 2002. Reduction of surface gravity data from global atmospheric pressure loading. *Geophysical Journal International*, **149**, 534–545.

Cazenave, et al., 2011. Gravity field of the, temporal variations from Satellite. In Gupta, H. K. (ed.), *Encyclopedia of Solid Earth Geophysics*, Springer 96.

Chen, Y. T., Gu, H. D., and Lu, Z. X., 1979. Variations of gravity before and after the Haicheng quake, 1975, and the Tangshan quake, 1976. *Physics of the Earth and Planetary Interiors*, **18**, 330–338.

Crossley, D., and Hinderer, J., 2005. Using SG arrays for hydrology in comparison with GRACE satellite data, with extension to seismic and volcanic hazards. *Korean Journal of Remote Sensing*, **21**, 31–49.

Crossley, D., Jensen, O. G., and Hinderer, J., 1995. Effective barometric admittance and gravity residuals. *Physics of the Earth and Planetary Interiors*, **90**, 221–241.

Crossley, D., Hinderer, J., Casula, G., Francis, O., Hsu, H.-T., Imanishi, Y., Jentzsch, G., Kaarianen, J., Merriam, J., Meurers, B., Neumeyer, J., Richter, B., Shibuya, K., Sato, T., and van Dam, T., 1999. Network of superconducting gravimeters benefits a number of disciplines. *EOS, Transactions of AGU*, **80**, 1, 121, 125–126.

Fukuda, Y., and Sato, T., 1997. Gravity effects of sea level variation at the superconducting Gravimeter sites, estimated from ERS-1 and Topex-Poseidon altimeter data. In Segawa, H., et al. (eds.), *Gravity, Geoid, and Marine Geodesy*. JAG Symposia. Springer, Berlin, Vol. 117, pp. 107–114.

Han, S. C., Shum, C. K., Bevis, M., and Kuo, C. Y., 2006. Crustal dilatation observed by GRACE after the 2004 Sumatra-Andaman quake. *Science*, **313**, 658–662, doi:10.1126/science.1128661.

Hare, J. L., Ferguson, J. F., Carlos, L. V., and Aikenand Brady, J. L., 1999. The 4-D microgravity method for water flood surveillance: a model study for the Prudhoe Bay reservoir, Alaska. *Geophysics*, **64**, 78–87.

Hinderer, J., and Crossley, D., 2000. Time variations in gravity and inferences on the Earth's structure and dynamics. *Surveys in Geophysics*, **21**, 1–45.

Hinderer, J., Andersen, O., Lemoine, F., Crossley, D., and Boy, J.-P., 2006. Seasonal changes in the European gravity field from GRACE: a comparison with superconducting gravimeters and hydrology model predictions. *Journal of Geodynamics*, **41**, 59–68.

Hinderer, J., Crossley, D., and Warburton, R., 2007. Superconducting gravimetry. In Schubert, G., (ed. in chief) *Treatise on Geophysics*. Amsterdam: Elsevier, Vol. 3, pp. 65–122.
Imanishi, Y., Sato, Y., Higashi, T., Sun, W., and Okubo S., 2004. A network of superconducting gravimeters detects submicrogal coseismic gravity changes. *Science*, **306**, 476–478.
Jousset, P., Dwipa, S., Beauducel, F., Duquesnoy, T., and Diament, M., 2000. Temporal gravity at Merapi during the 1993–1995 crisis: an insight into the dynamical behaviour of volcanoes. *Journal of Volcanology and Geothermal Research*, **100**, 289–320.
Kroner, C., and Jentzsch, G., 1999. Comparison of different barometric pressure reductions for gravity data and resulting consequences. *Physics of the Earth and Planetary Interiors*, **115**, 205–218.
Lambert, A., Courtier, N., Sasagawa, G. S., Klopping, F., Winester, D., James, T. S., and Liord, J. O., 2001. New constraints on Laurentide postglacial rebound from absolute gravity measurements. *Geophysical Research Letters*, **28**, 2109–2112.
Nawa, K., Suda, N., Fukao, Y., Sato, T., Aoyama, Y., and Shibuya, K., 1998. Incessant excitation of the Earth's free oscillations. *Earth, Planets and Space*, **50**, 3–8.
Niebauer, T. M., Sasegawa, G. S., Faller, J. E., Hilt, R., and Klopping, F., 1995. A new generation of absolute gravimeters. *Metrologia*, **32**, 159–180.
NRC, 1997. In Dickney, J. O., et al. (eds.), *Satellite gravity and the geosphere: National Research Council Report*. Washington: National Academy Press, p. 112.
Rosat, S., Hinderer, J., Crossley, D., and Boy, J.-P., 2004. Performance of superconducting gravimeters from long-period seismology to tides. *Journal of Geodynamics*, **38**, 461–476.
Sato, T., Okuno, J., Hinderer, J., MacMillan, D. S., Plag, H.-P., Francis, O., Falk, R., and Fukuda, Y., 2006. A geophysical interpretation of the secular displacement and gravity rates observed at Ny-Alesund, Svalbard in the Arctic – effects of the post-glacial rebound and present-day ice melting. *Geophysical Journal International*, **165**, 729–743.
Tiwari, V. M., Singh, B., Vyaghreswara Rao, M. B. S., and Mishra, D. C., 2006. Absolute gravity measurements in India and Antarctica. *Current Science*, **91**, 686–689.
Tiwari, V. M., Wahr, J. M., Gross, R., and Swenson, S., 2008. GRACE-observed gravity changes in areas of large quakes. *GRACE Science Meeting*, December 11–12, 2008, San Francisco, CA.
Tiwari, V. M., Wahr, J. M., and Swenson, S., 2009. Dwindling groundwater resources in northern Indian region, from satellite gravity observations. *Geophysical Research Letters*, **36**, L18401, doi:10.1029/2009GL039401.
van Dam, T. M., Larson, K., Wahr, J. M., Gross, S., and Francis, O., 2000. Using GPS and gravity to infer ice mass changes in Greenland, EOS. *Transactions of American Geophysical Union*, **81**(37), 421, 426–427.
Wahr, J. M., 2007. Variable gravity from satellite. In Schubert, G. (ed. in chief) *Treatise on Geophysics*. Amsterdam: Elsevier, Vol. 3, pp. 213–238.
Warburton, R. J., and Goodkind, J. M., 1977. Detailed gravity *tide* spectrum between 1 and 4 cycles per day. *Geophysical Journal of the Royal Astronomical Society*, **52**, 117130.
Wenzel, H. G., 1996. The nanogal software: tide data pro-cessing package *ETERNA* 3.3. *Bulletin d'Information des Marées Terrestres*, **135**, 9425–9439.

Cross-references

Gravity Field of the Earth
Gravity Field, Temporal Variations from Space Techniques
Gravity Measurements, Absolute
Gravity Method, Surface
Gravimeters

GRAVITY MEASUREMENTS, ABSOLUTE

Mark A. Zumberge
Institute of Geophysics and Planetary Physics, Scripps Institution of Oceanography, University of California, La Jolla, San Diego, CA, USA

Synonyms

Absolute acceleration due to gravity; Absolute *g*

Definition

Gravity. The mutually attractive force between any two masses.
Acceleration due to gravity, g. The value of the acceleration of a mass due to the force of gravity observed in a reference frame rotating with Earth.
Absolute. Measured with calibrated instruments traceable to metrologic standards. For the case of gravity, this requires length and time standards.
Absolute gravity meter or *absolute gravimeter.* An instrument used to measure gravity absolutely, traceable to time and length standards.

Introduction

The value of the acceleration due to gravity on Earth's surface, *g*, depends on position and time. A worldwide variability of roughly 0.5% from 9.8 m/s^2 is caused by Earth's rotation, variations in density, and variations in geometry caused both by tidal fluctuations and tectonic deformation. Studying how *g* varies can reveal important facts about Earth as a whole and about a long list of local phenomena ranging from volcanic activity to evolution of oil and gas reservoirs.

Often scientists are interested only in changes in gravity with either position (gravity maps) or with time (time-lapse gravity). Many of these studies can be completed with simple, relative gravity meters, which commonly consist of an electromechanical spring-mass sensor (Torge, 1990). Relative gravity measurements (i.e., the difference in *g* between two sites) are more common because the interesting variations in *g* are small and its absolute value is not important for most geophysical applications. Almost all relative gravity meters undergo drift, however, making it difficult to separate real temporal gravity changes from drift in the sensor. Drift can be calculated and accounted for in cases where it is feasible to make redundant observations at multiple sites and assume that gravity values at reference sites are constant. In a number of important cases, however, such drift compensating efforts are inadequate and absolute measurements are needed – they are inherently free from drift because of their reliance on length and time standards.

A note on units: Practitioners in the field of gravity measurement use the unit *Gal* (after Galileo), which is defined as 1 cm/s^2. The nominal value of *g* is around 980 Gal. More common usage is the mGal (= 10^{-3} Gal)

and the μGal (= 10^{-6} Gal or about $10^{-9}g$). A precision of a few μGal is available from the best relative gravity meters (following a significant drift compensation effort within the measurement process) and 1 or 2 μGal with the best absolute gravity meters. Absolute determinations are characterized by both precision (the smallest observable signal or change) and accuracy (level of uncertainty or the discrepancy between the measured value and the unknown true value).

History

Late in the seventeenth century, it was discovered that a pendulum would swing with a slightly different period if it were moved to a different latitude. This effect, caused by gravity's dependence on latitude (primarily because of Earth rotation), was used to make some of the earliest estimates of the deviation of Earth's figure from a sphere (Lenzen and Multhauf, 1965). Two kinds of pendulum gravity meters, the reversible pendulum and the invariable pendulum, provide a good example of the difference between absolute and relative sensors. A reversible pendulum, used extensively by Henry Kater in the nineteenth century, exploits the curious fact that two pivot points on an asymmetric pendulum, which are found to yield equal oscillation periods are separated by the length of a simple pendulum (a point mass suspended from a massless rod) having that same period. This allows an absolute calculation of gravity based on the measurement of the period and the distance between the two oscillation points, which have a common period, and is therefore an absolute measurement. Invariable pendulums, on the other hand, were simply swung around a single point and changes in gravity were observed as changes in the period. Pendulum measurements were made extensively by geodesists to refine the figure of the Earth and for mineral exploration up until the early twentieth century when electronic methods became adequately advanced to measure accurately the time needed for a mass in a vacuum chamber to fall a known distance. Cook (1965) and Faller (1963) pioneered some of the early free-fall measurement techniques. Toward the end of the twentieth century, the advent of the laser allowed for great improvements in absolute gravity measurement (Hammond and Faller, 1967; Alasia et al., 1983; Zumberge et al., 1982; and Niebauer et al., 1995). Now, in the beginning of the twenty-first century, atomic methods are becoming viable. The most commonly used method today is to determine the acceleration of a mass freely falling in a vacuum by means of laser interferometry.

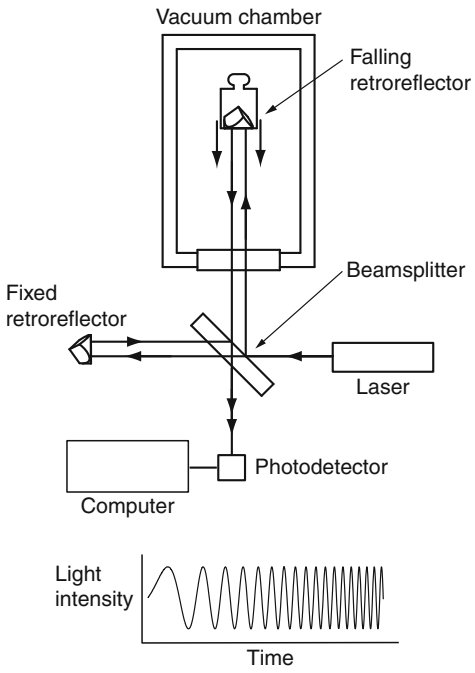
Principles of modern measurements

Macroscopic falling test mass

Figure 1 is a schematic diagram of the mainstay of modern absolute gravity measurements. A mechanical system inside a high vacuum chamber effects the release of a test mass. A retro-reflecting prism mounted to the test mass returns laser light entering and exiting the chamber through a window in the bottom. A beam splitter (or half-silvered mirror) combines this reflected light with that similarly reflected from a fixed retro-reflector. The two reflected beams interfere with one another, evermore rapidly producing successive light and dark fringes as the falling mass accelerates. Two cycles of light-to-dark interference fringes are completed for each advance in the fall of a distance equaling the laser's wavelength. A photodetector senses the passage of interference fringes and generates an electrical signal from which the local value of gravity (the rate at which the falling mass accelerates) can be calculated through appropriate timing analysis performed by a computer. Typically, the mass falls a distance of 10–20 cm and a thousand or more drops over a period of a day provide a statistical uncertainty of a few μGal. Because the measurement is based on calibrated length and time standards – the wavelength of the laser light and the frequency of the timing oscillator – the measurement is absolute.

Microscopic falling test masses

A new and promising method for the absolute determination of gravity makes use of cold atom interferometry (Kasevich, 2002). Toward the end of the twentieth century, physicists learned to use lasers to trap and manipulate clouds of atoms. Atoms can now be confined in an evacuated magneto-optical trap, cooled to very low



Gravity Measurements, Absolute, Figure 1 A schematic diagram of a free-fall absolute gravity meter. A mechanical system (not shown) inside the vacuum chamber releases and catches the falling object. An interferometer determines the position of the test mass as it falls; the interference fringe signal is an increasing frequency sine wave.

temperatures with laser cooling, and then launched vertically (again, with lasers) to create an "atomic fountain." Exploiting the fact that all matter can be described as waves, atomic physicists have learned to interfere atomic wave functions with one another and, in so doing, perform a number of highly precise measurements using them. In particular, because the phase of an atom's wave function depends on gravity, interfering atoms that have traversed separate paths through a gravitational field can yield an absolute measure of gravity. The first experiments performed by Kasevitch and Chu (1992) and followed by others (Peters et al., 2001; Lamporesi et al., 2008) required a laboratory full of complex optical, vacuum, and electronic equipment. Recently, however, the technology has advanced to a state where portability may become feasible (Merlet et al., 2010). Commercial versions of cold atom interferometric absolute gravity meters are currently under development (e.g., AOSense, Sunnyvale, California, USA).

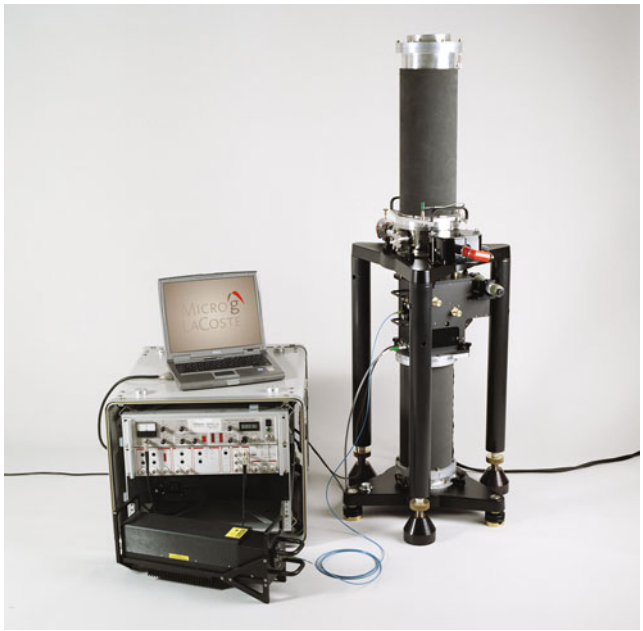
Instrumentation

The mechanical details of conventional free-fall apparatuses vary. In some, the test mass's fall is initiated by a small elevator inside the vacuum chamber, which also surrounds the mass as it falls and acts as a shield against minute non-gravitational forces. In others, the mechanical system resides at the bottom of the vacuum chamber and launches the test mass upward, providing a mass in free flight, the acceleration of which can be measured during both its rise and fall. Another important variation is the use of a spring suspension system to isolate the reference reflector from vibrations.

Currently, the most widely used absolute gravity meter is the model FG5, manufactured by Micro-g LaCoste (Lafayette, Colorado, USA) and shown in Figure 2. With laser interferometry, it tracks the descent of a 300 g mass during a 15-cm drop. The portable system weighs 320 kg, can be assembled in an hour or so, requires 500 watts of electrical power, and can produce an absolute gravity value accurate to 2 μGal after an hour of data collection. Significant care has been taken by its developers to ensure that systematic errors – those whose signs do not vary and cannot be reduced by averaging – are less than the quoted accuracy. Such errors include forces on the falling mass other than gravity (e.g., residual air-drag and magnetic or electrostatic forces) and imperfections in the optical and timing systems. Periodically, many absolute gravity meters make measurements simultaneously in one of the world's standards lab for an intercomparison to test the assessments of the instruments' accuracies (Sasagawa et al., 1995).

Applications to geophysics

Near Earth's surface, the rate of change of gravity with altitude is about 3 μGal per cm. Therefore, a detection of a gravity change of a few μGal yields information on cm-level height changes. Observing gravity changes



Gravity Measurements, Absolute, Figure 2 The micro-g LaCoste FG5 is a commonly used portable absolute gravity meter.

together with height information obtained by other means can provide important information for modeling the causes of surface deformation. An increase in density of 0.1 g/cm^3 occurring throughout a 25-cm thick layer below the observer (such as might be caused by water saturation) increases gravity by 1 μGal. These changes can occur very slowly, which makes it desirable to use absolute measurements to detect such changes. Examples of long time series of absolute gravity used to study geophysical phenomena include observations of post-glacial rebound (Lambert et al., 2001), water flooding in a natural gas reservoir (Hare et al., 1999), vertical land movement near tide-gauges (Teferle et al., 2006), and observations of coseismic offsets from an earthquake (Tanaka et al., 2001).

A very important geophysical application of absolute gravity measurement is the establishment of calibration lines for relative gravity meters (Sasagawa et al., 1989). Absolute g measurements are also necessary in certain metrology applications, including piston-cylinder pressure calibrations and in defining the standard kilogram.

Summary

Absolute gravity measurements are much less common than those made with relative gravity meters. However, because of their freedom from drift, made possible through the use of modern atomic length and time standards, absolute g meters are extremely important for monitoring long-term changes associated with crustal deformation and density change and for calibrating

relative meters. New advances in atomic physics may result in an absolute gravity meter, which is similar in size to typical relative gravity meters, advancing the use of absolute gravity measurements in geophysics.

Bibliography

Alasia, F., Cannizzo, L., Cerutti, G., and Marson, I., 1982. Absolute gravity acceleration measurements: experiences with a transportable gravimeter. *Metrologia*, **18**, 221.

Cook, A. H., 1965. The absolute determination of the acceleration due to gravity. *Metrologia*, **1**, 84.

Faller, J. E., 1963. *An Absolute Interferometric Determination of the Acceleration of Gravity*. PhD thesis, Princeton, NJ, Princeton University.

Hammond, J. A., and Faller, J. E., 1967. Laser-interferometer system for the determination of the acceleration of gravity. *IEEE Journal of Quantum Electronics*, **QE-3**, 597.

Hare, J. L., Ferguson, J. F., Aiken, C. L. V., and Brady, J. L., 1999. The 4-D microgravity method for waterflood surveillance: A model study for the Prudhoe Bay reservoir, Alaska. *Geophysics*, **64**, 78.

Kasevich, M. A., 2002. Coherence with atoms. *Science*, **298**, 1363.

Kasevitch, M., and Chu, S., 1992. Measurement of the gravitational acceleration of an atom with a light-pulse atom interferometer. *Applied Physics B*, **54**, 321.

Lambert, A., Courtier, N., Sasagawa, G. S., Klopping, F., Winester, D., James, T. S., and Liard, J. O., 2001. New constraints on Laurentide postglacial rebound from absolute gravity measurements. *Geophysical Research Letters*, **28**, 2109.

Lamporesi, G., Bertoldi, A., Cacciapuoti, L., Prevedelli, M., and Tino, G. M., 2008. Determination of the Newtonian gravitational constant using atom interferometry. *Physical Review Letters*, **100**, 050801.

Lenzen, V. F., and Multhauf, R. P., 1965. Development of gravity pendulums in the 19th century. *US National Museum Bulletin*, **240**, 301.

Merlet, S., Bodart, Q., Malossi, N., Landragin, A., Pereira Dos Santos, F., Gitlein, O., and Timmen, L., 2010. Comparison between two mobile absolute gravimeters: optical versus atomic interferometers. *Metrologia*, **47**, L9–L11.

Niebauer, T., Sasagawa, G., Faller, J., Hilt, R., and Klopping, F., 1995. A new generation of absolute gravimeters. *Metrologia*, **32**, 159.

Peters, A., Chung, K. Y., and Chu, S., 2001. High-precision gravity measurements using atom interferometry. *Metrologia*, **38**, 25.

Sasagawa, G., Zumberge, M., Stevenson, J., Lautzenhiser, T., Wirtz, J., and Ander, M., 1989. The 1987 southeastern Alaska gravity calibration range: absolute and relative gravity measurements. *Journal of Geophysical Research*, **94**, 7661.

Sasagawa, G., Klopping, F., and van Dam, T., 1995. Analysis of data from the 1994 international comparison of absolute gravimeters with a single computational protocol. *Metrologia*, **32**, 192.

Tanaka, Y., Okubo, S., Machida, M., Kimura, I., and Kosuge, T., 2001. First detection of absolute gravity change caused by earthquake. *Geophysical Research Letters*, **28**, 2979.

Teferle, F. N., Bingley, R. M., Williams, S. D. P., Baker, T. F., and Dodson, A. H., 2006. Using continuous GPS and absolute gravity to separate vertical land movements and changes in sea-level at tide-gauges in the UK. *Philosophical Transactions of the Royal Society*, **364**, 917.

Torge, W., 1989. *Gravimetry*. Berlin: de Gruyter.

Zumberge, M. A., Rinker, R. L., and Faller, J. E., 1982. A portable apparatus for absolute measurements of the Earth's gravity. *Metrologia*, **18**, 145.

Cross-references

Gravity Field of the Earth
Gravity Field, Time Variations from Surface Measurements

GRAVITY METHOD, AIRBORNE

Uwe Meyer
Sub-Department Geophysical Reconnaissance – Resources and Near Surface Processes, Federal Agency for Geosciences and Mineral Resources, Hannover, Germany

Synonyms

Aerogravity; Airborne gravimetry

Definition

Airborne. Anything carried by a fixed wing aircraft, a helicopter, an airship, or an unmanned aircraft.
Aero. Anything related to the atmosphere above ground.
Gravity. In this context nonrelativistic mass attraction as defined by Newton.
Gravimetry. In this context the measurement of the gravity field variation by mass attraction.

Introduction

The law of gravity is one the most fundamental in natural sciences. Instruments for gravity measurements have been developed first for use on land surface, later for use on ships. The instruments used for maritime surveying needed a special platform to compensate the movements of the vessel that add additional components to the main value of the measurement: the acceleration of a well-defined mass due the gravity field of the Earth. The same accounts for airborne applications of gravity meters: The aircraft's variations in roll, pitch, and yaw angles and height changes must be either compensated for or measured for later reductions and corrections. Therefore, most instruments used in the early stages of airborne gravity meter development were basically upgraded marine gravity meters (LaCoste, 1967). Due to the problem of the fast aircraft movements and therefore the even larger accelerations implied, only after the practical implementation of a global satellite-based positioning system (as GPS) non-inertial navigation became precise enough to enable satisfactory corrections and reductions (Bruton et al., 1999). At present, three different categories of airborne gravity instruments evolved: scalar gravity meters, vector gravity meters, and gradient or full tensor meters. All instruments are used to measure the regional variation of the Earth's gravity field over often inaccessible areas in a quick fashion.

Principle idea of airborne gravity

In the still most common airborne gravity applications, scalar gravity meters are used. Such meters usually

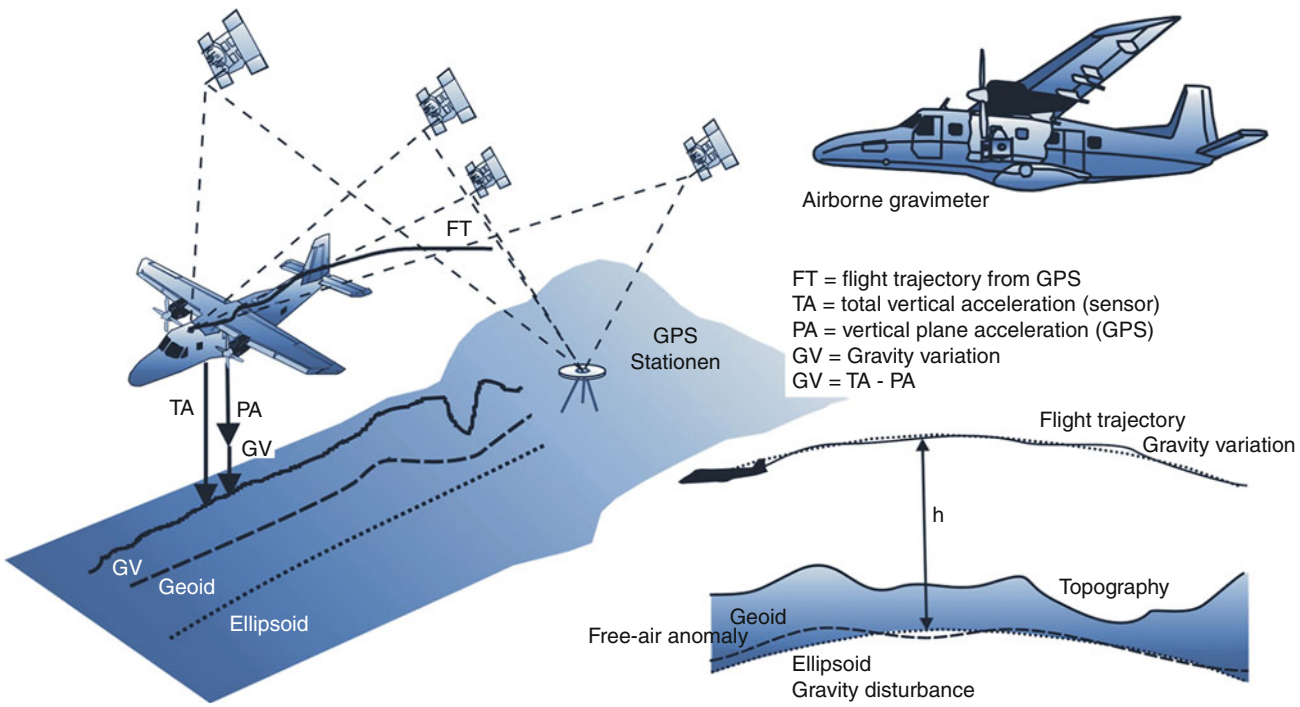
measure only the relative vertical component of the Earth's gravity field using a mass coupled to a spring, a principle that was adopted from gravity land-based surveying (Figure 1). In airborne gravity, much larger accelerations than the desired signal are introduced by attitude and height changes of the aircraft, carrying the gravity meter. Beyond these inertial accelerations, vibrations occur that are usually damped by suitable mechanical or electromechanical means. The residual effect of the vibrations in the aircraft's cabin is usually dealt with as a noise signal. Therefore, we focus on the gravitational and inertial accelerations only. The instrument and the navigation equipment are placed in a rigid aircraft that defines the body frame. Within this body frame, we primarily measure the accelerations in the axes of the aircraft: forward, down, and sideward in a right-hand system. The gravity sensor is fixed on a platform that compensates some of the aircrafts forward and sideward accelerations that are coupled to attitude variations but not the vertical acceleration components. The main aim of the platform is to keep the gravity sensor in horizontal level in order to keep the measurement range and sensitivity of the gravity meter high. Additional navigation equipment as global positioning system antennas and inertial measurement units monitor the movements of the aircraft's body frame in the geodetic frame. Consequently, lever arms between the navigation measurement units and the gravity sensor must be corrected for in the body frame and the resulting accelerations and force must be transferred to the geodetic reference frame for later gravity mapping.

Attitude compensation platforms

Most platforms designed to carry a gravity sensor on a moving vehicle are two axes gimbal constructions to compensate coupled roll, pitch, and yaw variations (LaCoste et al., 1967). The attitude information is commonly acquired by a gyro for short wavelengths and an accelerometer for wavelengths up to some minutes for each directional component. Some platforms compensate heading variations as well. Such platforms are Schuler-tuned. The platforms are designed to compensate attitude changes almost instantaneously in the dynamic range of the gravity sensor. Therefore, special filters are applied that define the platform period, which is usually in the order of some minutes.

Scalar gravity meters

Most meters in use for airborne gravity surveying are based on a test mass coupled to a spring or on torsion balance. Some new developments utilize superconductive gravity sensing systems. Such instruments measure the relative changes of the main vertical component of the Earth's gravity field. Beyond the residual platform errors, an independent non-inertial vertical acceleration determination is needed to reduce for the vertical accelerations of the body frame. This is possible only since global satellite-based positioning system navigation data is available. Scalar systems generally have a very low drift rate and therefore can be used for long wavelength gravity mapping as required for geodesy. The shortest detectable



Gravity Method, Airborne, Figure 1 Principle of airborne gravity.

wavelengths are dependent on flight velocity, the dynamic range of the gravity sensor, and the platform period. Generally, some few mGal (1 mGal = 10^{-5} m/s^2) are achieved in accuracy related to some few kilometers wavelength resolution at an average speed of about 160 km/h above ground.

Vector gravity meters

Vector gravity meters are in most cases adapted inertial measurement units that measure the three vertical components of the Earth's gravity field in the body frame of the aircraft (Jekeli, 2001). Analogue to most attitude-compensating platform designs, the signal is gathered from a combination of directional gyro and accelerometer measurements. In this configuration, global satellite-based positioning system navigation data must be applied for all directional acceleration reductions. The drift rates of such instruments are usually higher than those of scalar meters. At present, only very few systems are available, that are not only able to measure the significant vertical component but also the horizontal components in sufficient accuracy and resolution. Most vector gravity meters are not installed on an attitude-compensating platform but are placed on dampers only. The resolution and accuracy for the vertical component of available vector systems is better than those of scalar spring tension or torsion balance meters.

Gravity gradient meters

In order to enhance the short wavelengths resolution and the measurement accuracy significantly, gravity gradient meters were developed (Heiland, 1940). The main idea is to eliminate the impact of the aircrafts attitudes and accelerations in the measurements by measuring the directional gradients directly. To achieve this, diagonally arranged pairs of highly accurate accelerometers are placed on a slowly rotating disk, one for each direction. Since the two accelerometers measure in opposite directions, the largest effects of the body frame accelerations are ruled out by the acceleration difference between the two accelerometers and mainly the directional gravity gradient remains. Such gradients are not stable enough for large-scale integration and therefore are only suitable for short wavelength detections in local to small-scale regional areas. This sensing system needs an exceptionally accurate attitude compensation platform. If all directional gradient components are measured, this system is called a full tensor gradient gravity meter. Available resolutions are in the order of 0.5–1 Eotvoes (1 Eotvoes (E) = 0.1 mGal/km = 10^{-9} s^{-2}).

Processing, reductions, and corrections

The original data set is the sum of all measured accelerations in one or more components within the body frame of the aircraft. In order to get the gravity variation on flight level, the directional components of the body frame accelerations in a chosen geodetic frame have to be reduced for. As a next step, an Eotvoes correction has to be applied to compensate for the gravitational impact due to varying centrifugal forces (Harlan, 1968). Next, a platform error correction has to be applied, usually a Harrison correction for scalar meters.

The gravity disturbance (gravity variation on ellipsoidal level) is determined by the normal gravity reduction and the free-air gravity reduction between observation level and ellipsoid.

In order to achieve the gravity anomaly, the same is done with reference to the geoid instead of the ellipsoid.

Applications and aircraft

Airborne gravity can either be used to determine the regional geoid more precisely or for geophysical reconnaissance and prospection. For geodetic purposes rather high altitude straight and level fixed wing aircraft flights utilizing scalar gravity meters are most common (Forsberg et al., 1996). In geophysics, two aspects occur: large to mid-scale reconnaissance surveys to get an overview on tectonic or geological settings and small-scale prospection in search for mineral deposits or similar tasks (Veryaskin and McRae, 2008). For reconnaissance, fixed wing aircraft with long-range capabilities are preferred. In prospection, high-resolution vector or gradient gravity meters are used on small fixed wing aircraft, helicopters, or airships. First attempts to use unmanned airborne vehicles (UAV) are under way. New instrumental developments try to utilize modern superconductive meters, absolute gravity meters, or devices based on cold atom interferometry.

Summary

Airborne gravity has become a practical and accurate method to measure the variations of the Earth's gravitational field over regional areas. In geodesy, fixed wing aircraft equipped with scalar gravity meters are preferred. In geophysics, the same configuration is often used for reconnaissance surveys. For aerogeophysical prospection, low- and slow-flying aircraft are favored combined with vector or gradient gravity meters.

Bibliography

Bruton, A. M., Glennie, C. L., and Schwarz, K. P., 1999. Differentiation for high precession GPS-velocity and acceleration determination. *GPS Solutions*, **2**(4), 7–21.

Forsberg, R., Hehl, K., Bastos, L., Gidskehaug, A., and Meyer, U., 1996. Development of a geoid mapping system for coastal oceanography (AGMASCO), *Gravity, Geoid and Marine Geodesy*, IAG Symposium Series, Springer, Berlin, Vol. 117, pp. 163–170.

Harlan, R. B., 1968. Eotvoes corrections for airborne gravimetry. *Journal of Geophysical Research*, **73**(14), 4675–4679.

Heiland, C. A., 1940. *Geophysical Exploration*. New York: Prentice-Hall, Inc.

Jekeli, C. H., 2001. *Inertial Navigation Systems with Geodetic Applications*. Berlin: Walter de Gruyter.

LaCoste, L. J. B., 1967. Measurement of gravity at sea and in the air. *Reviews of Geophysics*, **5**(4), 477–526.

LaCoste, L. J. B., Clarkson, N., and Hamilton, G., 1967. LaCoste & Romberg stabilized platform shipboard gravity meter. *Geophysics*, **32**, 99–109.
Veryaskin, A., and McRae, W., 2008. On combined gravity gradient components modelling for applied geophysics. *Journal of Geophysical Engineering*, **5**, 348.

Cross-references

Gravity Measurements, Absolute
Gravity Method, Principles
Gravity, Gradiometry

GRAVITY METHOD, PRINCIPLES

Hans-Jürgen Götze
Abteilung Geophysik, Institut für Geowissenschaften,
Christian-Albrechts-Universität zu Kiel, Kiel, Germany

Introduction

By gravimetry (Latin "gravis") methods are identified, which can be used to measure the gravity field of the Earth. The determination of this potential field is of greater importance for geodesy, geophysics, and geotechnics, *Gravity Method, Surface*. In particular, in geophysics and geodynamics, gravity is used to investigate the Earth's interior and its dynamic processes. Furthermore, precise balances and other technical equipment need gravity data for their calibration.

The main objective of gravity/gravimetry is the determination of the Earth's gravity field and of other celestial bodies (e.g., the Earth's moon). The gravity field and its gravity gradients are measured as a function of position and time on, below, and above the Earth's surface (e.g., by gravity meters on ships, in boreholes, or on board aircrafts and, nowadays, satellites: CHAMP, GRACE, and GOCE Mission), *Gravity Method, Satellite*.

In the seventeenth and eighteenth centuries, scientific interest and first measurements have been caused by the understanding and development of the physics of deformable and rigid bodies. At that time, determinations of position (distances) and time became accurate enough to enable gravity measurements. Later, after scientists discovered the position dependency of gravity, it has been used to determine the figure of the Earth – then until today. Developments of the gravity method were always correlated with the interaction of technical possibilities and scientific objectives in physics (geophysics) and land surveying (later geodesy) because geodetic surveying and geophysical exploration became a strong influence on growing societies (e.g., Torge, 1989). Today the gravity field and its variations – measured by satellites with very high and unexpected accuracy – are closely related with the distribution and transport of masses in the Earth's system: the ocean circulation, changes of ground water level and solid Earth moisture, melting of continental ice sheets, river run off, changes of sea level, and mass flows in the Earth lithosphere and mantle, all of which cause transport and redistribution of masses. Using innovative and extremely precise gravity sensor systems dedicated to observe the gravity field, the implication of these processes can be established. Even in the investigation of other celestial bodies' gravity gains in importance: e.g., in the 1970s the "Lunar surface gravity meter" was deployed at the Apollo 17 ALSEP site (Talwani, 2003). The state of lunar and planetary gravity field investigations are described and discussed (e.g., by Janle and Meissner, 1986; Konopliv et al., 2001).

In the following chapters, all values of parameters of the Earth's gravity field are related to the World Geodetic System 1984 (WGS 84), Geodesy, Figure of the Earth.

Potential function

It is accepted that the force **F** of a mass point m, which is influenced by a timely constant gravity acceleration **g** can be expressed by

$$\mathbf{F} = m\mathbf{g}$$

To move this mass along a path C with path elements d**r** from a space point P_1 to a second point P_2, an energy U_{12} has to be applied:

$$U_{12} = -\int_{P_1}^{P_2} \mathbf{F} \cdot d\mathbf{r} = -m \int_{P_1}^{P_2} \mathbf{g} \cdot d\mathbf{r} \tag{1}$$

There is no loss of energy and we call U_{12} independent from its path and only dependent from start and end point P_1 respective to P_2. Equation 1 shows that only the difference of potential energy at points P_1 and P_2 can be calculated. The force **F** can be derived from the potential energy as:

$$\mathbf{F} = -\text{grad } U = -\nabla \cdot U \tag{2}$$

F is a potential field. Equation 2 holds even if we add a constant U_0. This constant will be selected in a way that potential energy disappears in the infinite. In this case, U(P) defines the energy in the field of gravity acceleration which has to be applied to m to move it from point P to the infinite:

$$U(P) = -\int_{P}^{\infty} \mathbf{F} \cdot d\mathbf{r}.$$

In geophysics and geodesy, the potential energy U is replaced by the potential (potential function):

$$V = -U/m. \tag{3}$$

This is the negative potential energy per mass unit, and with (Equation 2) one has:

$$\mathbf{g} = \nabla V.$$

Therefore, gravity field can be described in entire space by a single function V. The potential function fulfills the POISSON differential equation:

$$\Delta V = -4\pi f \rho \tag{4}$$

with:

$f = (6.67428 \pm 0.00067) \cdot 10^{-11}\ m^3/kg\ s^2$ (gravitational constant)

ρ = mass density.

The POISSON equation forms the base of potential theory and is an elliptic differential equation. If space is mass free ($\rho = 0$) the LAPLACE equation holds in Cartesian coordinates:

$$\Delta V = \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = 0 \tag{5}$$

It became the most universal differential equation in the physical sciences because of the wide range of phenomena that is described by the equation. There exists a multitude of solutions of this differential equation which are not always potential functions. For both differential equations hold the following *definition*:

> A space function and its second derivatives which are continuous and bounded in a region where defined are called harmonic functions if they fulfill the LAPLACE and – in the interior of a mass filled space – the POISSON differential equation.

Newton's law and Newtonian potential

In 1687 Sir Isaac Newton published his famous treatise "Philosophiae Naturalis Principia Mathemetica" which contains the Newton's law of gravitational attraction:

> The magnitude of a gravitational force between two masses is proportional to each mass and inversely proportional to the square of their distance.

In a Cartesian coordinate system, a mass m_0 is centered at point $P_0\,(x', y', z')$ and at point P (x, y, z) a mass m. Then a force, F(P), acts on mass m_0 (Figure 1):

$$\mathbf{F}(P) = -f\frac{mm_0}{R^2}\mathbf{e}_R \tag{6}$$

with

$\mathbf{R} = P_0\,P$ (vector which points from P_0 to P)

$\mathbf{e}_R = \mathbf{R}/R$ (unit vector) of $\mathbf{R}$

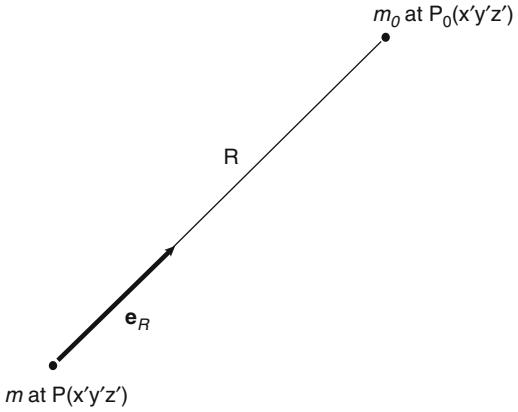
$R = [(x - x')^2 + (y - y')^2 + (z - z')^2\,]$ (norm of vector $\mathbf{R}$)

The gravity acceleration $\mathbf{g}$ which is caused by mass m is:

$$\mathbf{g}(P) = \frac{\mathbf{F}(P)}{m_0} = f\frac{m}{R^2}\mathbf{e}_R \tag{7}$$

and in Cartesian coordinates one can write:

$$\mathbf{g}(P) = fm/R^3[(x - \acute{x})\mathbf{i} + (y - \acute{y})\mathbf{j} + (z - \acute{z})\mathbf{k}] \tag{8}$$



Gravity Method, Principles, Figure 1 Masses m_0 and m experience a gravitational force which follows Newton's law: it is proportional to m_0 and m and $1/r^2$. By convention the unit vector $\mathbf{e}_R$ is directed from the gravitational source point (P) to the "observation" point P_0 which is located at the mass m_0.

$\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ are the components of unit vector in the coordinate system.

At the Earth's surface, gravity acceleration is approximately $g = 9.81\ m/s^2$. The potential V is a scalar and an additive quantity. Therefore, the gravity potential of n point masses can be written as:

$$V(P) = f\sum_{i=1}^{n}\frac{m_i}{R_i},\ \text{with } R_i \text{ (spaces of the n points P)}$$

At the point masses, V(P) is not defined due to $R_i = 0$. However, except for these singular points the LAPLACE equation holds.

For a mass continuum we obtain from Equation 6 the Earth's gravitation by Newton's law of gravitation:

$$\mathbf{b}(\mathbf{r}) = f\iiint_{\text{Earth}}\frac{\mathbf{r}' - \mathbf{r}}{|\mathbf{r}' - \mathbf{r}|^3}\,dm$$

where r' and $\mathbf{r}'$ are the geocentric position vectors of a mass element dm and the attracted point P (unit mass = 1) and $|\mathbf{r}' - \mathbf{r}| = R$. The mass element dm can be expressed by the volume density $\rho = \rho(\mathbf{r}')$ and the volume element $dm = \rho\,dv$.

$$\mathbf{b} = \text{grad } V$$

For the Earth's gravitational potential with v = volume of the Earth we have:

$$V(\mathbf{r}) = f\iiint_{\text{Earth}}\frac{\rho(\mathbf{r}')}{|\mathbf{r}' - \mathbf{r}|}\,dv \tag{9}$$

Besides the attraction of the Earth at points P generally all cosmologic masses contribute to this attraction.

However, in most cases only the influence of sun and moon is considered. The entire gravity attraction can be obtained by:

$$\nabla \cdot \mathrm{V(P)} = \mathbf{a}_\mathrm{E}(\mathrm{P}) + \mathbf{a}_\mathrm{C}(\mathrm{P}) \tag{10}$$

$\mathbf{a}_\mathrm{E}$ – attraction of Earth
$\mathbf{a}_\mathrm{C}$ – attraction of cosmological masses

$\mathbf{a}_\mathrm{E}$ causes time-dependent deformations of the Earth by tidal effects and each mass point P is dislocated by a very small value $\mathrm{R_D}$. Each mass point is influenced by the rotation of the Earth, and its angular velocity is given by $\omega_E = 7.292115 \cdot 10^{-5}$ rad/s, Gravity field of the Earth.

In an Earth-centered coordinate system, the position vector $\mathbf{R}_\mathrm{P}$, which points to point P is given by (refer also to Figure 2):

$$\mathbf{R}_\mathrm{P} = \mathbf{R}_1 + \mathbf{R}_\mathrm{D}$$

Then, the gravity vector $\mathbf{g}(\mathrm{P})$ can be calculated by: $\mathbf{g}(\mathrm{P}) = \nabla \cdot \mathrm{V(P)} - \ddot{\mathbf{R}}_p$

It holds:

$$\dot{\mathbf{R}}_\mathrm{P} = \dot{\mathbf{R}}_\mathrm{D} + \dot{\mathbf{R}}_1 = \dot{\mathbf{R}}_\mathrm{D} + (\omega_E \times \mathbf{R}_1)$$

$$\ddot{\mathbf{R}}_\mathrm{P} = \ddot{\mathbf{R}}_\mathrm{D} + (\omega_E \times (\omega_E \times \mathbf{R}_1))$$

disregarding the small changes of rotational velocity $\dot{\omega}_E$. The gravity vector can be written as:

$$\mathbf{g}(\mathrm{P}) = \mathbf{a}_\mathrm{E}(\mathrm{P}) + \mathbf{a}_\mathrm{c}(\mathrm{P}) - \ddot{\mathbf{R}}_\mathrm{D} + (\omega_E \times (\omega_E \times \mathbf{R}_1)) \tag{11}$$

The first term represents the attraction of Earth, the latter the centrifugal acceleration which is caused by Earth rotation.

The middle terms $\mathrm{a_c(P)} - \ddot{\mathbf{R}}_\mathrm{D}$ display the time-dependent part of Earth-tides. For precise gravity field determinations it must be observed and eliminated from the observations. The tidal correction which is caused by the sun and the moon never exceeds 3×10^{-5} m/s^2.

For the centrifugal acceleration $\mathbf{a}_z$, we can write:

$$\begin{aligned}\mathbf{a}_z &= (\omega_E \times \mathbf{R}_1) \times \omega_E \\ &= \omega_E{}^2 \mathbf{R}_1 - (\omega_E{}^2 \cdot \mathbf{R}_1)\omega_E = \omega_E{}^2 \cdot \mathbf{R}_\perp\end{aligned} \tag{12}$$

with

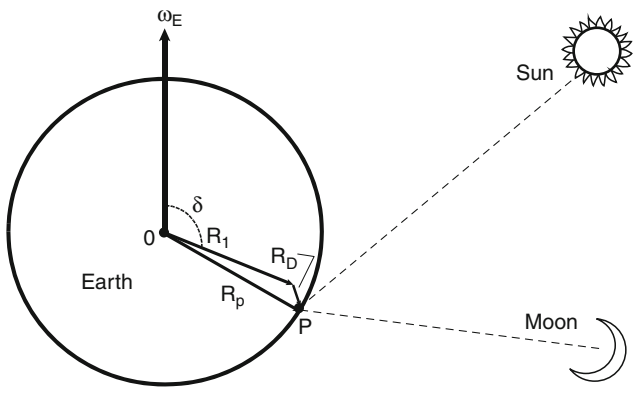
$\mathbf{R}_\perp$ = to the Earth's rotation axis perpendicularly oriented component of $\mathbf{R}_1$.

Because $\mathbf{R}_\perp = 1/2 \ \nabla \ \mathrm{R_\perp}^2$ also $\mathbf{a}_z$ is a gradient field:

$$\mathbf{a}_z = \nabla \cdot \mathrm{Z} \text{ and } \mathrm{Z} = 1/2 \omega_\mathrm{E}{}^2 \mathrm{R_\perp}{}^2$$

Z is called the potential of the centrifugal acceleration and one obtains for the complete potential of the Earth with (Equation 9):

$$\mathrm{W} = \mathrm{V} + \mathrm{Z} \text{ and } \mathrm{g} = \nabla \cdot \mathrm{W}. \tag{13}$$



Gravity Method, Principles, Figure 2 To illustrate the calculation of gravity effects in a point P at the Earth's surface, which depend on the attraction of celestial masses (sun and moon) and the centrifugal acceleration (δ = pol distance, ω_E = angular velocity vector).

Anomalies

From Equation 13 the normal gravity field (γ_0) can be deduced from a "normal Earth body" (among others, e.g., Heiskanen and Moritz, 1967; Torge, 1989; Blakely, 1996). Geodesists and geophysicists agreed that the "normal gravity" is defined at an equipotential surface which is part of a spheroidal surface that bounds a rotating, uniformly dense Earth. Due to the gravity (attraction) and rotation forces, such a spheroid has a very comparable shape to an ellipse of rotation. Consequently, it is called a reference ellipsoid: in the WGS84 geodetic system, the flattening of the ellipsoid is f = 1/298.257223563. Readers wishing additional information are referred to publications by Li and Götze (2001) and Hackney and Featherstone (2003).

Local variations in altitudes and densities of crust and mantle, and geological formations cause variations in the Earth's gravitational field, known as gravitational anomalies (see *Gravity, Data to Anomalies*). Some of these anomalies can be very extensive and cause even bulges in sea level. The study of these anomalies forms the basis of the gravity discipline in geophysics. The anomalies are measured with extremely sensitive gravimeters, and today detected also with gradiometers. The effect of topography and other known factors (see below) are subtracted from the observations. Such techniques are used by geophysics to explore oil and mineral deposits and global structures. Denser rocks (often containing mineral ores) cause higher than normal local gravitational fields on the Earth's surface; less-dense sedimentary rocks cause negative anomalies (see *Gravity Anomalies, Interpretation*). The separation of gravity anomalies which are caused by local density variations of rocks and geological formations implicate corrections of the observed gravity. It was suggested by Blakely (1996) to describe these corrections as contributions to the field which was measured:

Observed gravity	= Attraction of the reference ellipsoid (normal gravity) + Effect of heights above geoid (FREE – AIR correction) + Effect of topographic masses above geoid (BOUGUER and TOPOGRAPHIC correction) + Effect of time-dependent variations (TIDAL correction) + Effect of moving platforms (EÖTVÖS corrections) + Effect of masses that support loads (ISOSTATIC corrections) + Effect of both known and unknown crust/mantle density variations (GEOLOGY)

It is the goal of each interpretation to isolate the effect of geological bodies from all other terms. We have to consider that the magnitude of this effect is relatively small if compared with the value of normal gravity:

Normal gravity approximately is	$9.81\ m/s^2$ and
Gravity effects of "Geology" approximately is	$10^{-5}\ m/s^2 - 10^{-3}\ m/s^2$

This is less than 0.01 % of the observed gravity field magnitude. In geophysics, one distinguishes the standard anomalies: free air, Bouguer, Isostatic, explained in the entry *Gravity, Data to Anomalies* in this encyclopedia.

Ambiguity and principle of equivalence

The interpretation of gravity anomalies is not unique and boundary condition/information is always required, mainly from other geophysical disciplines and/or geology. It can be easily shown by modeling that for a given anomaly and a given density contrast, a wide range of possible interpretations can be made: at various depths, based upon the different geometrical shapes. Nor does the method of interpretation by the gravity field gradients allow us to make a unique interpretation, or to distinguish deep from shallow anomalies as has been claimed: It has been shown that we cannot escape the ambiguity by using second derivative quantities (gradients) or curvature, and that, in fact, gravity and its derivatives are related by a corollary of Green's theorem. This theorem provides an analytical proof of ambiguity not only for the case of gravity data but for magnetic data as well (e.g., Skeels, 1947).

The inversion problem

The ambiguity of the inverse problem of gravimetry originates from potential theory (Stokes). It is related with the computation of density functions from the observed gravity field. Density functions yield information of location, shape, and densities of causing masses. The computation leads to an integral equation, for which no unique solution exists (see *Inverse Theory, Global Optimization*). For a gravity anomaly generated by a particular mass distribution, infinitely many equivalent disturbing masses can be constructed at depths above the maximum depth of a disturbing body (Figure 3). However, the total mass of the disturbing body (surface integral of the gravity anomaly), the lateral position of its mass center, and the corresponding coating on the surface of the Earth (special case of an equivalent disturbing mass) can be determined uniquely. If certain conditions are met, limiting values of maximal depth, density (difference), thickness, and lateral extension of the disturbing body can be found. The necessary additional independent information (constraints) is taken from other geophysical methods and geological and/or petrological investigations.

In practical solutions of inversion problems, the *direct problem* of gravity appears, i.e., the computation of gravitational effects caused by particular mass contributions given in terms of location, shape, and density. In this case, a unique solution is always possible by application of the law of gravitation, Equation 9.

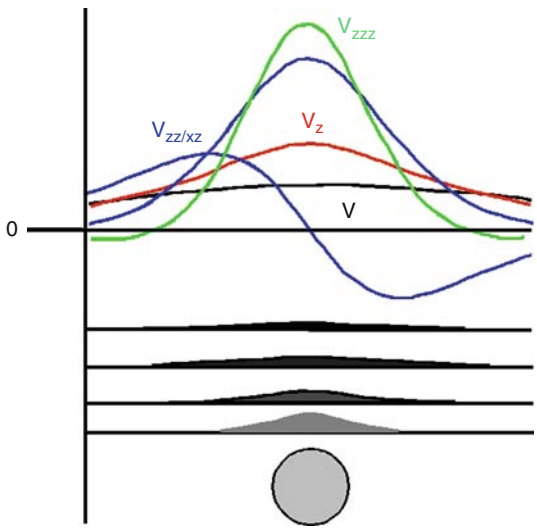
Units and gravitational constant

Gravity is measured in ms^{-2} which is the standard unit in the "système international d`Unités" (SI units are used in physics since 1972). The six independent components of the gravity gradients are measured in s^{-2}. For many applications of gravity field, the following units are useful:

$$1\ nm\ s^{-2} = 10^{-9} ms^{-2} \text{ or}$$

$$1\ \mu m\ s^{-2} = 10^{-6} ms^{-2}.$$

Also the older cgm – units are still used in geophysics and geodesy:



Gravity Method, Principles, Figure 3 Gravitational potential (V), vertical gravity component, and derivatives of gravity over a homogeneous sphere with a depth 5 km, a radius of 1 km, a density contrast of 1 t/m^3, and the equivalent masses between surface and maximum depth (After Torge, 1989).

$1\ \text{Gal} = 1\text{cm s}^{-2}$

$1\ \text{mGal} = 10^{-5}\text{ms}^{-2}$ or

$1\ \mu\text{Gal} = 10^{-8}\text{ms}^{-2}$.

"Gal" was (is) used in honor of the famous Renaissance Physicist Galileo Galilei (1564–1642). Although these older cgm units are no longer admitted, in applied geophysics they are still in use, together with "gravity units" (g.u. = 1 μm^{-2}).

Measurements and refinements of the gravitational constant have progressed since the time of Newton. However, up to now, laboratory experiments and physical research have not rendered a significant or generally accepted proof of $f \neq 0$. Fixler et al. (2003) reported on a new experiment. The Newtonian constant of gravity, f, was measured by using a gravity gradiometer based on atom interferometry. They report a value of $f = 6.693 \times 10^{-11}$ m / kg s^2, with a standard error of the mean of $\pm 0.027 \times 10^{-11}$ and a systematic error of $\pm 0.021 \times 10^{-11}$ m^3 / kg s^2.

Summary

This contribution describes the principles of the gravity method that was developed by Sir Isaac Newton in the seventeenth century and holds until today. The theory bases on the definition of a potential function which replaces potential energy. A space function and its second derivatives which are continuous and bounded in a region where defined are called harmonic functions if they fulfill the LAPLACE and – in the interior of a mass filled space – the POISSON differential equation. A normal gravity field can be deduced from a "normal Earth body" and geodesists and geophysicists agreed that a "normal gravity" is defined at an equipotential surface which is part of a spheroidal surface that bounds a rotating, uniformly dense Earth. Due to the gravity (attraction) and rotation forces such a spheroid has a very comparable shape to an ellipse of rotation. Consequently, it is called a reference ellipsoid. Local variations in altitudes and densities of the Earth's crust and mantle, and geological formations cause variations in the Earth's gravitational field, known as gravitational anomalies. In geophysics elevation is used in all the corrections including the height correction and the complete Bouguer correction. In addition, one should correct observed gravity for the geoid shape. The gravity effects due to the geoid undulations (N) are called the "indirect effects." Any gravity field interpretation effected by the ambiguity of the inverse problem of gravimetry which originates from potential theory; it is related with the computation of density functions from the observed gravity field.

Bibliography

Blakely, R., 1996. *Potential Theory in Gravity and Magnetic Applications*. Cambridge: Cambridge University Press.

Chapman, M. E., and Bordine, J. H., 1979. Considerations of the indirect effect in marine gravity modeling. *Journal of Geophysical Research*, **84**, 3889.

Fixler, J. B., Foster, G. T., McGuirk, J. M., and Kasevich, M. A., 2003. Atom interferometer measurement of the Newtonian constant of gravity. *Science*, **315**(5808), 74, doi:10.1126/science.1135459.

Götze, H.-J., and Li, X., 1997. Topography and Geoid effects on gravity anomalies in mountainous areas as inferred from the gravity field of the central Andes. *Physics and Chemistry of the Earth*, **21**(4), 295.

Hackney, R. I., and Featherstone, W. E., 2003. Geodetic versus geophysical perspectives of the 'gravity anomaly'. *Geophysical Journal International*, **154**, 35.

Heiskanen, W. A., and Moritz, H., 1967. *Physical Geodesy*. San Francisco: Freeman.

Janle, P., and Meissner, R., 1986. Structure and evolution of the terrestrial planets. *Geophysical Surveys*, **8**(2), 107.

Konopliv, A. S., Asmar, S. W., Carranza, E., Sjogren, W. L., and Yuan, D. N., 2001. Recent gravity models as a result of the lunar prospector mission. *Icarus*, **150**(1), 1, doi:10.1006/icar.2000.6573.

Li, X., and Götze, H.-J., 2001. Ellipsoid, geoid, gravity, geodesy and geophysics - a tutorial. *Geophysics*, **66**(6), 1660.

Mikuška, J., Pašteka, R., and Marušiak, I., 2006. Estimation of distant relief effect in gravimetry. *Geophysics*, **71**(6), 59, doi:10.1190/1.2338333.

Schmidt, S., and Götze, H.-J., 2002. Gravity data processing. In Flüh, E. R., Kopp, H., and Schreckenberger, B., (eds.), *Cruise Report SO161 - 1 & 4*, Subduction processes off Chile, Kiel, Germany: Research Center for Marine Geosciences at Christian Albrechts-Universität-Kiel, GEOMAR Report 102.

Skeels, D. C., 1947. Ambiguity in gravity interpretation. *Geophysics*, **12**, 43, doi:10.1190/1.1437295.

Talwani, M., 2003. The Apollo 17 gravity measurements on the moon. *The Leading Edge*, **22**(8), 786, doi:10.1190/1.1605083.

Torge, W., 1989. *Gravimetry*. New York: de Gruyter Berlin.

Vening-Meinesz, F. A., 1931. Une nouvelle méthode pour la réduction isostasique régional de Íintensité de la pesanteur. *Bulletin Géodésique*, **29**, 33–51.

Watts, A. B., 2001. *Isostasy and Flexure of the Lithosphere*. New York: Cambridge University Press.

Cross-references

Geodesy, Figure of the Earth
Gravity Anomalies, Interpretation
Gravity Field of the Earth
Gravity Method, Satellite
Gravity Method, Surface
Gravity Modeling, Theory and Computation
Gravity, Data to Anomalies
Inverse Theory, Global Optimization
Isostasy

GRAVITY METHOD, SATELLITE

Georges Balmino
Centre National d'Etudes Spatiales, Toulouse, France

Definition

After recalling the need for measuring the Earth's gravity field at fine spatial scale and high precision with respect to

applications in the geosciences, we broadly review the classical ways of determining global gravity models. We then describe the new satellite gravity missions with their basic principles, and finally give our vision of the future in this domain.

Details in a gravity model recovery (functional representation, dynamical approach, inverse problem of celestial mechanics, disturbing forces, etc.) and in the principles of satellite-to-satellite tracking and satellite gradiometry (and instruments) are to be found in other chapters of this Encyclopedia. Broad concepts only and most important mathematical methods are given here.

Introduction

The knowledge of the gravity field of the Earth and of an associated reference surface of the altitudes (the geoid) is necessary for geodesy, for improving theories of the physics of our planet interior and surface, including various mass transfer mechanisms, and for modeling the ocean circulation in absolute. This knowledge comes from several observing techniques; on the global scale it mostly benefited from the observation of artificial satellite orbits and has been improved over decades. The current approach, based on the accumulation of information from a large number of satellites, one being observed at a time, reached its limits around the year 2000. Dramatic improvements now come from newly implemented space techniques already proposed in the 1970s and rendered possible thanks to technological progress; they are exemplified by the CHAMP, GRACE, and GOCE missions, which are realizations of the satellite-to-satellite (SST) tracking principles (in the high-low and low-low modes), of satellite gradiometry, together with the systematic use of micro-accelerometry to measure the surface forces that are a source of uncertainty in gravity field mapping from satellites.

Geodesy, gravity field, and scientific applications: why?

Geodesy is a science whose major goal is to define and study the shape of the Earth, its deformations (from relative movements of points on its surface), and its gravitational field (or its gravity field if one includes the rotational effects). It is an old discipline that acquired its fame in the eighteenth century when French scientists went measuring arcs of meridians – in Lapland and Peru, in order to determine whether the Earth was flattened at the poles or at the equator. Two hundred years later, the use of artificial satellites yielded a global and continuous view of our planet, and geodesy, too, benefited from this revolution. It became a science at the crossroads of others; for instance it is used daily for navigating satellites with high precision for oceanography, geophysics, and a variety of societal applications. In the mid-1960s first global models of the Earth's gravity field from satellites initiated a revolution in geophysics from space; about 40 years later we are witnessing a new era that sees the deciphering of our planet's gravity with unprecedented resolution and accuracy.

For a better understanding of both surface and deep phenomena that are related to the interior structure of the Earth and its temporal evolution, for studying the dynamics of the oceans and their interaction with meteorological and climate changes, for modeling the ice caps–oceans–continents relationships, and for predicting the long-term evolution of the mean sea level, also for unifying vertical reference systems and for the precise determination of satellite orbits in space, it is mandatory to know our planet gravity field to ever increasing resolution and precision – including its temporal variations induced by the aforementioned phenomena. Details are provided by Jekeli (*this Encyclopedia*); we limit ourselves to qualitative arguments on the dual role of the gravity field in Earth sciences.

On the one hand, by comparing the real field with the field of an idealized body (e.g., an ellipsoid in hydrostatic equilibrium) one defines gravity anomalies that characterize deviations from a state of internal equilibrium (that is non-radial density variations), which constitutes one method of sounding our planet's interior. The magnitude of gravity anomalies is in the range −1,000/+1,000 mgals ($1\ \text{mgal} = 10^{-5}\ \text{ms}^{-2}$). As a matter of fact this is one of four ways of "looking" inside, the other three being seismology, magnetic field studies, and the analysis of the Earth's rotation. However, the gravity analysis approach is unique in providing direct information on the density field (although this is integrated and cannot be univocally inverted). When one strips the Earth from its topographic blocks (with assumed, or known density) one is left with residual gravity anomalies whose magnitude is not much different from the magnitude of the original ones; this exemplifies the phenomenon of isostasy, which is the concept of mechanisms of support of the topographic masses. Therefore the residual field mirrors mass excesses or deficits in depth, which inform us on the lateral structure of the lithosphere and upper mantle. A significant part of this information is today provided by seismology, by tomographic analysis, but still suffers from uncertainties due to hypothesis on which such inversions are based. The combination of both types of data (seismic velocities and gravity anomalies) is a powerful tool to get a better picture of the interior and to make progress in the understanding of several phenomena for instance the accumulation of stresses and triggering of earthquakes.

On the other hand a surface that is intimately related to gravity, the geoid, is used as the reference for defining and measuring the altitudes on the continents but also under the oceans or over the ice caps. The geoid is a particular equipotential surface of the gravity potential (the sum of the gravitational and centrifugal potentials), which may be viewed, in oceanic areas, as the surface of an ocean at rest. The Earth's rotation being sufficiently well known and its effects being smooth at the surface (they essentially decrease with increasing latitudes), the irregularities of the geoid (in the ±100 m range, as measured with respect to an ellipsoid of revolution that approximates the Earth's

shape) characterize the density field variations in a way similar to the gravity anomalies. In addition, on the topographic surface water naturally flows along the geoid (or other equipotential) slope; over the oceans sea water circulates (under the forcings of winds and density-salinity fluctuations) and these movements (different from the vertical displacements of water due to tides) and associated amounts of transported water (and heat, chemicals, nutrients, etc.) can be quantified with respect to the geoid: these are the ocean currents. One distinguishes between the mean and the variable parts of the circulation: the latter can nowadays be monitored by satellite altimetry, whereas the former (which is needed in the modeling of the climate by quantifying the long-term absolute transports of heat – especially in the shallow parts of the oceans) requires the precise knowledge of the difference between the mean sea surface and the geoid (the so-called sea surface topography); the mean sea surface itself can be deduced from long series of altimeter measurements but it is demanded that the geoid be independently determined.

For both solid Earth physics and oceanography, the requirements today are (at least) for a 100-km resolution geoid (and gravity field) with a cumulated error – that is, up to that resolution, of 1 cm (and 1 mgal) or better. Clearly the classical domains of geodesy (unification of vertical reference systems, inertial navigation, accurate satellite orbits for monitoring the Earth's deformations, the sea surface, the Earth's kinematics, etc.) will greatly benefit from such knowledge.

In addition to this, there is a growing interest, in relationship with environmental studies and prediction capabilities, in measuring the time variations of the gravity field (see *Gravity Field, Temporal Variations from Space Techniques*). Some of them (e.g., the oceanic and solid tides) are deduced from the measurements or modeling of phenomena that are responsible for them, with satisfactory precision and resolution – or close to be satisfactory. On the contrary others can only be determined from direct observations, from which the responsible phenomena are then modeled. These are the postglacial rebound of the lithosphere (therefore of the planet surface); the variations of the ice caps, of the sea level, of the large continental aquifers; and the distribution of precipitations (rain, snow), the soil moisture, the evapo-transpiration, that is, the water cycle in general.

Classical ways of measuring the Earth's gravity field

Up to a few years ago three types of information have been used to describe or model, at local, regional, and/or global scales, the Earth's gravitational potential and various functionals:

- Surface measurements of gravity, on land, at sea (with ships and submarines), by plane (more recently). This is the oldest approach, much improved along its history. The measurements are often reduced to gravity anomalies, then gridded and made available at a variable resolution that depends on the area, on the actual density of observations, also on restrictions still exerted by many countries for protecting national economic (sometimes military) interests. The resolution of such grids varies from a few to hundreds of kilometers, and the precision ranges from about 1–10 mgals – the accuracy may be worse due to systematic errors in many data sets (such as the old marine surveys). The data are archived at world level, mainly at the National Geo-Intelligence Agency (NGA) in the USA, and at the Bureau Gravimetrique International (BGI) in France. Airborne measurements are slowly expanding but the precision at ground level (after downward continuation) is still insufficient. More information on gravimetry is provided in cross references.
- The classical analysis of satellite orbit perturbations: since the 1960s a few groups in the world have determined global gravity field models from the inversion (via the law of the dynamics) of satellite orbit perturbations – which reflect the forces acting on a satellite. This has also been the current and very successful way of determining the gravity field of other solar system bodies (Nerem, 1995). Most used Earth's satellites were indeed never designed for such purpose: the space geodesists simply took benefit of the more or less accurate measurements made from ground stations on these satellites for orbit knowledge. Various techniques were exploited, essentially the Doppler effect on the signal transmitted from a stable oscillator on board a spacecraft to ground stations – or the reverse (case of the DORIS system), and roundtrip ground to satellite distances measured by laser systems – and optical determination of the satellite positions on the star background was also used at the beginning. The main limitations of these models come from the altitude of the observed satellites (due to the quasi-exponential decay of the perturbations with altitude), from the observation coverage (limited by the ground station networks and pass duration), and from uncertainties in modeling the surface forces (atmospheric drag, solar radiation pressure, and Earth albedo). If the data coverage has improved for some satellites thanks to the development of denser station networks (e.g., the DORIS system), or by starting to track satellites from "above," that is, by the GPS (Global Positioning System) constellation, the other limitations remain and prevent from making significant progress. Solutions derived from satellite data only up to the year ~2000, covered spatial scales from 500 to 40,000 km, but they were reliable (say at the centimeter level on the geoid knowledge) only down to a resolution of about 1,000–1,500 km. That is why they have been – and are still combined with surface gravimetric measurements and altimetric data (see below), which bring higher resolution if not precision – depending on the area; such combined models globally reach 50 km resolution. The same strategy recently applied with a much better satellite solution based on GRACE mission data yielded the EGM 2008 model,

which has, in principle, a 10 km resolution though it represents the field with such details solely over areas with enough surface data. A detailed account of global gravity modeling history, techniques, and satellites used so far is given by Pavlis (*this Encyclopedia*). In a subsequent section we give an overview of the mathematical methods used to retrieve gravity field information from satellite dynamics.

- Satellite altimetry over the oceans: this is a space technique by which one measures (by radar or laser) the distance between a spacecraft and the sea surface; provided that the spacecraft position is known with sufficient precision it provides the sea surface whose time average is close to the geoid but differs from it (up to about 1 m in areas of strong currents such as the Gulf Stream, the Arctic Circum-polar Current, the Kuro-Shio). For several applications in marine geodesy and geophysics, it is sufficient to correct for the mean circulation with the present (approximate) models, though coastal currents, uncertainties in tidal models, and larger error measurements close to the shore limit the usage of the information to the off-shore and open ocean areas. Such restriction being understood, satellite altimetry data have had an important contribution to global Earth's gravity models in increasing their resolution over the oceans – but without providing the true geoid.

New satellite missions

Basic principles were found years ago; technological advances have allowed for their implementation.

To get rid of the limitations of past systems, any new satellite approach should satisfy as many of the following criteria as possible: quasi-polar orbit (for the coverage), 3-D uninterrupted measurements (for recovering the gravity signals with an isotropic error), mean altitude as low as possible (for higher resolution), counteract the effects of signal attenuation with altitude by a differencing measurement method, isolate the gravitational signals by measuring and/or compensating the surface forces (which involves ultra-sensitive accelerometers). The missions launched during the 2000–2010 decade satisfy more or less these criteria. Working at low altitude is the constraint that has received most attention and all missions fly at or below 450 km (compared to the altitudes of satellites used so far: 800 km or larger). Surface forces are measured by micro-accelerometers (and even compensated in the case of the gradiometric mission). The other criteria have been considered differently according to the type of mission and the used technology.

The first dedicated satellite mission, launched in 2000, was CHAMP (Challenging Mini-Satellite Payload for Geophysical Research and Application) (Reigber, 1996); its two main objectives were the mapping of the magnetic field and of the gravity field of the Earth. It was developed by Germany, with French contributions (magnetometers, and a tri-axial electrostatic micro-accelerometer) and the provision of a GPS receiver by the USA; this is one realization of the SST approach in the high-low mode (with a lower segment altitude of 450 km, decreasing to about 320 km toward the end – probably in 2011), which resulted in the first significant leap in gravity modeling precision.

CHAMP appears as the precursor of the second gravity mission GRACE (Gravity Recovery and Climate Experiment), which is a US mission (JPL, 1998) conducted in cooperation with Germany and launched in 2002. It consists in flying two satellites separated by about 200 km from each other, on the same mean polar orbit; the satellites are each tracked by the GPS (like CHAMP) and besides measure their relative distance and velocity with a micrometric accuracy – this observable (of the low-low SST type) may be viewed as an extremely fine measurement of the gravitational potential difference at the satellites. Each spacecraft is equipped with a micro-accelerometer (of a class superior to the CHAMP one). The mission has exceeded its designed life time (like CHAMP); besides improving our knowledge of the mean gravity field (by more than one order of magnitude), it is bringing a wealth of information on the seasonal and inter-annual time variations of the gravity field, especially those related to the hydrological cycle and ice cap volume changes.

The third mission, which was selected in November 1999 as the first core mission of the Earth Explorer program of ESA (European Space Agency) is GOCE (Gravity field and steady-state Ocean Circulation Explorer) (ESA, 1999) launched in 2009. For the first time, a gradiometer is flown on board a satellite on a very low altitude orbit (around 250 km) of 96.5° inclination, with drag-free control. The gradiometer instrument consists in six ultra-sensitive micro-accelerometers of a new class (capable of measuring accelerations of 5.10^{-13} ms^{-2}) and delivers gravity gradients (second derivatives of the gravitational potential) in spacecraft axes with an accuracy of a few milli-Eötvös (1 E = 10^{-9} s^{-2}) per $Hz^{1/2}$ in the 0.005–0.1 Hz measurement bandwidth. The satellite is also tracked by the GPS (and by ground laser stations, actually like CHAMP and GRACE – for control and safety), and the combination of these trajectory observations and of the gradiometer measurements should yield a global gravity model with a resolution of about 100 km at least with a total uncertainty of ~2 cm on the geoid.

Overview of satellite dynamics methods

In the following we describe the main mathematical concepts and tools that are in common use in satellite geodesy, that is:

1. The representation of the gravitational potential
2. The solution of the dynamical problem, with two subparts:
   - The equations of motion and their solution

- The equations, derived from the preceding ones, necessary for obtaining the trajectory and physical parameters at stake (i.e., for the inverse problem)

3. The setting up of the observation equations and their linearization in the inverse problem
4. The solution of the inverse problem

Some of the described material may be redundant with Pavlis (ibid).

The representation of the gravitational potential

For most global studies the potential of the Earth's attracted masses is usually modeled by a spherical harmonic series, written at each point P as:

$$U = \frac{GM}{r} + \sum_{l>0} \frac{GM}{r}\left(\frac{R}{r}\right)^l \sum_{m=-l}^{+l} K_{lm} P_{lm}(\sin\phi) e^{im\lambda}$$

with G: gravitational constant, M: mass of the Earth, K_{lm}: dimensionless harmonic coefficient of degree l and order m, and r, φ, λ are the spherical coordinates of the point P in a reference system fixed with respect to the Earth; the P_{lm}'s are the Legendre functions of the first kind (polynomials when $m = 0$) and, for computational stability reason, are usually normalized (like the K_{lm}'s) – which we will assume in the following.

Alternatively, the sum of all terms of the same degree l can be written with real (normalized) coefficients and functions, in the form:

$$U_l = \frac{GM}{r}\left(\frac{R}{r}\right)^l \sum_{m=0}^{+l} (C_{lm}\cos m\lambda + s_{lm}\sin m\lambda) P_{lm}(\sin\varphi)$$

The C_{l0} coefficients are called zonals, the C_{lm}, S_{lm} ($m > 0$) are the tesseral harmonics with the sectorials ($l = m$) being a special subclass. They may vary with time.

Practically the potential series is truncated at maximum degree L, that is, $|m| \le l \le L$, and therefore dim $\{K_{lm}\} = (L + 1)^2$. L depends on the type of satellite orbits and data sensitivity.

When the K_{lm} are known, it is straightforward to compute any geodetic functional F, which is linear in these parameters:

$$F(U) = \sum_l f_l U_l$$

For instance:

- For the gravity anomaly field: $F = \Delta g$ and $f_l = (l\text{-}1)/r_E^2$
- For the geoid height: $F = N$ and $f_l = 1/\gamma\ (r_E, \varphi)$

where r_E is the radius vector and $\gamma\ (r_E, \varphi)$ is the normal gravity at the evaluation point on some adopted reference ellipsoid (E) approximating the shape of the geoid.

Variances and covariances can also be propagated to these functions from the covariance matrix (see further down) of the spherical harmonic coefficients (in particular the variances δC^2_{lm} and δS^2_{lm}). Useful indicators of the precision of a model are the mean errors per degree on some given geodetic functional – usually the geoid and/or the gravity anomaly field, computed as:

$$< \delta F >_l = f_l \left[\sum_{m=0}^{l} \left(\delta C_{lm}^2 + \delta S_{lm}^2\right)\right]^{\frac{1}{2}}$$

The solution of the dynamical problem

The equations of motion are most often considered in rectangular coordinates with respect to a given reference system, and written in a specific time scale. Although approximate analytical theories exist for describing the evolution of an orbit subject to disturbing forces (they are still in use for mission analysis and first order scientific investigations), numerical integration is systematically used for higher precision. Forces acting on a satellite are not only of gravitational origin and their complete modeling involves parameters, such as the K_{lm} coefficients, whose determination requires integrating other systems of equations associated with those providing the trajectory.

Reference system and time scale

One makes a choice of a time scale, such as the Geocentric Coordinate Time or the Terrestrial Time, and of reference systems: one related to the Earth's surface (the International Terrestrial Reference Frame, or ITRF), and one fixed in space (the International Celestial Reference Frame, or ICRF) as defined by the International Astronomical Union and the International Association of Geodesy (IERS, 2004). The K_{lm} coefficients are expressed in ITRF, and the equations of motion of a satellite are usually written in a system (EME2000) defined by the Earth mean equator and mean equinox of epoch J2000 (January 1, 2000 at 12 h) and fixed in ICRF. The transformation from ITRF to EME2000 involves all phenomena related to the rotational motion of Earth in space, that is precession, nutation, sidereal time, and polar motion – of which parameters may be part of the unknowns in the inverse problem.

Equations of motion

The spacecraft motion obeys the differential equations, in EME2000:

$$\ddot{\bar{r}} = \bar{\nabla} U + \bar{\nabla} U^* + \bar{\Gamma}_{SP} + \bar{\Gamma}_D + \bar{\Gamma}_{th}$$

with $\bar{r}(t_0) = \bar{r}_0$; $\dot{\bar{r}}(t_0) = \dot{\bar{r}}_0$ (initial conditions at beginning of a piece of trajectory – an arc).

In the above, U is the Earth's gravitational potential, U^* the so-called third body gravitational potential, Γ_{SP} is the acceleration vector due to solar pressures (direct from the Sun, indirect, i.e., reflected by the Earth – known as the albedo effect) and infrared pressure (radiated by the Earth considered as a black body), Γ_D is the acceleration due to drag from the atmosphere, and Γ_{th} designates the residual acceleration due to poorly known effects

(e.g., thrusters accelerations in the case of an orbit controlled spacecraft). U^* includes the Sun, the Moon, and planets.

For computing the surface accelerations Γ_{SP} and Γ_D, a so-called box-and-wing (macro) spacecraft model (with a few planar faces) is used (together with ad hoc physical coefficients), and one has also to use an attitude model. Alternatively, Γ_{SP} and Γ_D may be measured by very sensitive micro-accelerometers such as in the CHAMP, GRACE, and GOCE satellites. The Γ_{th} effects may be modeled via time series or as window (rectangular) functions at some specific epochs (for instance the times of attitude maneuvers).

Integration of the equations of motion

This is the direct dynamical problem. Numerical integration is systematically used. Robust multistep integrators, of the Adams, Adams–Moulton, or Cowell type, in normal form or in summed form (which reduces the accumulation of round-off errors), are preferred. Such integrators have a starting phase, then a running phase with a predictor–corrector iterative scheme. One big advantage of such class of methods is that one can easily speed up the algorithm, by sometimes a dramatic factor in performing a pseudo-correction by which the full force function (right-hand side member of the dynamical equations) is replaced (after the first loop of the corrector part) by an approximation (e.g., the central term of the gravity field plus a few other important terms).

Variational equations and their integration

The inverse dynamical problem consists in determining the physical model parameters (or a subset of them) from the observation of the satellite trajectory – by any mean. It also implies the determination of the equations of motion initial conditions in a first step or simultaneously (usually by iteration).

As it is known, and as we describe it in the next subsection, this requires to obtain at each time t the partial (first) derivatives of the satellite state vector $(\bar{r}, \dot{\bar{r}})$ with respect to the components of the initial conditions vector (which we call $\bar{Z}$) and similarly with respect to the vector of all searched for physical parameters (which we call $\bar{P}$). This is achieved via the variational equations.

We rewrite the equations of motion as follows:

$$\ddot{\bar{r}} = \bar{\gamma}\ (\bar{r}, \dot{\bar{r}}, t, \bar{P})$$
$$\text{with } (\bar{r}, \dot{\bar{r}})_{t0} = \bar{Z}$$

$\bar{r}$ and $\dot{\bar{r}}$ are functions of t, t_0, $\bar{Z}$, and $\bar{P}$. Let w be any Z_j or P_k. Then:

$$\frac{\partial}{\partial w}\ddot{\bar{r}} = \frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(\frac{\partial \bar{r}}{\partial w}\right) = \frac{\partial \bar{\gamma}}{\partial \bar{r}}\left(\frac{\partial \bar{r}}{\partial w}\right) + \frac{\partial \bar{\gamma}}{\partial \dot{\bar{r}}}\left(\frac{\partial \dot{\bar{r}}}{\partial w}\right) + \frac{\partial \bar{\gamma}}{\partial w}$$

We note that $\partial \dot{\bar{r}}/\partial w = d(\partial \bar{r}/\partial w)/\mathrm{d}t$ and that the last term is present only if $w = P_k$.

Therefore any vector $\bar{g} = \partial \bar{r}/\partial w$ is solution of the system of so-called variational equations:

$$\ddot{\bar{g}} = \frac{\partial \bar{\gamma}}{\partial \bar{r}}\,\bar{g} + \frac{\partial \bar{\gamma}}{\partial \dot{\bar{r}}}\,\dot{\bar{g}} + \begin{Bmatrix} 0 \\ \partial \bar{\gamma}/\partial w \end{Bmatrix}_{w=P_k}^{w=Z_j}$$

with

$$g_i(t_0) = \dot{g}_i(t_0) = 0 \quad \forall i \text{ if } w = P_k, = \delta_{ij} \text{ if } w = Z_j,$$

since this is nothing but the Jacobian $\left[\frac{D(r_i, \dot{r}_i)}{D[r_i(t_0), \dot{r}_i(t_0)]}\right]_{t0}$.

We have one such system for each parameter to be estimated, which could be very costly to integrate, but we note that: $\partial \bar{\gamma}/\partial(\bar{r}, \dot{\bar{r}})$ is common to all systems $(\forall w)$ and can be evaluated once (at convergence of the predictor–corrector loop) at each stepsize in the course of the integration of the main system; these variational equations are linear (with coefficients function of time); therefore they can be economically integrated by quadrature or, more efficiently, by using solely the predictor part of the integrator (another approach uses the matrizant of the associated homogeneous system, followed by quadratures).

The observation equations for the inverse problem

The recovery of the gravity field model parameters is an inverse problem that consists, here, in deriving the gravity coefficients (K_{lm}) from a set of observation equations.

General form of the equations

Each measurement q of a functional f of the satellite position (or velocity) is of the type:

$$q = f(\bar{r}, \dot{\bar{r}}, t, \bar{Q})$$

or:

$$q = f(\bar{r}_1, \bar{r}_2, t_1, t_2, \bar{Q})$$

Direct measurements of a gravity field functional (such as gravity gradients along one or several spacecraft axis) are of the form:

$$q = f(\bar{r}, \dot{\bar{r}}, t, \bar{P})$$

In the above, $\bar{r}, \dot{\bar{r}}, \bar{r}_1, \bar{r}_2$ are usually functions of $\bar{Z}$ and $\bar{P}$, respectively, the initial condition (state) vector (often comprising other "fudge" parameters – e.g., empirical multiplying factors for the surface forces) and the dynamical parameter vector (of which the gravity field parameter is a subset). $\bar{Q}$ is a vector of geometric quantities (for instance coordinates of observing stations) and/or other physical parameters (e.g., propagation delays in a medium).

The observation equations are usually nonlinear with respect to the unknown parameters. The classical approach is to linearize them around starting values:

$$\Delta q = q^{\text{obs}} - q^{\text{comp}} = \frac{\partial q^{\text{comp}}}{\partial \bar{Z}} \Delta \bar{Z} + \frac{\partial q^{\text{comp}}}{\partial \bar{P}} \Delta \bar{P} + \frac{\partial q^{\text{comp}}}{\partial \bar{Q}} \Delta \bar{Q}$$

with

$$\Delta \bar{Z} = \bar{Z} - \bar{Z}_0, \quad \Delta \bar{P} = \bar{P} - \bar{P}_0, \quad \Delta \bar{Q} = \bar{Q} - \bar{Q}_0$$

Partial derivatives are computed via the variational equations of the equations of motion of the satellite(s) when q is a distance, velocity between an observing reference point (a tracking station on Earth, another spacecraft – like in the Earth gravity mapping missions CHAMP, GRACE and GOCE), or are directly evaluated if q can be expressed as a simple function of the physical phenomena at stake at the satellite location, such as a gravity gradient in the case of a gradiometer sensing the gravitational potential (case of the GOCE mission).

More precisely, we have

$$\frac{\partial q^{\text{comp}}}{\partial \bar{Q}} = \frac{\partial f}{\partial \bar{Q}}$$

$$\frac{\partial q^{\text{comp}}}{\partial (\bar{Z}, \bar{P})} = \frac{\partial f}{\partial \bar{r}} \frac{\partial \bar{r}}{\partial (\bar{Z}, \bar{P})} + \frac{\partial f}{\partial \dot{\bar{r}}} \frac{\partial \dot{\bar{r}}}{\partial (\bar{Z}, \bar{P})}$$

(or $\partial f / \partial \bar{r}_1$, $\partial f / \partial \bar{r}_2$ if the observable involves two spacecrafts with both orbits being adjusted simultaneously like in GRACE). In current observation systems $\partial f / \partial \dot{\bar{r}}$ is not present: the only case would be a Doppler type measurement, but this is always treated as a difference of ranges (between the beginning and the end of a given count time).

The quantities $\partial f / \partial \bar{Q}$, $\partial f / \partial \bar{r}$ are in principle easy to compute at the observation epoch although in practice one has to take account of the signal propagation duration, of small motions at the observing stations (solid and fluid tides loading effects, tectonic movements), of relativistic clock effects, etc.

Finally $\partial \bar{r} / \partial (\bar{Z}, \bar{P})$ is obtained by integration of the variational equations as described in the previous subsection.

Example: range measurement generic observation equation

All tracking observations of a satellite orbit involve one or several range values. Consequently, it is logical to construct a generic observation equation for a single range measurement between two points S_1, S_2, being understood that the final equation for a real measurement is a combination of generic equations. Besides this allows an easy generalization to the case when the observing station is replaced by another spacecraft (a high altitude navigation satellite – e.g., from the GPS constellation, or a co-orbiting satellite like in the GRACE mission). Therefore let us consider the case where S_2 is the satellite of interest and S_1 can be a station or another spacecraft.

Firstly propagation times must be taken into account by an iterative scheme that allows to refer the quantities to the true epochs (t_1 for S_1 and t_2 for S_2). Various corrections being properly applied the range S_1S_2 then reduces to the geometric quantity:

$$r_{12} = \|\overline{r_2} - \overline{r_1}\|$$

The observation equation is therefore:

$$r_{12}^{\text{obs}} - r_{12}^{\text{comp}} \approx \mathrm{d}r_{12} = \hat{u}_{12.}(\mathrm{d}\overline{r_2} - \mathrm{d}\overline{r_1})$$

with $\hat{u}_{12} = \bar{r}_{12}/r_{12}$ and:

$$\mathrm{d}\bar{r}_{n,i} = \sum\nolimits_j \frac{\partial \bar{r}_{n,i}}{\partial (PInt^j_{n,i})} \Delta (PInt^j_{n,i}) + \sum\nolimits_k \frac{\partial \bar{r}_{n,i}}{\partial (PExt^k)} \Delta (PExt^k)$$

($n = 1, 2$).

The quantities with comp superscript are derived from the numerical integration of the equations of motion of the satellite(s), the Jacobian matrices $\partial\{...\}/\partial\{P...\}$ are obtained from the integration of the variational equations; the i subscript refers to the ith arc in a strategy where the ensemble of the processed data is split into several batches, and we have distinguished between internal (or inner) parameters (proper to the ith arc) and external parameters (common to all arcs). For each batch of data the inner parameters are solely solved for in an iterative process (because of significant non linearities); at the end all parameters are considered and the equations are solved in ensemble. The whole procedure may then be iterated.

The solution of the inverse problem

The observation equations have been linearized, which yield the system:

$$Ax = b$$

where x is the vector of unknowns (of dimension N), A is a M*N matrix (M is the number of observations and usually M> > N), and b the residual vector (of the Δqs). A weight matrix Π (M*M) is associated to the observations; it is very often reduced to its diagonal part for practical reasons.

These equations can be solved by least squares, usually by forming the normal equation system:

$$\left(A^T \prod A\right) x = A^T \prod b$$

from which x is computed; $\boldsymbol{N} = A^T \Pi A$ is a (N*N) matrix.

A regularizing procedure may be used when the observability and/or sufficient decorrelation of all parameters are not guaranteed; in such a case $\boldsymbol{N}$ is replaced by $\boldsymbol{N} + \boldsymbol{R}$, where $\boldsymbol{R}$ is the regularizing matrix. This is usually

necessary for the inner unknowns (initial orbital elements, empirical parameters, etc., which may not be fully determinable). For the physical parameters, $\boldsymbol{R}$ may be as follows:

- Derived from an approximate physical law (e.g., Kaula's law for the gravitational harmonic coefficients – usually applied above a certain degree or for particular classes of coefficients such as the zonals)
- Obtained from a priori physical information

The covariance matrix of x is finally computed as:

$$\mathrm{cov}(x) = \sigma_0^2 (N + R)^{-1}$$

where σ_0 is the variance of unit weight estimated from the system and residual vector.

Other methods (orthogonal decomposition – e.g., using QR factorization) may be used for computational accuracy reasons.

The computational cost of the formation of the normal system and of its solution is as follows:

- Proportional to MN^2 for forming the normal equations, or for an orthogonal decomposition
- Proportional to N^3 for obtaining the solution *and* the covariance matrix

The first step is obviously the most demanding and this aspect has always driven the strategy of research teams working on the determination of global gravity field models.

Progress made: a short summary

Early in the space age several research groups started working on the determination of the gravity field of the Earth. One of the first global models published was the Standard Earth I, by an American team at the Smithsonian Astrophysical Observatory (Lundquist et al., 1966); it was a solution derived solely from analytical theory and satellite tracking (optical) data and complete to degree and order 8 (plus a few so-called resonant harmonics of higher degree and order). It was quickly followed by a combined solution (to degree-order 16), which was based on a little more satellite data and on mean surface gravity values (Gaposchkin and Lambeck, 1969); this model was enriched a few years later (Gaposchkin et al., 1973). Meanwhile similar work (but based on numerical integration approach) had started at the NASA Goddard Space Flight Centre (GSFC), and in Europe where two groups (at CNES, the French Space Centre, and at the Technical University in Munich) joined their efforts. This resulted in two series of models: the GEM models on the American side (Lerch et al., 1972a, b), the GRIM models on the European side (Balmino et al., 1976a, b). For several years these two series of models were improved, the GSFC team joined efforts with a group from the University of Texas at Austin, the German side of the European alliance was transferred at the DGFI (Deutsches Geodatisches ForschungsInstitute) then at the GFZ (GeoForschungsZentrum). All teams involved in this domain produced many global models, each milestone being marked in general by a new satellite solution, followed by a combined model, that is, with the added (higher spatial frequency) contribution of surface gravimetry and ocean altimetry (see Pavlis, ibid.).

At the end of the twentieth century, models were limited (mostly in the precision of knowledge of their long and medium wavelength components) by the weaknesses of the satellite solutions; for instance the cumulated geoid error at degree 50 (400 km resolution) was still over 1 m. The revolution came with the advent of the new dedicated gravity missions: CHAMP quickly provided a gain of a factor 5–10 (depending on the resolution) to the models derived from it (and from other good past satellite data), GRACE yielded models whose quality now reaches a centimeter error at degree 100 (allowing the study of time variations); GOCE looks very promising after 1 year in orbit, with its capability of complementing GRACE mean (static) models in the higher degrees (very likely up to degree 250 or even above). Due to the difficulty and importance of this type of research (whose outputs require extremely thorough validation – by cross-verifications), many more groups have engaged in the processing of the new missions data, such as the Institute of Theoretical Geodesy at the University of Bonn, which recently produced a very good gravity model from GRACE (Mayer-Gürr et al., 2010).

Finally, Table 1 lists some of the main satellite only solutions published since 1966.

Conclusion and vision of the future of gravity mapping from satellite

New progress can only come from a space approach for this only can satisfy the scientific needs (homogeneous data, full coverage, higher resolution and precision) within reasonable time and cost.

From the results already obtained, which demonstrate that the new gravity missions are (close to) meeting the expected outcome, one can say that several of the objectives previously discussed have been reached. However, and like in other domains, new requirements are emerging, encouraged by the findings of these missions: higher resolution and precision models and therefore measurements are needed.

If we restrict ourselves to the approach by satellite-to-satellite tracking or satellite gravity gradiometry, which are used in the new satellite missions of this decade, we may foresee improvements:

– On the precision of deciphering the time variations of the gravity field at shorter spatial scale, by exploiting further the co-orbiting satellites case like in GRACE; this could be done by performing inter-satellite range measurements by laser interferometry, a technique that has been under study and development (at laboratory scale) for several years

Gravity Method, Satellite, Table 1 Some landmarks of global gravity field modeling

Model	Year	Degree	Data	Reference
ITG-Grace2010s	2010	180	S(Grace)	Mayer-Gürr et al., 2010
GGM03S	2008	180	S(Grace)	Tapley et al., 2007
ITG-Grace02s	2006	170	S(Grace)	Mayer-Gürr et al., 2006
EIGEN-GRACE02S	2004	150	S(Grace)	Reigber et al., 2005
EIGEN-GRACE01S	2003	140	S(Grace)	Reigber et al., 2003c
EIGEN-2	2003	140	S(Champ)	Reigber et al., 2003b
EIGEN-1	2002	119	S(Champ)	Reigber et al., 2003a
EIGEN-1S	2002	119	GRIM5,S + Champ	Reigber et al., 2002
GRIM5S1	1999	99	S	Biancale et al., 2000
GRIM4S4	1995	70	S	Schwintzer et al., 1997
JGM2S	1994	60	S	Nerem et al., 1994a
JGM1S	1993	60	S	Nerem et al., 1994a
GEMT1	1987	36	S	Marsh et al., 1988
GEM9	1977	20	S	Lerch et al., 1979
GEM7	1976	16	S	Wagner et al., 1976
GRIM1	1975	10	S	Balmino et al., 1976a
GEM5	1974	12	S	Lerch et al., 1974
GEM3	1972	12	S	Lerch et al., 1972b
GEM1	1972	12	S	Lerch et al., 1972a
SE1	1966	8	S	Lundquist and Veis, 1966

– On the resolution of the static field (mapping of finer lateral variations of gravity) by a satellite-borne gradiometer of enhanced sensitivity, such as a cryogenic instrument that had been studied in the USA and in Europe

Future progress may also be possible by revisiting the differential approaches, for instance by differencing (or double differencing) relative velocity measurements, or by flying two or more satellites in specific geometries (gravimetric "wheel," pendulum configuration, etc.).

But there comes a critical aspect: GRACE results have shown the great potential of such a mission for environmental studies and call for the need to continue monitoring the Earth's gravity field variations (much like we continue monitoring the mean sea level by altimetry). Considering that the GRACE mission may soon come to its end, the main issue is – and will be at each time a new gravity satellite system is flown – to decide on the follow-on mission as quickly as possible so as to avoid any discontinuity in the data series.

Bibliography

Balmino, G., Reigber, C., and Moynot, B., 1976a. A Geopotential Model Determined from Recent Satellite Observation Campaigns (GRIM1). *Manuscripta Geodaetica*, **1**, 4169, Berlin.

Balmino, G., Reigber, C., and Moynot, B., 1976b. *The GRIM2 Earth Gravity Field*. Deutsche Geodätische Kommission, Reihe A, Heft Nr. 86, München.

Biancale, R., Balmino, G., Lemoine, J. M., Marty, J. C., Moynot, B., Barlier, F., Exertier, P., Laurain, O., Gegout, P., Schwintzer, P., Reigber, C., Bode, A., Gruber, T., König, R., Massmann, F. H., Raimondo, J. C., Schmidt, R., and Zhu, S. Y., 2000. A new global Earth's gravity field model from satellite orbit perturbations: GRIM5-S1. *Geophysical Research Letters*, **27**(22), 3611–3614.

ESA (European Space Agency) 1999. Gravity Field and Steady-State Ocean Circulation Explorer Mission (GOCE). *ESA-SP*, **1233**(1), 1–217.

Gaposchkin, E. M., and Lambeck, K., 1969. *Smithonian Standard Earth (II)*. Special Report No. 315, Cambridge/Mass.: Smithonian Astrophysical Observatory.

Gaposchkin, E. M., 1973. *Smithonian Standard Earth (III)*. Special Report No. 353, Cambridge/Mass.: Smithonian Astrophysical Observatory.

IERS (International Earth Rotation Service), 2004. IERS Conventions (2003). IERS Techn. Note 32, Frankfurt.

JPL (Jet propulsion Laboratory), 1998. *GRACE: Gravity Recovery and Climate Experiment*. Science and Mission Requirements Document, revision A, JPLD-15928, NASA's Earth System Science Pathfinder Program, 1–84.

Lerch, F. J., Wagner, C. A., Smith, D. E., Sandson, M. L., Brownd, J. E., and Richardson, J. A., 1972a. *Gravitational Field Models for the Earth (GEM1&2); Report X55372146*. Greenbelt/Maryland: Goddard Space Flight Center.

Lerch, F. J., Wagner, C. A., Putney, M. L., Sandson, M. L., Brownd, J. E., Richardson, J. A., and Taylor, W. A., 1972b. *Gravitational Field Models GEM3 and 4. Report X59272476*. Greenbelt/Maryland: Goddard Space Flight Center.

Lerch, F. J., Wagner, C. A., Richardson, J. A., and Brownd, J. E., 1974. *Goddard Earth Models (5 and 6). Report X92174145*. Greenbelt/Maryland: Goddard Space Flight Center.

Lerch, F. J., Klosko, S. M., Laubscher, R. E., and Wagner, C. A., 1979. Gravity Model Improvement Using Geos3 (GEM9 and 10). *Journal of Geophysical Research*, **84**(B8), 3897–3916.

Lundquist, C. A., and Veis, G., 1966. *Geodetic Parameters for a 1966 Smithonian Institution Standard Earth*. Special Report No. 200, Cambridge/Mass.: Smithonian Astrophysical Observatory.

Marsh, J. G., Lerch, F. J., Putney, B. H., Christodoulidis, D. C., Smith, D. E., Felsentreger, T. L., Sanchez, B. V., Klosko, S. M., Pavlis, E. C., Martin, T. V., Robbins, J. W., Williamson, R. G., Colombo, O. L., Rowlands, D. D., Eddy, W. F., Chandler, N. L.,

Rachlin, K. E., Patel, G. B., Bhati, S., and Chinn, D. S., 1988. A new gravitational model for the earth from satellite tracking data: GEMT1. *Journal of Geophysical Research*, **93**(B6), 6169–6215.
Mayer-Gürr, T., Eicker, A., and Ilk, K. H., 2006. ITG-GRACE02s: a GRACE gravity field derived from short arcs of the satellite's orbit. In *Proceedings of the First Symposium of International Gravity Field Service*, Istanbul.
Mayer-Gürr, T., Kurtenbach, E., and Eicker, A., 2010. ITG-Grace2010 Gravity Field Model; http://www.igg.uni-bonn.de/apmg/index.php?id=itg-grace2010.
Nerem, R. S., Lerch, F. J., Marshall, J. A., Pavlis, E. C., Putney, B. H., Tapley, B. D., Eanses, R. J., Ries, J. C., Schutz, B. E., Shum, C. K., Watkins, M. M., Klosko, S. M., Chan, J. C., Luthcke, S. B., Patel, G. B., Pavlis, N. K., Williamson, R. G., Rapp, R. H., Biancale, R., and Nouel, F., 1994. Gravity model developments for topex/poseidon: joint gravity models 1 and 2. *Journal of Geophysical Research*, **99**(C12), 24421–24447.
Nerem, R. S., 1995. Terrestrial and planetary gravity field. *Reviews of Geophysics*, suppl., 469–476.
Reigber, C., Balmino, G., Moynot, B., and Müller, H., 1983. The GRIM3 Earth Gravity Field Model. *Manuscripta Geodaetica*, **8**, 93–138, Stuttgart.
Reigber, Ch., Bock, R., Forste, Ch., Grunwaldt, L., Jakowski, N., Lühr, H., Schwintzer, P., and Tilgner, C., 1996. CHAMP Phase B executive summary. G.F.Z., STR96/13, 1–37.
Reigber, C., Balmino, G., Schwintzer, P., Biancale, R., Bode, A., Lemoine, J.-M., König, R., Loyer, S., Neumayer, H., Marty, J.-C., Barthelmes, F., Perosanz, F., and Zhu, S. Y., 2002. A high quality global gravity field model from CHAMP GPS tracking data and Accelerometry (EIGEN-1S). *Geophysical Research Letters*, **29**(14), 10.1029/2002GL015064.
Reigber, C., Balmino, G., Schwintzer, P., Biancale, R., Bode, A., Lemoine, J. M., König, R., Loyer, S., Neumayer, H., Marty, J. C., Barthelmes, F., Perosanz, F., and Zhu, S. Y., 2003a. Global gravity field recovery using solely GPS tracking and accelerometer data from CHAMP. *Space Science Reviews*, **29**, 55–66.
Reigber, C., Schwintzer, P., Neumayer, K. H., Barthelmes, F., König, R., Förste, C., Balmino, G., Biancale, R., Lemoine, J.-M., Loyer, S., Bruinsma, S., Perosanz, F., and Fayard, T., 2003b. The CHAMP-only earth gravity field model EIGEN-2. *Advances in Space Research*, **31**(8), 1883–1888, doi:10.1016/S0273–1177(03)00162-5.
Reigber, C., Schmidt, R., Flechtner, F., König, R., Meyer, U., Neumayer, K. H., Schwintzer, P., and Zhu, S. Y., 2003c. First EIGEN Gravity Field Model based on GRACE Mission Data Only. (http://op.gfz-potsdam.de/grace/results/grav/g001_eigen-grace01s.html).
Reigber, C., Schmidt, R., Flechtner, F., König, R., Meyer, U., Neumayer, K. H., Schwintzer, P., and Zhu, S. Y., 2005. An Earth gravity field model complete to degree and order 150 from GRACE: EIGEN-GRACE02S. *Journal of Geodynamics*, **39**, 1–10.
Schwintzer, P., Reigber, C., Bode, A., Kang, Z., Zhu, S.Y., Massmann, F. H., Raimondo, J. C., Biancale, R., Balmino, G., Lemoine, J. M., Moynot, B., Marty, J. C., Barlier, F., and Boudon, Y., 1997. Long wavelength global gravity field models: GRIM4S4, GRIM4C4. *Journal of Geodesy*, **71**(4), 189–208.
Tapley, B., Ries, J., Bettadpur, S., Chambers, D., Cheng, M., Condi, F., and Poole, S., 2007. The GGM03 Mean Earth Gravity Model from GRACE. Eos Trans. AGU 88(52), Fall Meet. Suppl., Abstract G42A-03.
Wagner, C. A., Lerch, F. J., Brownd, J. E., and Richardson, J. E., 1976. Improvement in the Geopotential Derived from Satellite and Surface Data – GEM 7 and GEM 8. Report X9217620, Greenbelt/Maryland.

Cross-references

Gravimeters
Gravity Anomalies, Interpretation
Gravity Data, Advanced Processing
Gravity Data, Regional – Residual Separation
Gravity Measurements, Absolute
Gravity Method, Airborne
Gravity Method, Principles
Gravity, Data to Anomalies
Gravity, Gradiometry

GRAVITY METHOD, SURFACE

Dinesh Chandra Mishra, Virendra M. Tiwari
Gravity and Magnetic Studies Group, National Geophysical Research Institute (CSIR), Hyderabad, Andra Pradesh, India

Definition

Gravity method. Geophysical method based on measuring the variations in the earth's gravity field.
Gravity stations. Locations, where gravity measurements are made.
Base gravity stations. Locations, where absolute value of the earth's gravity field is known.
Primary base. Gravity observation stations, which are made on stable platform and absolute value of the gravity field is known to a greater accuracy.
Secondary bases. Temporary bases established for a particular survey.
Closed loop. Bases that are connected in a loop for base loop correction.
Drift correction. Correction applied for nontidal variation of gravity, which is mainly dependent on the behavior of gravitymeters with time.
EOTVOS correction. Correction applied to marine gravity data for the movement of ship.
mGal; 1 mGal = 10^{-5} m/s^2
μGal; 1 μGal = 10^{-8} m/s^2

Introduction

Gravity method in geophysics is employed to determine subsurface density distribution by measuring the variation of the earth's gravity field. Measurements of gravity field are usually made on the earth or ocean surface. The scale of investigations (accuracy and spatial distance of gravity measurements) depends on the type of geological problems being investigated. These vary from a few μGal to mGal of measurements accuracy, which is a few millionth of average earth's gravity field (9.8 m/s^2) and station intervals of a few meters to a few kilometers. Based on applications, gravity survey can be broadly classified in the following three groups:

1. Regional gravity surveys: Station spacing is a few kilometers and survey accuracy is of about 1 mGal. It is generally used for tectonic and geodynamic studies.

2. Reconnaissance gravity surveys: Station spacing is a kilometer or so and survey accuracy is of 0.1–0.5 mGal. Such surveys are normally planned for delineating promising zones for hydrocarbon and mineral exploration, tectonic and lithological mapping.
3. Detailed gravity surveys: Mainly done to demarcate local geological structures favorable for mineral deposits, engineering applications, and groundwater studies. Survey accuracy is in the order of ~0.01 mGal and station interval is in the order of a few meters.

Measuring equipments

Gravity measurements

There are primarily three methods of measuring earth's gravity field:

(a) A method, in which period of pendulum is used to find absolute value of gravity
(b) Method based on measuring velocity of free-falling body in the vacuum chamber that also provides absolute value of gravity
(c) Sensitive spring balance, in which change in the spring length offers variation in the earth's gravity. They are relative measurements.

The spring balances are relative instruments, however they are main source of gravity data recording. They measure the differences in the gravity field between two or more points. Worden gravity meter is one such gravimeter, which is employed extensively in several countries. Scintrex gravity meters are now only commercially available gravity meters, which have least count of 1 μGal. Similar kind of gravity meters are also used on the ship, however they are specially designed to take care of vertical and horizontal accelerations with gyroscopic adjustments.

Elevation and position measurements

Gravity field of the earth increases as we move toward poles due to flattening and centrifugal force. The factor that affects the earth's gravity field most is the distance from center of the earth (R). Therefore, the elevation of stations and its position should be precisely known for gravity surveys. There are several ways to know the position and elevation but nowadays Global Positioning System (GPS) provides a best method of determining position and elevation by utilizing information from several orbiting satellites. Sometimes, when measurements are made in shadow regions particularly in the forests, the GPS navigations are not very good. In such cases, theodolite, navigation based on maps, and elevation using pressure measurement devices are often used as per the requirements of survey accuracy. GPS measurements are improved by using them in differential GPS mode (DGPS). DGPS navigation systems that provide position, speed, and heading information of ship are common now for shipborne gravity measurements.

Land gravity survey

Gravity survey often refers to the measurement of the earth's gravity field along with the position and elevation of the observation point. Elevation effect is approximately 0.2–0.3 mGal/m implying that an error of 1 m in elevation of the stations will affect the gravity measurement by about 0.2 mGal which is quite significant, especially when the accuracy of gravimeters is 0.01–0.001 mGal. Therefore, ample efforts are made to record elevation of the stations of the desired accuracy for a particular survey. In general, gravity measurements are made over a plain area a few meters away from depression, ponds, mound, etc., which can affect the gravity readings.

Preprocessing of field data

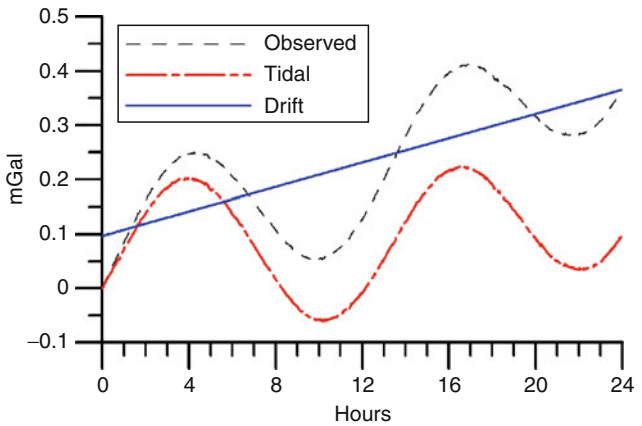
Gravimeters are highly sensitive and they are affected by the changes caused due to the mechanical and environmental changes such as jerks, pressure, and temperature. Gravimeters measure variations in the gravity field, thus the measurements using them are required to tie with the base stations where the absolute values of the gravity field are known. Therefore, measurements need to be corrected for factors influencing measurements to achieve the needed accuracy. There are mainly two corrections, namely, (1) drift correction and (2) base loop correction, which are applied during course of the field work.

Drift correction

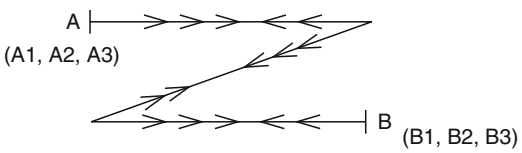
When the repeat observations are made by a gravimeter at the same location, it is found to be varying with time. The variation in the gravity field with time is caused due to several geophysical phenomena (Tiwari and Hinderer, 2011). In a gravity survey, time-varying gravity field is primarily attributed to the tidal effect and nontidal changes, primarily caused due to the mechanical behavior of the gravimeter and contributions from atmospheric or hydrological effects during few hours of repeat observations. The peak to peak tidal effect is about 300 μGal during new moon day and can be precisely calculated for specific locations using standard available softwares. The nontidal variation is referred as drift of the gravimeter (Figure 1). This is normally ~0.0–0.03 mGal/h and is corrected after removing tidal effects computed for that specific location. The drift correction is estimated by making repeat observations at the interval of 1–2 h and the difference in the two nontidal gravity values, known as drift of the instrument, is distributed proportionally to the other observations made during that period. Figure 1 shows the variation of the gravity field with time, tidal effect and drift at Hyderabad. Repeat observations at a fixed location are also important to understand the condition of gravimeters. If drift of the gravimeter is more than a permissible limit of <0.05 mGal/h, the instrument needs to be thoroughly checked for scale adjustments, battery, etc., or survey needs to be redone.

Base loop correction

Gravity base stations are the points where the absolute value of the gravity field is known to an accuracy of 0.01 mGal (10 μGal) and are used to obtain the absolute value of the gravity field at the stations in the field by measuring its variations with respect to base stations. There are primary base stations in the country usually at fixed locations such as railway stations, airports, etc. In any gravity survey, a network of secondary bases is established encompassing survey block, which is connected to any of the known primary base. To achieve the accuracy, a particular procedure of measurements is followed. The bases are connected by looping three times between two bases A and B as A-B-A-B-A-B (A and B being two stations, Figure 2) and occupying them usually within an hour to ensure that the instrument drift is linear and minimum. This measurement procedure is referred to as double tie of bases. This kind of looping provides three observations at every base, which are corrected for the drift of the gravimeter based on consecutive five readings. The consistent value from the three observations or the average of close-by readings if they are not consistent is chosen for every base station. In case of large survey area such as in a regional survey or hydrocarbon exploration, bases are first established in the region that are subsequently used to tie the gravity stations for their absolute values and simultaneously used for drift correction so as to avoid reoccupying same base after 2 h. In these cases, the survey starts from a base station and closed at some other base station after every 1–2 h. The difference in the gravity field is adjusted with respect to their absolute values distributed to the station occupied during that period.



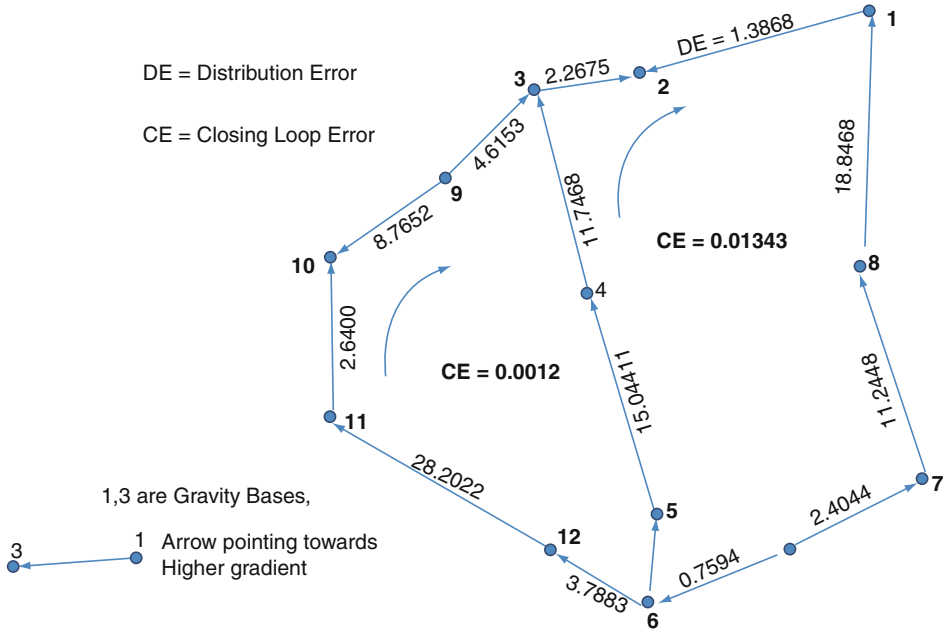
Gravity Method, Surface, Figure 1 Observations of gravity changes at Hyderabad for a day. Tidal and drift of the instrument are plotted separately.



Gravity Method, Surface, Figure 2 A and B are two bases that are occupied five times starting from A to B and going back and forth that provide three diurnal corrected readings at each station which are used to obtain correct reading at base stations.

The bases are established on permanent structures for easy identification during the survey and afterward. Figure 3 shows a typical example of a network of bases, which shows the differences in the gravity values between two bases along the arms with arrows showing toward higher values. These differences along the arms of a network (loop) are added or subtracted as the case may be depending on the direction of the arrows to obtain the difference at closing base of the loop. Ideally, it should be zero but in all practical surveys a difference is found, which is known as closed loop error (CE). This difference is divided by number of arms that gives the distribution error (DE). The arrow inside the loop points in the direction of the positive loop error. In this case, station number 1 is a primary base that is used to establish secondary bases in the survey block for ease of recording the gravity readings in the survey block. Initial checks for loop closure errors are done in the field itself in order to repeat the erroneous bases in case the loop error is found to be more than the permissible error. The permissible error for the closed loop is $0.02\ \sqrt{n}$ mGal, where n is number of arms in a loop. Sazhina and Grushiny (1971) have provided detailed description of these corrections under "Adjustment of Gravity Networks" similar to geodetic networks. Having established the requisite number of base stations in an area, the gravity observations at various stations are tied to these bases within a maximum time limit of 1–2 h for drift correction during that period. In regional surveys where survey is conducted along roads at station spacing of 3–5 km, it may not be possible to make loops. However, as the desired accuracy of the gravity data in this survey is limited to 1 mGal, bases can be tied along the roads at about 20–30 km interval as A-B-A-B which provides two drift corrected readings at every base. The average of the two readings is taken as the base value. In case of mineral exploration, if the area is small and isolated far away from any known base, an arbitrary base is chosen near the camp with any arbitrary value, for example, 0, 100, 1,000, etc., and all the readings are reduced to this arbitrary base value. This provides the gravity anomalies in the survey block with reference to the arbitrary base. In case it is desired, this arbitrary base can be tied to standard bases in the nearby region even after the survey and observed gravity values at stations are updated with reference to the absolute value of the gravity field at the base station accounting for the arbitrary value adopted previously for the survey.

There are several other corrections like latitude, terrain, elevation corrections, etc., that are applied to the observed gravity field at a station to obtain Bouguer anomaly or simply gravity anomaly at a station (Lafehr, 1991; Gotze and Li, 1996). These corrections are dealt under data reduction or preparation of gravity anomaly maps.



Gravity Method, Surface, Figure 3 Closed loops of gravity base stations for loop correction. Station numbers are marked as filled circle with arrows indicating toward the increasing values. The closing loop error (CE) is the departure of vector algebraic sum of the each arm (differences in the observed gravity field at adjoining bases) from zero. Observed gravity difference between two adjoining stations (values of arm) is written on the corresponding arm and CE is also written for each loop. Distribution error (DE) is CE divided by number of arms in the loop, which is added to each loop to arrive final values.

Marine gravity survey

The fundamental difference of marine gravity survey from the land survey is measurements on moving platform, which distorts the measurement accuracy. However, if the speed and accelerations are measured, the corresponding corrections might be applied to the gravity data. Nevertheless, the measurement accuracy on the moving platform like ship will not be as accurate as measurements made on ground. The standard accuracy of marine gravity survey is an order of magnitude less than ground survey. The sea state pose another difficulty in the marine survey, it is apparently calm in the deep water and more predictable than the shallow sea or near the coast. Thus, marine gravity measurements are generally corrected for effects of moving platform, earth's tide, drift, and adjusted for crossover tracks as briefly explained below.

Moving platform correction

The gravity field measurements are distorted due to movement of ship, as well as due to horizontal and vertical accelerations, which can be corrected and eliminated by using supplementary data. They are classified in three categories (Dehlinger, 1978).

EOTVOS corrections: When a ship moves, particularly in the east west direction, it experiences centripetal accelerations about the earth's axis that modifies the acceleration due to gravity. This can be precisely corrected if the speed and heading of ship is known.

Corrections for horizontal and vertical accelerations: These corrections are applied to gravity measurements for any horizontal and vertical motion of the platform. In the modern type of gravimeter, they are adjusted by designing stabilized platform or gimbal-suspended systems. They are also removed using GPS data and then averaging of several gravity values.

Corrections for cross coupling accelerations: This correction is applied to a beam type of gravity measuring instruments when it is rigidly attached to stabilized platform. The error occurs when horizontal and vertical components of ship accelerations have equal periods and have a phase difference.

Earth tide correction

As discussed above, earth and ocean are subject to the gravitational attraction of Sun and Moon. Over the land, major part of tidal effect is perturbation of solid earth; however, over ocean the perturbation of outer water surface also has pronounced effect. Therefore, tidal correction applied to shipborne gravity data should have a model of both solid earth and ocean.

Drift correction

All the marine gravity surveys start with known absolute gravity value at the port of sailing and closed to the known gravity station at the end of survey. Since the time of reoccupation of known gravity bases (reference stations) are generally weeks to month, the drift computed are not

very accurate. Nevertheless, all the spring associated gravity meters also have long time drift that are generally taken into account while closing the marine survey though they are often small.

Network or crossover adjustments

The problem of offset in the marine gravity survey is quite known. It arises due to several reasons (Wessel and Watts, 1988). One of the main reasons of offset is mechanical tare or jar of the spring type of gravity meters. Cross over corrections in the same Leg of the marine survey is similar to close loop correction discussed for land survey. In case of marine survey over a block, a cross cutting profile is recorded that is used for instrumental check, drift, and data accuracy.

Design of survey

Designing of gravity survey is a vital component to optimize the resources according to the geological problems. Following are the details.

Regional surveys for geodynamic studies

As the name indicates these surveys are basically regional in nature and are conducted in large regions for geodynamic studies especially for crustal and upper mantle structures. These surveys are carried out along roads and tracks at a station spacing of 3–5 km, which delineates long wavelength anomalies related to large scale structures and density inhomogeneities. Anomalies recorded in these surveys are of the order of tens of mGal (10–100 mGal) spread over a few hundred km (~100–1,000 km). The accuracy of such surveys is limited to ±1–2 mGal that can be obtained by controlling elevation of stations using Differential Geographic Positioning Systems (differential GPS) with accuracy better than 1 m in elevation.

Hydrocarbon exploration

There are two aspects of gravity surveys for hydrocarbon exploration on land, namely, (a) reconnaissance gravity survey and (b) detailed/integrated gravity surveys (Nettleton, 1976).

Reconnaissance gravity survey

Gravity surveys for exploration of hydrocarbons are basically conducted as reconnaissance surveys to delineate basement structures and structures in sedimentary basins to delimit the areas for seismic surveys. It is generally, carried out in a large area covering a specific basin along roads and tracks at a station spacing of approximately 2.0 km. It is planned in such a way that network of stations at about 2.0 km spacing are available in the entire region. The order of gravity anomalies expected from such structures are of a few mGal in amplitude and therefore, the desired accuracy of such surveys are ±0.1 mGal. Station elevation is therefore obtained by geodetic leveling or Differential GPS providing elevation to an accuracy of a few cm (<10 cm). To check the accuracy of recorded gravity value, 5–10% of gravity stations are repeated on different days and preferably by different observers. In ocean, surveys are conducted along profiles 4–5 km apart over a specific block and bathymetry is recorded with echo sounder.

Detailed/integrated gravity survey

While conducting seismic surveys for exploration of hydrocarbons, it is a general practice to follow the same profiles by gravity surveys at closer stations spacing of 0.5–1 km. These surveys are basically conducted for an integrated approach to supplement the information derived from seismic surveys and delineate structures within the sediment. The desired accuracy of such surveys are ±0.01 mGal (10 μGal) and therefore, station location and their elevation to accuracy of a few centimeter (<5 cm) are usually obtained from geodetic surveys or dual frequency differential GPS. In case of shallow gas occurrences, such surveys can be conducted over a grid with both profile and station spacing of 10–100 m as the case may be. At the places, where reflection seismic data is poor, gravity survey is conducted to supplement the seismic.

Surveys for minerals, groundwater, and engineering sites

The application of gravity survey for mineral and groundwater exploration primarily lies in delineating surface/shallow structures such as faults, fractures, etc., which may be mineralized (Parasnis, 1997). Same is the case with engineering sites where shallow structures such as faults, fractures, etc., are delineated to avoid those sections for large scale constructions such as dams, power plants, etc. The survey procedures for application to these studies are therefore, similar in logistics. They fall under two categories, namely, reconnaissance and detailed/integrated surveys.

Reconnaissance survey

This kind of survey is carried out to limit the area of interest from a larger area. It is conducted along profiles separated by a few hundred meters a part (100–500 m) at stations spacing of a few tens of meters (10–100 m). If the area is large and inaccessible to lay down specific profiles, it is conducted along available roads and motorable tracks at a station spacing of a few tens of meters (10–100 m). Due to desired accuracy of gravity data, elevation of stations in these surveys are obtained by geodetic leveling or differential GPS, providing elevation to an accuracy of a few centimeters (<10 cm). If the area is not large and accessible, a base line is laid out parallel to the general strike of the geological formation in the center of the block and profiles are laid out perpendicular to it at desired intervals. Stations are marked along the profiles at regular intervals by geodetic survey or using differential GPS. Based on these reconnaissance surveys, promising zones and geologically significant structures such as faults are delineated which are followed up by detailed gravity and other geophysical surveys such as electrical, electromagnetic, etc. In case of surveys for

engineering sites, same reconnaissance procedure is followed to delimit the area as described above for mineral exploration.

Detailed/integrated gravity survey
Detailed gravity survey for the mineral exploration is basically limited in areas of a few square meters and are carried out in grids with profiles and station spacing of a few meters (1–10 m) or may be even less depending on the size of the area and the target. A baseline is laid out parallel to the general strike of the geological formations, and profiles are laid perpendicular to it at almost same spacing of 1–10 m as per requirement of particular survey. Stations along the profiles are marked by the geodetic survey at the same interval as profiles forming a grid. In certain cases, the observed gravity anomaly may be less than 0.1 mGal, which require survey accuracy of about 0.01 mGal (10 μGal).

Summary

Gravity method, surface includes both onshore and offshore measurements of the gravity field of the earth for geodynamic studies and resource exploration. Accuracy of the gravity survey required for various purposes and precautions to be observed in this regard are discussed. Geodynamic studies to an accuracy of 1–2 mGal can be carried out using normal gravity survey but resource exploration usually requires high-resolution gravity survey to an accuracy of ~10 μGal using microGal gravimeters. In such cases, base loop correction in ground surveys and correction for crossover points in marine surveys are essential and are emphasized.

Bibliography

Dehlinger, P., 1978. *Marine Gravity*. Amsterdam: Elsevier.
Gotze, H. J., and Li, X., 1996. Topography and geoid effects on gravity anomalies in mountainous areas as inferred from the gravity field of the Central Andes. *Physics and Chemistry of the Earth*, **21**, 295–297.
Lafehr, T. R., 1991. Standardization in gravity reduction. *Geophysics*, **56**, 1170–1178.
Nettleton, L. L., 1976. *Gravity and Magnetics in Oil Exploration*. USA: McGraw-Hill.
Parasnis, D. S., 1997. *Principles of Applied Geophysics*. London: Chapman & Hall.
Sazhina, N., and Grushiny, N., 1971. *Gravity Prospecting*. Moscow: Mir Publishers.
Tiwari, V. M., and Hinderer, J., 2011. Time variation of earth's gravity field: surface measurements. In Gupta, H. (ed.), *Encyclopedia of Solid Earth Geophysics*. Berlin: Springer (this volume).
Wessel, P., and Watt, A. B., 1988. On the accuracy of marine gravity measurements. *Journal of Geophysical Research*, **94**, 7685–7729.

Cross-references

Earth Tides
GPS, Data Acquisition and Analysis
Gravity Field of the Earth
Gravity Field, Temporal Variations from Space Techniques
Gravity Field, Time Variations from Surface Measurements
Gravity Method, Principles
Gravity, Data to Anomalies

GRAVITY MODELING, THEORY AND COMPUTATION

Jean-Pierre Barriot, Lydie Sichoix
Observatoire Géodésique de Tahiti, University of French Polynesia, Faaa-Tahiti, French Polynesia

Definition

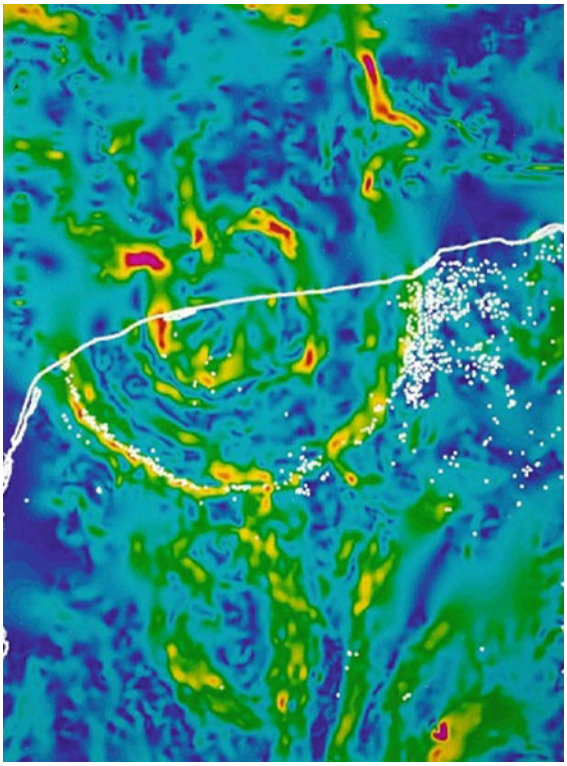
Gravity modeling. Explaining the variations of the external gravity field of a body in terms of its internal density variations. Applications range from oil prospecting to the deep structure of the Earth.
Inverse problem. In Gravity modeling, the forward problem is to compute the external gravity field of a body knowing its internal density distribution. The inverse problem is the opposite, and is much more difficult to solve for.

The gravity modeling from a theoretical point of view

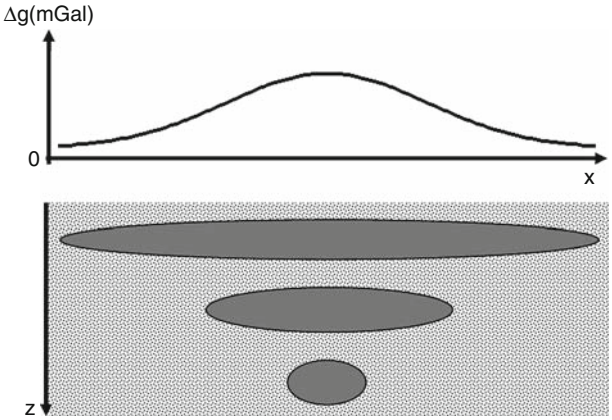
Gravity is an elusive quantity to deal with, but a powerful one. This is mainly because unlike electromagnetic forces, gravity cannot be shielded as there are no negative masses. We refer to LaFehr (1980) and Milsom (1996) for an introduction to the gravimetric techniques (see also *Gravity Method, Principles*), and Zidarov (1990) for the subtleties of gravimetric inversion. Good monographs are available on the theory and computational aspects of gravity modeling. The one from Blakely (1996) is particularly exhaustive (see also *Gravity Anomalies, Interpretation*). Here we give just a short introduction, with emphasis on the modeling directly done on the data, and not on its Fourier transforms.

A landmark success of the application of gravimetric interpretation is the identification of the Chicxulub crater buried underneath the Yucatan Peninsula (Hildebrand et al., 1991, Figure 1, see *Impact Craters on Earth*).

The reasons why gravity is so elusive are profound and lie deep in potential theory. Let us start with a simple example. The radial gravity outside a homogeneous sphere of mass M and radius r is given by GM/R^2, where G is the Newtonian constant and R the distance from the center of the sphere. But r does not intervene in the computation, as long as $R > r$. This means that two homogeneous spheres of radius r and r' have the same external field, as long as they have the same mass. This nonuniqueness has been studied in detail and theory shows that a 3D repartition of density that causes a gravity anomaly can be split in two additive parts: one that is harmonic, that is, that satisfies the Laplace equation, and another that is anharmonic. Only the harmonic part of the anomalous density can be unambiguously



Gravity Modeling, Theory and Computation, Figure 1 The Chicxulub structure (diameter 180 km) is widely believed to be the remnant of the impact that killed the dinosaurs 65 millions of years ago. This structure was identified on a gravity map of the coast, with no other visible evidence. The impacting asteroid that formed the crater was at least 10 km in diameter (Image source: Shuttle radar topography mission, Chicxulub Crater Gravity Map with Cenotes Shown as White Dots with the Yucatan Coastline Being the White Line).



Gravity Modeling, Theory and Computation, Figure 2 Each one of these underground bodies (*gray*) creates the same gravity anomaly. If we suppose that the body is reduced to a point mass, one can infer from this assumption the maximum possible depth for the causative body (Image source: Dubois and Diament, 2005, redrawn).

reconstructed from the anomalous gravity signal (Moritz, 1989; Vanicek and Kleusberg, 1985; Ballani et al., 1993). The anharmonic part of the density belongs to the so-called null-space of the model. In terms of geophysics, a typical example of nonuniqueness is shown in Figure 2.

Far more useful is to note that the ambiguity in gravity modeling can be reduced, or even zeroed if we know the support, in mathematical terms, that is, the spatial extension, of the density anomaly. For example, if we know by other means the radius r of the homogeneous sphere of paragraph one, the density ρ is automatically fixed, with $g = GM/R^2$ by $M = 4/3 \times \pi \times \rho \times r^3$. Chao (2005) has extended this reasoning to the modeling of gravity in terms of spherical harmonics (see *Spherical Harmonic Analysis Applied to Potential Fields*). If we consider a 2D spherical skin covering the Earth, the inverse solution of the surface density distribution is unique. The latter applies quite readily in the inversion of time-variable gravity signals originating from within the crust (such as those observed by the GRACE/GOCE space missions, see *Gravity Method, Satellite*).

Another possibility is to suppose the density known a priori. In that case, the radius of the sphere of paragraph one instantly follows. This seems trivial, but it is not. It has been demonstrated that the gravity field outside a body can be totally explained by a combination of point masses, linear masses, and sheet masses that lie deep inside the body (Zhdanov, 1988). But here also, this result is of little help for the geophysicist. Moreover, the inverse gravity problem is unstable: small variations in the data (errors) can cause unbounded variations in the solution. Technically, the inverse gravity problem is ill-posed in Hadamard's sense (1902, see also Sabatier, 2000 and Menke, 1989, and Inverse Theory, Global Optimization).

The core fact that permits a useful application of gravity anomalies in geophysics is that these anomalies, even contaminated by noise, reflect *real* structures, that is, structures that *we know* have a reasonable density contrast with the surroundings, and/or exhibit a density that varies smoothly inside the expected structure – except at the edges or along faults – and/or are located inside a given volume, for example, we know from borehole measurements that the iron ore is located between two depths: one at −1,000 m and the other one at −2,000 m. To have bounds on the possible densities and/or the locations reduce tremendously the world of possibilities. In short, we are then looking for *plausible* structures.

With such cautions and restrictions, gravity modeling is a powerful tool if correctly used. The best use of gravity measurements is to team them up with other measurements with different null-spaces, like magnetic surveys or even seismic imaging. For example, the velocity of seismic waves increases with density, and the anomalous delays of propagation are linked to the anomalous velocity field (and so to the density through an equation of state) by an averaging integral along the ray known as the Radon

transform (Troyan and Hayakawa, 2002). A joint seismic/gravity inversion is a powerful tool and probably the best way to remove nonuniqueness (Strykowski, 1998, see also *Seismic Tomography*).

But this is hardly done in practice. The driver for field measurements is cost. A gravity survey, even on remote areas, is cheap, and often done as a precursor to more expensive and accurate methods, like seismic tomography. Gravity modeling must thus be able to identify promising structures on its own. The game is tricky, as the observed gravity is of course affected by noise, but also by data reductions (see *Gravity, Data to Anomalies*) that ensure that the gravity data is corrected for drifts, referred to some "constant" reference in altitude (usually the geoid), and is corrected from the gravity signature of the topography (Bouguer correction, see *Gravimeters*). Besides, the gravity signal is by nature nonlocal, and sums up the signature of distant bodies or interfaces, like the discontinuity of the Mohorovic boundary, for example. These distant signatures must be guessed and withdrawn from the gravity signal *before* any local geological interpretation is undertaken (see *Gravity Data, Regional – Residual Separation*).

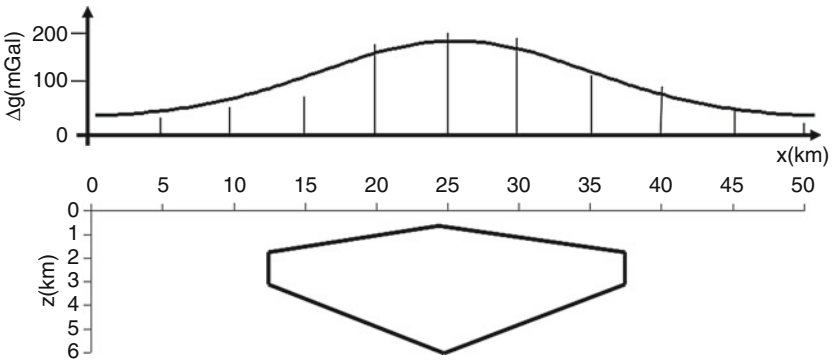
There are two ways to have the job done. One is to use direct modeling with feedback. Technically, an operator draws a 2D *educated* guess of the plausible structure see (Figure 3) on a tablet or other graphic device, and the computer program derives the anomalous gravity model caused by the structure and draws it against the real signal. The operator then modifies slightly the guessed structure in order to obtain a better match with the real signal. For the geophysicist, "anomalous gravity" is synonymous with "Bouguer anomaly," that is to say a measurement of gravity corrected for the variation of altitude and for the topography-induced variations. Other definitions of "anomalous gravity" are possible. For example, it is theoretically possible to work indifferently with geoid undulations or Bouguer anomalies. In practice, the modeling of geoid undulations is reserved for very-large-scale structures in isostatic equilibrium (see *Isostasy*), while Bouguer anomalies are reserved for small- and medium-scale structures (see *Gravity, Data to Anomalies* and *Geoid Undulation, Interpretation*).

This process can be automated with the program adjusting step by step the shape of the structure to exactly match the observed gravity anomaly. With a little bit more complexity in terms of calculus, 3D modeling can be done in almost the same manner, with polyhedra replacing polygons. But there is a huge difference at another level. A human operator has always trouble to visualize and modify 3D solid models, even with the assistance of powerful 3D graphic representations, and such a tedious tool is used with parsimony. Efficient formula that replaces volume integrals by integrals along vertices have been developed, from Talwani et al. (1959), to Tsoulis and Petrovi (2001) to implement polygon/polyhedron calculus.

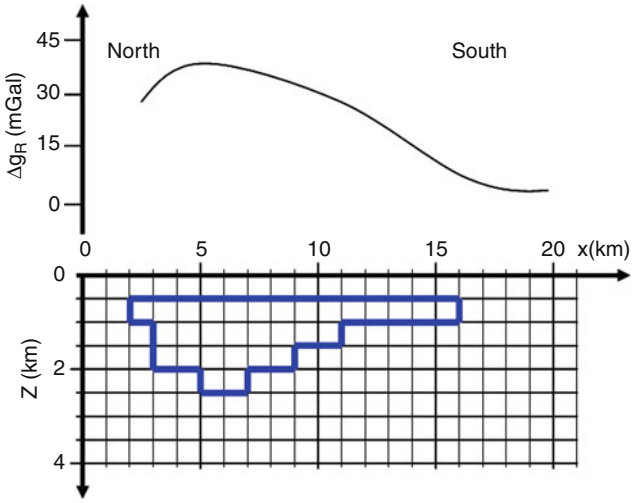
For 3D modeling, a better approach is to discretize the volume where we are seeking for the geological structure in tiny adjacent cells of simple fixed shape (e.g., cubic) but of unknown densities. It is important to note the difference with the previous approach: by looking at the densities, we obtain a mathematically linear problem, in the sense that the effect of each cell sums up additively in the gravity model. To the contrary, altering the skeleton of a polyhedron of constant density (even only a single corner point) has a nonadditive effect on the computed gravity. Technically, we are dealing with a nonlinear model. Nonlinear problems are intrinsically much more difficult to solve for.

The linear inverse gravimetric problem

Let us focus on the linear gravimetric inverse problem, illustrated in Figure 4. We have to define how the discrepancies between measurements and computed gravity from the model are accounted for. The usual way is to measure this discrepancy as a Euclidean distance, by summing up the square of the differences per data point between measurements and computed gravity and then taking the square root of the sum. This is called the



Gravity Modeling, Theory and Computation, Figure 3 The geological body is approximated by a polygonal structure of infinite length perpendicular to the paper sheet and of constant density contrast with the surroundings. The game is to move the summits of the polygon in order to match (*solid curve*) the observed gravity (*needles*).



Gravity Modeling, Theory and Computation, Figure 4 The linear inverse gravimetric problem. The subsurface is divided in small squares (2D) or small cubes (3D), and we are looking for the anomalous densities for each *square* or *cube* (Image source: Blakely, 1996, redrawn).

L2 norm of discrepancies. Another way is to sum up the absolute values of the discrepancies. This is called the L1 norm of discrepancies.

The set of all unknown densities x over the mesh and the set g of the gravity measurements on the ground are linked through the matrix relation $Ax = g$. In the frame of the L2 norm, to solve for the ill-posed inverse linear gravimetric problem can be cast as finding a set u of densities that verifies at the best this relation, and fulfilling at the same time another relation of the form $Bx = k$, where k is another physical or mathematical quantity. For example k can be magnetic measurements and B a design matrix linking densities to magnetic measurements. If we add in such way "sufficient" physical information, the solution of this augmented problem becomes unique. Very often, additional physical information is lacking, and uniqueness is imposed by setting $B = I$ and $k = 0$, where I is the identity matrix. This is equivalent to looking for a set u of densities that realizes a compromise between $Ax = g$ and $x = 0$, that is, an x that explains the gravity signal and is not too large (i.e., in the neighboring of zero). This is known in the literature as a Tikonov–Arsenine (1976) regularized inverse or Marquardt-Levenberg inverse (Levenberg, 1944; Marquardt, 1963). Technically, the solution depends also on the way that the hidden Euclidean spaces used to set up the problem are defined. They are not necessarily orthonormal spaces. The angular separation of their axes and unit lengths can carry additional information about the physical problem (called in the literature as "a priori weights"). Tarantola (2005) gives a complete survey of the probabilistic setting for the Tikonov–Arsenine formula, with the weights seen as the inverses of a priori covariance matrices. The covariance linked to I is especially important as it controls to which extent the density contrasts found are "reasonable." A density contrast of 0.2 is sound; a density contrast of 20 is not, even if the computed gravity matches more closely the observed signal!

The simplest form of the Tikonov–Arsenine inverse formula is of the form:

$$x^*(p) = \text{inv}(\text{trans}(A)A + pI)\text{trans}(A)g$$

where $x^*(p)$ is the solution of the inverse problem (here densities over the mesh), A is the design matrix defined above, p is a scalar parameter, I the identity matrix, and g the set of gravity measurements. In the Tarantola–Valette approach, p is the ratio between the variance of the noise error and the "variance" (the size) of the expected solution wrt zero. The scalar p is the "smoothing" parameter. The geophysicist chooses "by hand" p in order to have both "reasonable" L2 residuals, defined as $\text{trans}(Ax^* - g)(Ax^* - g)$, and a reasonable L2 size (densities) of x, defined as $\text{trans}(x^*)x^*$.

This solution x^* minimizes, with a unique minimum, the following functional (or cost function) $Q(x)$ with respect to x:

$$Q(x) = \text{trans}(Ax - g)(Ax - g) + p\ \text{trans}(x)x$$

Without doubt, the celebrated Tikonov–Arsenine formula is the workhorse of gravity inversion.

Another approach, also very popular, but unimplementable for large problems, is to look at the so-called singular value decomposition of A, a generalization of the eigenvalue decomposition of square matrices (Campbell and Meyer, 1979, see *Inverse Theory, Singular Value Decomposition*).

Looking at a solution of the inverse problem in terms of the L1 norm seems just a little adjustment with respect to the L2 norm. However, it is a totally different approach. Let us imagine that we have a set of n gravity measurements and a mesh of m cells, assuming that n is larger than m. Technically, the problem is overdetermined, with a matrix A with far more lines than columns. To solve for this system in the L1 sense can be casted as solving with respect to x each subset of m lines of A (each subsystem is now square with a trivial inverse), and looking at the solution that gives the best L1 residuals with respect to the complete matrix A. There are $n!/m!/(n-m)!$ subsystems. For 1,000 measurements and 50 cells, a computer crushing numbers from the dawn of the universe will have trouble to handle this inversion. Thankfully, Dantzig and colleagues (see Branham, 1990) found a way to carry out the L1 inversion with on-the-shelf computers, called the simplex algorithm (linear programming, see *Inverse Theory, Linear*). This type of inversion of gravity data has also to be supplemented with a priori (external) constraints to obtain a geologically "sound" solution, for example, by imposing that the inertia moments of the repartition of densities are minima (Cuer and Bayer, 1980), or by looking at extrema solutions, for example,

the "ideal" body of Parker (1975, see also Goodacre, 2006), that is, the set of densities characterizing the geological structure that gives a minimum density contrast with the surroundings.

The nonlinear inverse gravimetric problem

Let us return to the polygon/polyhedron modeling of gravity anomalies, where we assume that the density of the body is uniform.

Mathematically, the problem can be cast as $g = f(x)$, where now x is the set of coordinates of the corner points, and is nonlinear. We want to find a set of coordinates that explain "at the best" the gravity signal, that is, g(observed) $- f(x)$ is kept to a minimum, for example, with respect to the L2 norm.

We then have to minimize a cost function of the form:

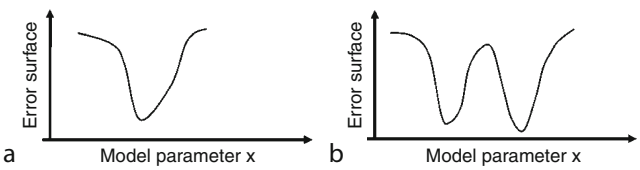
$$Q(x) = \text{trans}(g - f(x))(g - f(x))$$

There is a big difference with the linear case. We can have several well-separated, inherent minima to this functional (Figure 5). There are two ways to solve for this inverse problem. The first one is to linearize the problem, that is, to look at a solution close to a known estimate x_0, that is, by writing that $x = x_0 + dx$, with a dx bounded in some way.

A typical illustration in gravimetry is to look at the shape of an unknown interface between two strata of different and constant densities see (Figure 6). This is an extension of the polygonal model where some points are pushed to infinity and has been discussed by Tarantola and Valette (1982) in their milestone paper as an extension of the pioneering work of Cordell and Henderson (1968). This is often used to refine a model obtained by the modeler from a priori geological evidence, but can be hardly used without any a priori knowledge of the solution. The Backus and Gilbert approach (1968), very popular during the 1970s, can also be seen as a nonlinear problem.

The other approach, more radical, consists in trying to explore the whole set of possible minima of the nonlinear cost-functional. A popular mechanism to do so is to mimic the biological process of evolution. The genetic algorithm of Holland (1975) can be summarized in four basic steps: coding, selection, crossover, and mutation.

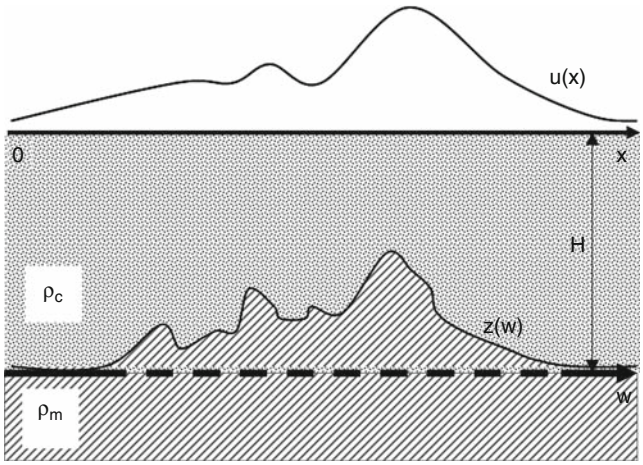
Common to any genetic algorithm is the digitization of an a priori and sufficiently populated list of model parameters using a binary code scheme (for example, we know that the density we are looking for is bounded by 2.5 and 2.8. Assume that the desired resolution (discretization) is 0.01. Then we can make a one-to-one correspondence between a binary word of 5 bits (a "chromosome") and this discretization. A subset of models is then selected from the model population by using some global fitness criterion, for example, the ratio of each model fitness (evaluated in the L2 sense with respect to the observed gravity and using the one-to-one correspondence) to the sum of all fitnesses for all the models. We then take the models the ratio of which is over a given threshold. For example, for a population of L models, we choose a subset of L/2 of "best fit" models. We then mate the L/2 models by pairs, each pair bearing two offsprings by recombination of their chromosomes by processes called crossover and mutation that mimic the meiosis of real chromosomes, giving birth to a complete new generation of L models that replaces the old one. Generally, the average fitness increases between generations, since only the best models from each generation are selected for breeding. Other algorithms exist to locate the "best" minimum, like the tunneling algorithm (Levy and Montalvo, 1985) or the particle swarm optimization (Monteiro Santos, 2010). See also in this Encyclopedia *Inverse Theory, Artificial Neural Networks* and *Inverse Theory, Monte Carlo Method*.



Gravity Modeling, Theory and Computation, Figure 5 (**a**) The cost function of the linear inverse problem presents only one minimum. (**b**) The cost function of the nonlinear inverse problem may present several minima, indicating a nonunique solution (Image source: Dimri, 1992, redrawn).

Spectral approaches

Spectral approaches rely on the expression of the gravity along a basis of given functions: very often a Fourier decomposition in the planar case or spherical harmonics on a global basis. New basis are gaining a wide acceptance, like wavelets, that are like "sine" or "cosine" waves but with a limited spatial support (Chui, 1992). Large



Gravity Modeling, Theory and Computation, Figure 6 Modeling an interface between two homogeneous separate strata. The problem is a special case of the polygon/polyhedron inverse problem (Image source: Tarantola and Valette, 1982, redrawn).

classes of inverse gravimetric problems can then be recast as deconvolution problems (Dimri, 1992), with all the artillery available from Fourier analysis (Dean, 1958; Blakely, 1996). Nevertheless, this approach suffers from all the drawbacks of the Fourier analysis, especially the need to smooth the discontinuities of the gravity signal at the edges of the zone of interest (this is of course not the case for spherical harmonics). Such approaches have been used in planetary geophysics to estimate the global and local average depth of the Moho interface on Mars (Zuber et al., 2000) from models of the gravity field and of the topography of the planets in spherical harmonics, without any other surface or subsurface information.

Summary

Gravity modeling is a cheap tool in geophysical prospecting. It can also be a very potent tool if used with some precautions. This always implies the introduction, both in forward and inverse modeling, of some a priori knowledge of the shape of structures that are guessed, or some bounds about their dimensions or about acceptable density contrasts.

Bibliography

Backus, G. E., and Gilbert, J. F., 1968. The resolving power of gross earth data. *Geophysical Journal of the Royal Astronomical Society*, **16**, 169–205.

Ballani, L., Engelis, J., and Grafarend, G., 1993. Global base functions for the mass density in the interior of a massive body (Earth). *Manuscripta Geodaetica*, **18**(2), 99–114.

Blakely, R. J., 1996. *Potential Theory in Gravity and Magnetic Applications*. Cambridge: Cambridge University Press. ISBN 0-521-57547-8.

Branham, R. L., Jr., 1990. *Scientific Data Analysis: An Introduction to Overdetermined Systems*. New York: Springer. ISBN 0-387-97201-3.

Campbell, S. L., and Meyer, C. D., Jr., 1979. *Generalized Inverses of Linear Transformations, Surveys and Reference Works in Mathematics 4*. Boston: Pitman. ISBN 0-273-08422-4.

Chao, B. F., 2005. On inversion for mass distribution from global (time-variable) gravity field. *Journal of Geodynamics*, **39**(3), 223–230.

Chui, C. K., 1992. *An Introduction to Wavelets*. San Diego: Academic Press.

Cordell, L., and Henderson, R. G., 1968. Iterative three-dimensional solution of gravity anomaly data using a digital computer. *Geophysics*, **33**, 596–601.

Cuer, M., and Bayer, R., 1980. Fortran routines for linear inverse problems. *Geophysics*, **45**, 1706–1719.

Dean, W. C., 1958. Frequency analysis for gravity and magnetic interpretation. *Geophysics*, **23**, 97–127.

Dimri, V., 1992. *Deconvolution and Inverse Theory: Application to Geophysical Problems, Methods in Geochemistry and Geophysics*. Amsterdam: Elsevier. ISBN 0-444-89493-4.

Dubois, J., and Diament, M., 2005. Géophysique – Cours et exercices corrigés, 3ème édition. Dunod, ISBN: 9782100491858.

Goodacre, A. K., 2006. Estimation of the minimum density contrast of a homogeneous body as an aid to the interpretation of gravity anomalies. *Geophysical Prospecting*, **28**(3), 408–414.

Hadamard, J., 1902. Sur les problèmes aux derivées partielles et leur signification physique. *Princeton University Bulletin*, **13**, 49–52.

Hildebrand, A. R., Penfield, G. T., Glen, T., Kring, D. A., Pilkington, M., Zanoguera, A. C., Jacobsen, S. B., and Boynton, W. V., 1991. Chicxulub Crater: a possible Cretaceous/Tertiary boundary impact crater on the Yucatan Peninsula, Mexico. *Geology*, **19**(9), 867–871.

Holland, J. H., 1975. *Adaptation in Natural and Artificial Systems*. Ann Arbor: Michigan University Press.

LaFehr, T. R., 1980. Gravity method. *Geophysics*, **45**(11), 1634–1639. ISSN 0016-8033.

Levenberg, K., 1944. A method for the solution of certain problems in least squares. *Quarterly of Applied Mathematics*, **2**, 164–168.

Levy, A. V., and Montalvo, A., 1985. The tunneling algorithm for the global minimization of functions. *SIAM Journal on Scientific and Statistical Computing*, **6**(1), 15–29.

Marquardt, D., 1963. An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, **11**, 431–441.

Menke, W., 1989. *Geophysical Data Analysis: Discrete Inverse Theory*. San Diego: Academic. International Geophysics Series, Vol. 45. ISBN 0-12-490921-3.

Milsom, J., 1996. *Field Geophysics*. Chichester: Wiley. ISBN 0-471-96634-7.

Monteiro Santos, F. A., 2010. Inversion of self-potential of idealized bodies anomalies using particle swarm optimization. *Computers and Geosciences*, **36**(9), 1185–1190.

Moritz, H., 1989. A set of continuous density distributions within a sphere compatible with a given external gravitational potential. *Gerlands Beitr Geophysik*, **98**, 185–192.

Parker, R. L., 1975. The theory of ideal bodies for gravity interpretation. *Geophysical Journal of the Royal Astronomical Society*, **42**, 315–334.

Sabatier, P. C., 2000. Past and future of inverse problems. *Journal of Mathematical Physics*, **41**(6), 4082–4124.

Strykowski, G., 1998. Some technical details concerning a new method of gravimetric-seismic inversion. In *Procedings of the XXIII General Assembly of the European Geophysical Society*, Nice. Physics and Chemistry of the Earth, Vol. 24, pp. 207–214.

Talwani, M., Worzel, J. L., and Landisman, M., 1959. Rapid gravity computations for two-dimensional bodies with application to the Mendocino submarine fracture zone. *Journal of Geophysical Research*, **64**, 49–61.

Tarantola, A., 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Philadelphia: SIAM. ISBN 978-0-898715-72-9.

Tarantola, A., and Valette, B., 1982. Generalized nonlinear inverse problems solved using the least squares criterion. *Reviews of Geophysics and Space Physics*, **20**(2), 219–232.

Tikhonov, A., and Arsenine, V., 1976. *Méthodes de Résolution des Problèmes Mal Posés*. Moscou: Editions Mir.

Troyan, V. N., and Hayakawa, M., 2002. *Inverse Geophysical Problems*. Tokyo: Terrapub Editor. ISBN 4-88704-131-4.

Tsoulis, D., and Petrovi, S., 2001. Short Note On the singularities of the gravity field of a homogeneous polyhedral body. *Geophysics*, **66**(2), 535–539.

Vanicek, P., and Kleusberg, A., 1985. What an external gravitational potential can really tell us about mass distribution. *Bolletino Geofisica Teoretica et Applicata*, **XXCII**(108), 243–250.

Zhdanov, M. S., 1988. *Integral Transform in Geophysics*. Berlin: Springer. ISBN 3-540-17759-0.

Zidarov, D., 1990. *Inverse Gravimetric Problems in GeoProspecting and Geodesy*. Amsterdam: Elsevier. ISBN 0-444-98777-0.

Zuber, M. T., Solomon, S. C., Phillips, R. J., Smith, D. E., Tyler, G. L., Aharonson, O., Balmino, G., Banerdt, W. B., Head, J. W., Johnson, C. L., Lemoine, F. G., McGovern, P. J., Neumann, G. A., Rowlands, D. D., and Zhong, S., 2000. Internal structure and early thermal evolution of mars from mars global surveyor topography and gravity. *Science*, **287**(5459), 1788–1793.

Cross-references

Geoid Undulation, Interpretation
Gravimeters
Gravity Anomalies, Interpretation
Gravity Data, Regional – Residual Separation
Gravity Method, Principles
Gravity Method, Satellite
Gravity, Data to Anomalies
Impact Craters on Earth
Inverse Theory, Artificial Neural Networks
Inverse Theory, Global Optimization
Inverse Theory, Linear
Inverse Theory, Singular Value Decomposition
Isostasy
Seismic Tomography
Seismic, Ray Theory
Spherical Harmonic Analysis Applied to Potential Fields

GRAVITY, DATA TO ANOMALIES*

Ron Hackney
Petroleum and Marine Division, Geoscience Australia, Canberra, ACT, Australia

Definition

Gravity anomaly. The difference between gravity measured at a point and a model value at that point that is based on the normal gravity of a reference ellipsoid, corrected for the gravity effects of elevation above the reference ellipsoid and the mass of rock between the point and the reference ellipsoid.

Introduction

The study of anomalous gravity has its roots in geodesy where it is used to determine the shape of the Earth (see *Geodesy, Physical* and *Geodesy, Figure of the Earth*). Gravity anomalies have also proved extremely useful in the interpretation of subsurface geological structure at various scales. Like many geophysical techniques, resource exploration has been the greatest driver of the use of gravity data (e.g., Nabighian et al., 2005). Gravity anomalies are often useful in the early stages of an exploration program as they provide insight into the form of low-density sediment accumulations (basins) or the location of high-density ore deposits. At larger scales, gravity anomalies provide insight into crustal structure and, most recently, time variations in the long-wavelength gravity field measured using satellites (see *Gravity Method, Satellite*) provide insight into changes in polar ice mass and groundwater fluctuations.

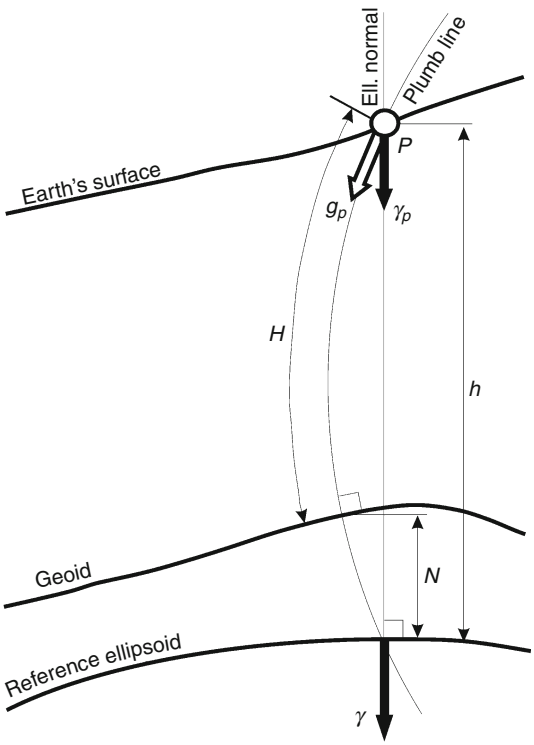
Before gravity data are useful in resource exploration or crustal studies, non-geological contributions to the gravity field that obscure the anomalies of interest need to be accounted for and effectively removed. This chapter explains the process of converting gravity measurements to gravity anomalies for geophysical applications. It is important to stress the geophysical context as there are often subtle differences in the way geodesists and geophysicists think of and use the various gravity-related quantities (cf. Li and Götze, 2001; Hackney and Featherstone, 2003; Hofmann-Wellenhof and Moritz, 2006).

Key parameters

Figure 1 shows the key parameters required for computing anomalous gravity. The relevant gravity quantities for geophysical applications are:

- Gravity measured at the point P, g_P
- Normal gravity on the surface of the reference ellipsoid, γ
- A model or predicted gravity value at the measurement point P, γ_P, derived from γ

The elevation of the point P is also required to account for the decrease in gravity with increasing elevation and the gravity effect of the mass of rock between the measurement point and the reference ellipsoid. Two height quantities need to be considered (Figure 1). The height of the point P above the geoid, H, is measured along a plumb line, a curved line that is perpendicular to equipotential surfaces (the geoid is the equipotential surface that most closely approximates mean sea level – see *Geoid Determination, Theory and Principles*). The ellipsoidal height, h,



Gravity, Data to Anomalies, Figure 1 Diagram illustrating the key parameters required for computing geophysical gravity anomalies.

is measured along an ellipsoidal normal, the line through P that is perpendicular to the surface of the reference ellipsoid. When computed algebraically (i.e., disregarding the vector nature of gravity), the difference between h and H gives the geoid–ellipsoid separation, N, where

$$N = h - H. \quad (1)$$

Observed gravity

The measured value of gravity at a point P is the value typically determined from relative gravity measurements that are converted to an observed value by tying to an international network of reference gravity stations, currently the International Gravity Standardization Network of 1971 (IGSN71: Morelli et al., 1974). Country-specific datums may also be used (e.g., Tracey et al., 2007), but these are also generally tied to IGSN71. The relative measurements may be made on land or at sea (see *Gravity Method, Surface*) or even in the air (see *Gravity Method, Airborne*). The relative measurements will have been corrected for instrument drift and Earth tide variations (see *Gravity Method, Surface*).

Anomalous gravity in geophysics

An anomaly is the difference between something that we measure at a particular place and what we expect from a reference model at that place. In seismology, measured travel times are compared to those expected from simplified models of the whole Earth (see *Seismic Tomography*). The resulting travel-time anomalies tell us about regions through which seismic waves travel faster or slower than through the simplified representation of reality. This in turn tells us something about the temperature or composition of the Earth's interior. In geophysical applications, gravity anomalies are best thought of in a similar way: Measurements of the gravity field at the Earth's surface are compared to the gravity expected from a reference model of the Earth. The differences in gravity between our measurements and the reference model tell us something about lateral mass-density variations within the Earth.

In this context, a geophysicist generally seeks to remove large-scale, non-geological gravity effects that mask the gravity anomalies that can be related to subsurface density variations (e.g., related to ore deposits, sedimentary basins, or crustal structure). The large-scale effects include Earth mass and rotation and, depending on the application, maybe some or all of the crust and mantle. Most of the large-scale effect is accounted for by normal gravity, the acceleration due to gravity on the surface of a reference ellipsoid, an ellipsoid of rotation that best fits the shape of the Earth (represented by the geoid – see *Geodesy, Figure of the Earth*).

If a model value is required at a point that does not lie on the surface of the reference ellipsoid, adjustments must be made to normal gravity to account for gravity effects related to (a) the difference in height between the measurement point and the surface of the reference ellipsoid (because gravity decreases with distance from the Earth proportionally to $1/r^2$) and (b) the body of rock (or water) lying between the measurement point and the reference ellipsoid. Other regional-scale gravity effects that arise from crustal structure might also be removed by a process of regional–residual field separation (see *Gravity Data, Regional – Residual Separation*) or filtering (see *Gravity Data, Advanced Processing*).

In the geophysical context, anomalous gravity is calculated by subtracting a predicted or model gravity value for a point P from the measured (observed) gravity at that point. This anomalous gravity quantity is technically referred to as a gravity disturbance (cf. Hackney and Featherstone, 2003), but the use of gravity "anomaly" is and will likely remain entrenched in the geophysical vernacular. In a geophysical sense, a gravity anomaly can be expressed as

$$\textit{Gravity anomaly} = (\textit{Measured value}) - (\textit{Model value}),$$

or mathematically as:

$$\Delta g_P = g_P - \gamma_P \quad (2)$$

where

- Δg_P is the anomalous gravity
- g_P is measured gravity at the point P
- γ_P is a predicted or model value of gravity *at the measurement point P* (our best estimate of the Earth's gravity field at P that considers Earth's mass and rotation and the elevation, h, of the point P with respect to the reference ellipsoid)

As illustrated in Figure 2, the predicted gravity value, γ_P, is computed by applying corrections to normal gravity that account for the height of the measurement point above the reference ellipsoid (free-air correction) and the mass of rock that lies between the measurement point and the reference ellipsoid (Bouguer and terrain corrections). Mathematically, the computation of γ_P can be expressed as:

$$\gamma_P = \gamma - \delta g_F + \delta g_B - \delta g_T \quad (3)$$

where

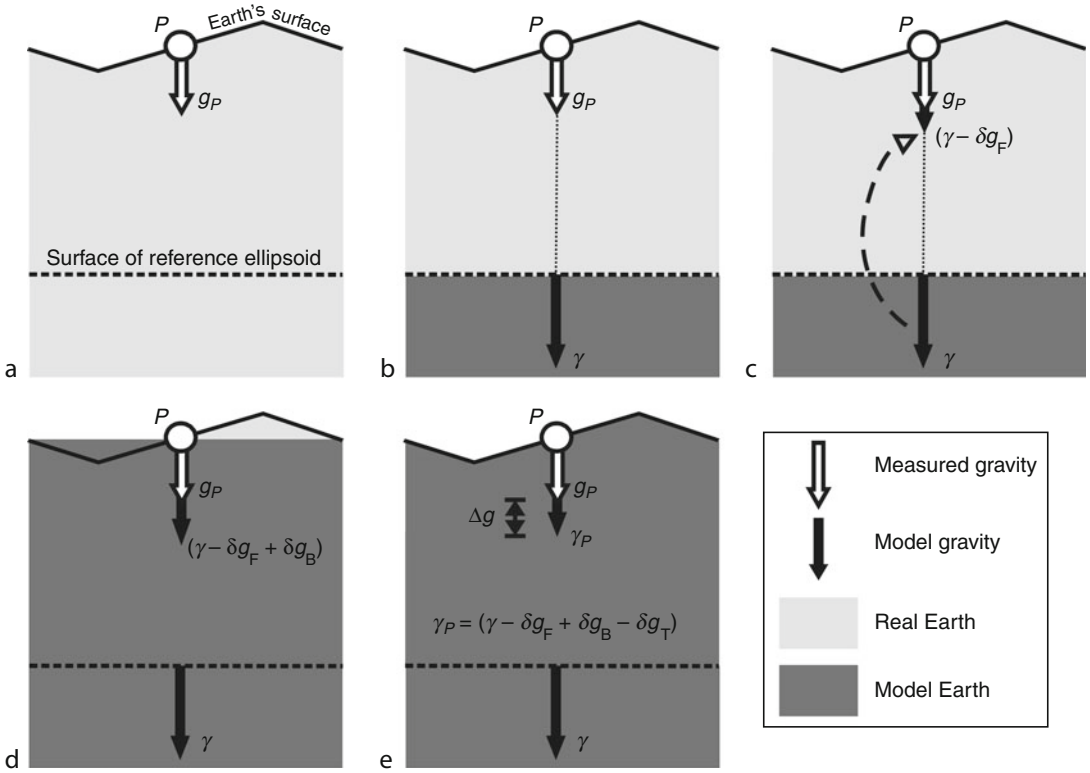
- γ is normal gravity on the reference ellipsoid
- δg_{F} is the free-air correction
- δg_{B} the Bouguer correction
- δg_{T} the terrain correction

The application of these corrections is described in the following section.

Computing model gravity at the measurement point

A note on units

The standard SI unit of acceleration due to gravity is $\mathrm{m{\cdot}s^{-2}}$. However, because gravity anomalies are generally about 3–7 orders of magnitude smaller than total gravity acceleration, the use of derivative units is more practical.



Gravity, Data to Anomalies, Figure 2 Illustration of the process of computing anomalous gravity in geophysics. (**a**) Observed gravity g_P reflects mass below the Earth's surface (*light gray shading*). (**b**) Normal gravity, γ, considers only mass below the reference ellipsoid (*dark gray shading*). (**c**) The free-air correction, δg_F, accounts for the elevation difference between the point *P* and the surface of the reference ellipsoid, but not the rock mass in that space. (**d**) The Bouguer correction, δg_B, adds the gravity effect of rock mass between *P* and the surface of the reference ellipsoid, but assumes that rock mass has a smooth upper surface. (**e**) The terrain correction, δg_T, corrects for the smooth-topped Bouguer plate or cap.

Historically, the mGal was the standard derivative unit (1 mGal = 10^{-5} m·s^{-2}). The mGal is still in common use, but gravity anomalies and related quantities are often also given in μm·s^{-2} (10^{-6} m·s^{-2}). In this description of the anomaly computation procedure, physical constants are given in standard SI units. This means that unless otherwise stated, the equations presented give quantities in m·s^{-2}.

Normal gravity (Figure 2b)

Normal gravity, γ, represents the gravity effect of a reference ellipsoid whose geometry is defined to best represent the shape of the Earth. Because an ellipsoid is a mathematically defined body, computing its gravity effect is straightforward. A closed-form expression for computing normal gravity on the surface of a *geocentric* reference ellipsoid is given by the Somigliana–Pizetti formula (e.g., Hofmann-Wellenhof and Moritz, 2006; see also *International Gravity Formula*):

$$\gamma = \gamma_a \frac{1 + k\sin^2\phi}{\sqrt{1 - e^2\sin^2\phi}}, \tag{4}$$

where γ_a is normal gravity at the equator of the reference ellipsoid, ϕ is the *geocentric* latitude of the measurement point, and k and e^2 are quantities derived from the parameters describing the form of the reference ellipsoid (Table 1).

Equation 4 is exact and has now been adopted in many countries as a standard when computing gravity anomalies for geophysical applications (e.g., Hinze et al., 2005; Tracey et al., 2007). In the past, simpler expressions were used for computing normal gravity (cf. Li and Götze, 2001). These approximate formulae were used largely for computational convenience, but given the measurement accuracy achievable today, their continued use is difficult to justify.

Free-air correction (Figure 2c)

Equation 4 gives normal gravity on the surface of the reference ellipsoid. However, the measurement point *P* is almost never located exactly on the surface of the reference ellipsoid. This means that the model value at a measurement point above the reference ellipsoid will be less than γ. The decrease in gravity above the reference ellipsoid is accounted for by the free-air correction. This

Gravity, Data to Anomalies, Table 1 Parameters defining the GRS80 geocentric reference ellipsoid (Moritz, 2000), the reference ellipsoid currently recommended by the International Union of Geodesy and Geophysics as representing the size, shape, and gravity field of the Earth with an accuracy sufficient for most geodetic, geophysical, astronomical, and hydrographic applications. Note that the parameters a, GM, J_2, and ω are exact defining parameters – the other parameters are derived geometric and physical constants

Parameter	Name	Equation	GRS80 value
a	Semi-major axis (equatorial radius)		6,378,137 m
GM	Geocentric gravitational constant		$3{,}986{,}005 \times 10^8$ m^3·s^{-2}
J_2	Dynamic form factor		$108{,}263 \times 10^{-8}$
ω	Angular velocity		$7{,}292{,}115 \times 10^{-11}$ rad·s^{-1}
b	Semi-minor axis (polar radius)		6,356,752.3141 m
γ_a	Equatorial normal gravity		9.780 326 7715 m·s^{-2}
γ_b	Polar normal gravity		9.832 186 3685 m·s^{-2}
k	Normal gravity constant	$(b\gamma_b/a\gamma_a) - 1$	0.001 931 851 353
e^2	First numerical eccentricity	$(a^2 - b^2)/a^2$	0.006 694 380 022 90
f	Ellipsoidal flattening	$(a - b)/a$	0.003 352 810 681 18
m	Geodetic parameter	$\omega^2 a^2 b/GM$	0.003 449 786 003 08
R_E	Arithmetic mean Earth radius	$(2a + b)/3$	6,371,008.7714 m

correction essentially uses the rate of change in gravity above the surface of the Earth to correct for the decrease in gravity with elevation.

Historically, the value of the free-air correction was approximated as 0.3086×10^{-5} m·s^{-2}/m (i.e., 0.3086 mGal/m), the gravity gradient above a spherical Earth. A more rigorous expression for the free-air correction, δg_F, can be derived from the gradient of normal gravity above the surface of the reference ellipsoid. Normal gravity at a height h above the ellipsoid can be computed directly from (Hofmann-Wellenhof and Moritz, 2006):

$$\gamma_h = \gamma\left[1 - \frac{2}{a}\left(1 + f + m - 2f\sin^2\phi\right)h + \frac{3}{a^2}h^2\right], \quad (5)$$

or the magnitude of the free-air correction, δg_F, to be subtracted from γ can be computed using

$$\delta g_F = \frac{2\gamma_a}{a}\left(1 + f + m + \left(-3f + \frac{5}{2}m\right)\sin^2\phi\right)h - \frac{3\gamma_a}{a^2}h^2. \quad (6)$$

The parameters f (ellipsoidal flattening) and m (geodetic parameter) are defined in Table 1.

Bouguer correction (Figure 2d)

Once normal gravity is corrected for the vertical separation between the reference ellipsoid and the measurement point, a further adjustment needs to be made for the body of rock between that point and the reference ellipsoid. The gravity effect of this rock is not represented in normal gravity and is not accounted for by the free-air correction. This correction, the Bouguer correction, δg_B, was historically computed by representing the body of rock between the measurement point and surface of the reference ellipsoid as a horizontal slab extending to infinity. The infinite-slab Bouguer correction is given by:

$$\delta g_B = 2\pi G\rho t \quad (7)$$

where ρ is the assumed density of the slab, t is the thickness of the slab, and G is the universal gravitational constant (6.67428×10^{-11} m^3·kg^{-1}·s^{-2}, the 2006 value recommended by the Committee on Data for Science and Technology, International Council of Science). In most cases, the slab thickness is equivalent to the height of the measurement point above the reference ellipsoid, h. The density of the slab generally takes on a standardized value of 2,670 kg·m^{-3}, a value that closely represents the average density of crustal rocks (Hinze, 2003). In some cases, a density that is more representative of local geology may be used to minimize the correlation between the Bouguer anomaly and topography. However, using standardized values facilitates the comparison of different datasets and avoids incorporating geological interpretation into the anomaly computation procedure (LaFehr, 1991a).

More sophisticated representations of the rock mass below the measurement point take into account the curvature of the Earth's surface. A commonly used representation is the spherical cap, the gravity effect of which is given by (LaFehr, 1991b):

$$\delta g_B^{cap} = 2\pi G\rho[\mu t - \lambda(R_E + t)] \quad (8)$$

where R_E is the mean radius of the reference ellipsoid (Table 1), t the thickness of the cap (generally equivalent to the height of the measurement point above the reference ellipsoid, h), and μ and λ are dimensionless coefficients related to R_E and the angle subtended by the spherical cap at the center of the Earth. The spherical cap is generally taken to extend to a distance of 166.735 km from the measurement point, the distance originally selected by Bullard (1936) to minimize the difference between the gravity effect of the spherical cap and the infinite slab

for a range of elevations (cf. LaFehr, 1991b). Equation 8 is valid for surveys made on land, but Argast et al. (2009) present an extended form of the equation that is valid for gravity measurements made in the air and at sea.

Terrain correction (Figure 2e)

Regardless of whether the Bouguer correction is computed from an infinite slab or from a spherical cap, it is computed for a body with a flat or curved top that does not match topography. Topographic variations can have a significant influence on measured gravity, particularly in areas of high relief (e.g., mountain ranges or near the continental slope). For this reason, it is often beneficial to compute a terrain correction that effectively imparts an irregular surface onto the top of the Bouguer plate or cap.

Various methods exist for computing the terrain correction (cf. Nowell, 1999; Hackney and Featherstone, 2003), but the methods generally involve dividing the region around the measurement point into a series of rectangular prisms or concentric ring segments. A digital elevation model is then used to determine the average elevation for each prism relative to the height of the measurement point. Accuracy can be improved by using prisms with sloping tops. The size of the prisms is generally increased away from the measurement point as the influence of distant prisms is less than for nearer prisms. To further simplify computations, some methods approximate the more distant prisms as a line mass and fast-Fourier transform techniques have also been developed.

Often the different methods for computing terrain corrections are combined. For example, Hinze et al. (2005) describe the procedure adopted for the North American gravity database that uses segmented rings and estimated topography within 100 m of the measurements point, segmented rings and high-resolution digital elevation data from 100 to 895 m, and vertical prisms and digital elevation data from 895 m to 166.7 km (including corrections for Earth curvature beyond 14 km).

Historically, the most commonly used terrain correction procedure was based on computing the gravity effect of a number of concentric rings divided into individual segments (Hammer, 1939). The gravity effect of each prism can be computed from:

$$\delta g_{\text{prism}} = 2\pi G\frac{\rho}{n}\left(r_2 - r_1 + \sqrt{r_1^2+z^2} - \sqrt{r_2^2+z^2}\right) \quad (9)$$

where n is the number of prism segments in each concentric ring (zone), r_2 is the outer radius of the concentric ring, r_1 the inner radius, and z is the elevation difference between the measurement point and the mean elevation of the ring segment. The other parameters are the same as in Equations 6 and 7. The density, ρ, used in Equation 9 should be the same as that used for the Bouguer correction.

The Hammer (1939) formulation includes concentric rings out to a distance of 22 km, but ideally terrain corrections should be computed out to 166.735 km (~1.5°), the full radius of the standard spherical cap. When terrain corrections are computed to this distance, Earth curvature must also be considered. However, the influence of terrain corrections beyond about 10–20 km from the station is often insignificant and sometimes ignored.

Anomaly computation

Once the free-air, Bouguer, and terrain corrections are applied to normal gravity, the gravity anomaly at the point P can be computed from Equations 2 and 3, i.e.,

$$\begin{aligned}\Delta g &= g_P - \gamma_P \\ &= g_P - (\gamma - \delta g_{\text{F}} + \delta g_{\text{B}} - \delta g_{\text{T}}) \\ &= g_P - (\gamma_h + \delta g_{\text{B}} - \delta g_{\text{T}})\end{aligned} \quad (10)$$

In Equation 10, the term in brackets represents the predicted or model value of gravity at the measurement point. The free-air correction is subtracted because normal gravity is reduced above the surface of the ellipsoid (Figure 2c), while the Bouguer correction is added because the extra gravity effect related to the body of rock between P and the reference ellipsoid is not included in γ (Figure 2d). The terrain correction must be subtracted from γ (Figure 2e) to remove the effect of the non-existent slab/cap mass assumed to fill valleys and to reduce the model valley of gravity for topography higher than the measurement point (mass above the measurement point reduces gravity).

Once the model value of gravity is subtracted from the measured value of gravity at P, the remaining quantity is the anomalous component of the gravity field that cannot be explained by large-scale gravity effects related to the Earth's shape and rotation, changes in gravity with elevation, and the bulk influence of rock below the measurement point. The resulting anomalous gravity reflects subsurface density variations and can be interpreted geologically (see *Gravity Anomalies, Interpretation*; *Gravity Modeling, Theory and Computation*).

Other factors

Atmospheric correction

The parameters that define the GRS80 reference ellipsoid (Table 1) rely on satellite-based measurements, meaning that normal gravity includes the mass of the atmosphere. The application of a correction to normal gravity for the atmospheric mass is usually insignificant in comparison to errors associated with terrain corrections, meaning that its application is not essential for many geophysical applications (e.g., Featherstone and Dentith, 1997). Nevertheless, the application of an atmospheric correction is included in new standards for gravity anomaly computation (e.g., Hinze et al., 2005; Tracey et al., 2007). If it is to be applied, the value to be subtracted from normal gravity can be computed using (Wenzel, 1985)

$$\delta g_{\text{A}} = 0.874 - 9.9\times10^{-5}h + 3.56\times10^{-9}h^2. \quad (11)$$

This equation has an accuracy of about 0.01×10^{-5} $m \cdot s^{-2}$ to 10 km elevation.

Further isolating anomalies of interest

The Bouguer anomaly computed using Equation 10 is dominated by the gravity effect of crustal-thickness variations that more-or-less isostatically compensate for topography (thick crust under mountains, thin crust under ocean basins – see *Isostasy*). While the Bouguer anomaly is useful for constraining crustal-scale studies, the gravity effect of smaller-scale density variations within the upper crust (e.g., ore deposits or sedimentary basins) is often obscured by the long-wavelength gravity signature related to crustal-thickness variations.

The long-wavelength component of the Bouguer anomaly can be removed using regional–residual field separation techniques (see *Gravity Data, Regional – Residual Separation*). Alternatively, some form of isostatic compensation can be assumed (see *Isostasy*) and the gravity effect of the Moho that compensates for topography can be computed and subsequently subtracted.

Distant relief

Computation of the Bouguer correction and associated terrain corrections are generally made out to a standard distance of 166.735 km. In certain circumstances, the gravity effect of topography beyond this distance can be significant (e.g., surveys in areas of high topography and continental- and global-scale studies: LaFehr, 1991a; Mikuška et al., 2006). Users of gravity data should be aware of these distant relief effects and consider them if necessary.

Efforts to include the effects of whole-Earth topography are emerging. For example, Kuhn et al. (2009) present a grid of terrain corrections for the Australian continent that considers topography globally and Tenzer et al. (2009) computed gravity anomalies for the entire Earth that incorporate corrections for global topography.

The geophysical "indirect effect"

The formulation for anomalous gravity described here (Equation 2) is the most logical way in which to think of gravity anomalies for geophysical applications. However, the process of anomaly computation is often described as a reduction of g_P to the sea level datum. While this process is useful in geodesy, it is not strictly correct in geophysical applications (Hackney and Featherstone, 2003).

The misconception that gravity anomaly computation for geophysical purposes involves reducing gravity measurements to the geoid (or sea level) is partly related to the measurement of elevation. Before the advent of GPS (see *GPS, Data Acquisition and Analysis*), elevations were normally determined by leveling over long traverses from reference tide gauges at the coastline (see *Geodesy, Ground Positioning and Leveling*). Leveling thus provided heights relative to sea level, or the geoid. However, normal gravity is derived from a mathematically defined body, the reference ellipsoid, which does not coincide exactly with the geoid.

The reference ellipsoid is generally chosen to best fit the geoid either regionally or globally, but the ellipsoid never perfectly matches the geoid. This means that if heights relative to sea level are used to compute anomalous gravity, then the corrections applied will either be an under- or over-correction, depending on whether the ellipsoid is above or below the geoid (Figure 3). The magnitude of this under- or overcorrection is referred to as the *geophysical indirect effect* (e.g., Hackney and Featherstone, 2003). Globally, the geoid–ellipsoid separation is about ± 100 m at most, meaning that the geophysical indirect effect varies by about $\pm 30 \times 10^{-5}$ $m \cdot s^{-2}$. This variation can induce long-wavelength errors in gravity data, especially in areas of high geoid gradient. For more regional surveys, the effect is generally minor and, in many cases, essentially constant. For regional surveys, the geophysical indirect effect is a bias that will often be accounted for during regional–residual field separation.

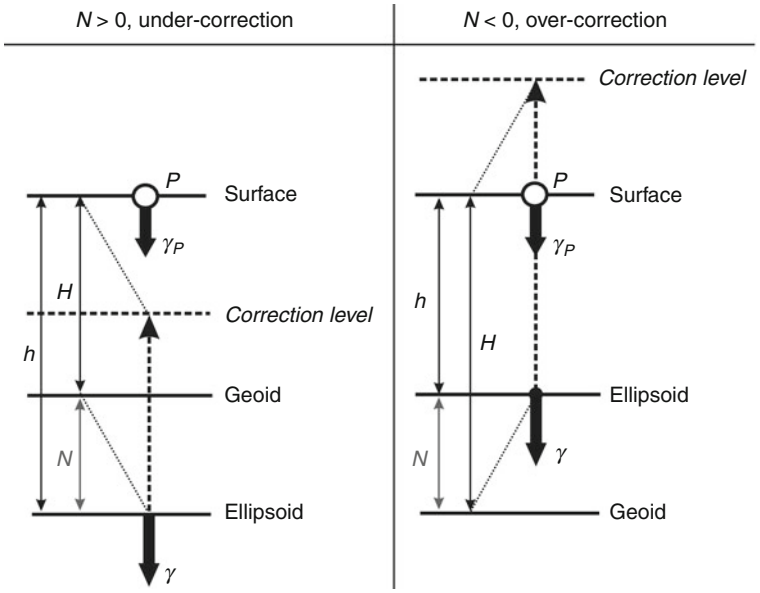
With the advent of GPS, elevations tied directly to the WGS84 ellipsoid became readily available. Because the differences in the dimensions of the WGS84 and GRS80 ellipsoids are minor, anomalous gravity can be computed directly from the heights measured by GPS. However, care is required in order to ensure that consistent heights are used for both older and newer datasets. Updated procedures adopted in some countries recognize the need to use heights tied to the reference ellipsoid (e.g., Hinze et al., 2005; Tracey et al., 2007).

Gravity anomalies for geodetic purposes

This entry focuses on the geophysical approach to computation of gravity anomalies, but it is worth outlining the ways in which gravity anomalies are computed for geodetic purposes. In geodesy (see *Geodesy, Physical* and *Geoid, Computational Method*), gravity anomalies are used to determine the shape of the Earth using either a "classical" or "modern" approach (e.g., Hofmann-Wellenhof and Moritz, 2006). In the classical approach, the geoid is determined using gravity anomalies computed from the difference between surface gravity measurements reduced (downward) to the geoid and normal gravity on the reference ellipsoid. The modern approach leads to the computation of a surface referred to as the telluroid and requires gravity anomalies computed from the difference between surface gravity and normal gravity corrected (upward) to the telluroid. The separation between the telluroid and the surface (the "height anomaly") can then be used to determine a geoid-proxy referred to as the quasigeoid.

Computation example: the Central Andes

The Central Andes in South America are an ideal region to illustrate the process of computing anomalous gravity for geophysical purposes. This is because the high topography and associated thick crust lead to large corrections



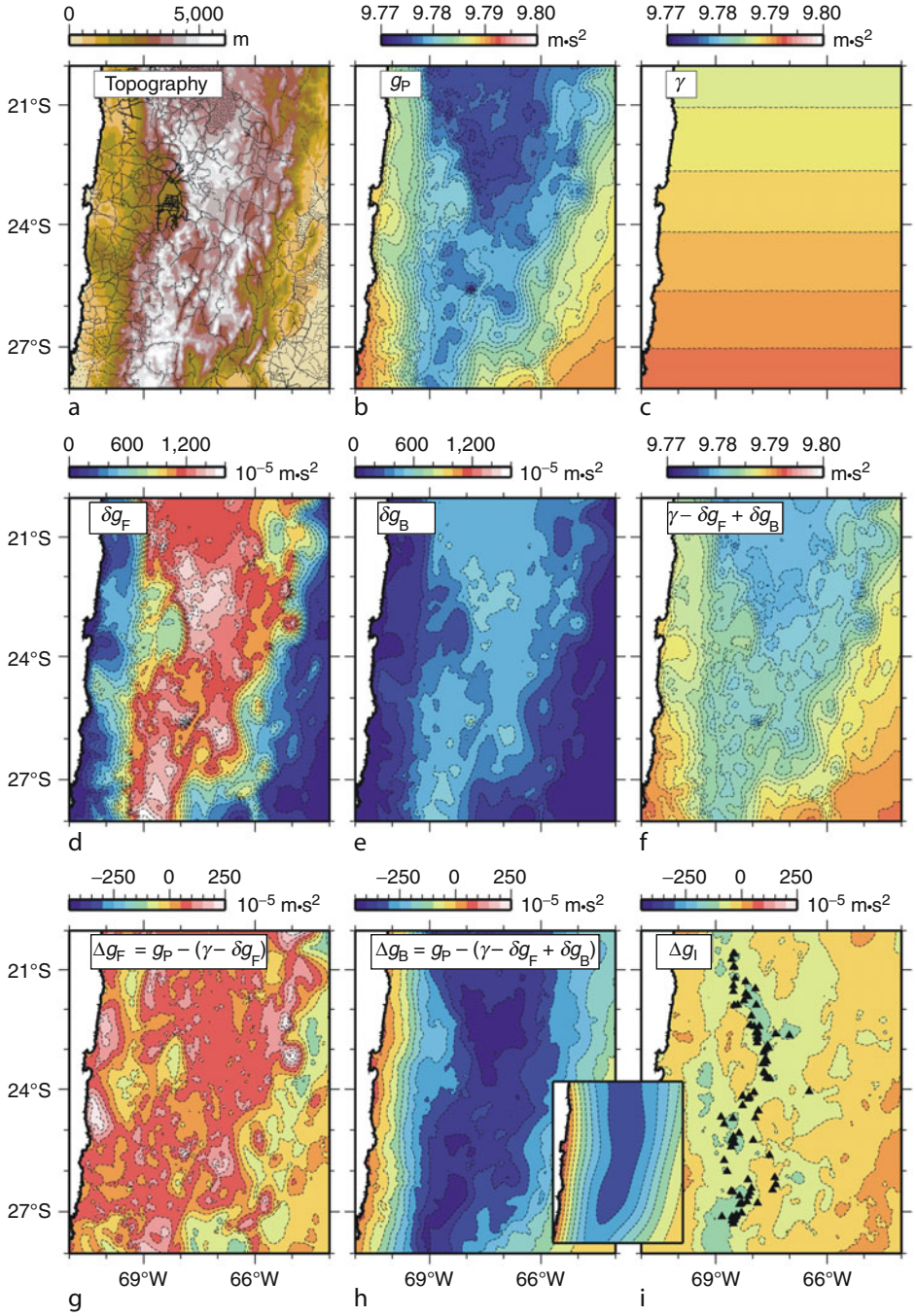
Gravity, Data to Anomalies, Figure 3 Illustration of the under- or over-correction of model predicted gravity that results when the free-air and Bouguer corrections are computed using heights relative to mean sea level (geoid) rather than relative to the reference ellipsoid. *N* is defined by Equation 1.

and high-amplitude gravity anomalies. Observed gravity and the different types of corrections, model values, and anomalies for the Central Andes are illustrated in Figure 4. The gridded data shown in Figure 4 are based on a dataset that contains more than 4,000 surface gravity measurements (Schmidt and Götze, 2006) (Figure 4a). Figure 4b shows the observed (measured) values of gravity, g_P, that are tied to reference stations linked to the IGSN71 gravity datum. Observed gravity values are large (9.773–9.796 $m{\cdot}s^{-2}$) and the highest parts of the Central Andes are associated with the lowest measured gravity. This is due mostly to the pronounced decrease in gravity with elevation, but the mass deficit associated with the thick, low density crust (with respect to the underlying mantle) also contributes to the decreased gravity in the mountainous areas.

Normal gravity on the GRS80 reference ellipsoid (Equation 4 and Figure 4c) shows a north-to-south increasing trend that reflects the pole-ward increase in gravity related to the shape and rotation of the reference ellipsoid. To derive a model value directly at each of the data points shown in Figure 4a, free-air (Equation 6) and Bouguer (Equation 7 or 8) corrections are applied to normal gravity. The magnitude of the free-air correction, δg_F, is shown in Figure 4d – the correction is largest where topography is highest. The free-air correction is subtracted from normal gravity to account for the decrease in normal gravity with elevation above the reference ellipsoid. The free-air anomaly results when the height-corrected normal gravity is subtracted from observed gravity (Figure 4g). Compared to observed gravity, the amplitude range of the free-air anomaly is significantly reduced ($\pm 300 \times 10^{-5}$ $m{\cdot}s^{-2}$).

A map of free-air anomalies (Figure 4g) is difficult to interpret geologically because it retains a strong correlation with topography. This correlation is modified when the gravity effect of the rock between the surface and the reference ellipsoid is accounted for using the Bouguer correction (Figure 4e). The Bouguer correction is added to normal gravity (because normal gravity does not include the gravity effect of the rocks) and, together with the addition of the free-air correction, results in a model value, γ_P, that resembles the observed gravity (Figure 4f). When this model value is subtracted from the observed gravity, the Bouguer anomaly results (Figure 4h). The Bouguer anomaly of the Central Andes is large and negative (minimum values reach about -450×10^{-5} $m{\cdot}s^{-2}$), which reflects the thick crustal root beneath the Central Andes.

An isostatic residual gravity map is shown for the Central Andes in Figure 4i. The regional field (inset between Figure 4h, i) is computed by assuming regional isostatic compensation of topography using an effective elastic thickness of 20 km (see *Isostasy*). Crustal thickness is computed for this model and then the gravity effect of the isostatic Moho is estimated and subtracted from the Bouguer anomaly. The Central Andean isostatic residual field has a vastly reduced amplitude (-150 to $+ 60 \times 10^{-5}$ $m{\cdot}s^{-2}$) and bears a closer resemblance to surface geology, particularly the location of active volcanoes. The coincidence between active volcanoes and negative isostatic residual anomalies suggests the presence of low-density material within the crust or that the assumption of regional isostatic compensation is not applicable to the volcanic chain.



Gravity, Data to Anomalies, Figure 4 Illustration of the various components of anomalous gravity for the Central Andes in South America. (**a**) Topography from the GTOPO30 global digital elevation model – points show the location of gravity measurements (Schmidt and Götze, 2006) where the various corrections are computed and gridded using a cell size of one arc minute; (**b**) observed gravity; (**c**) normal gravity on the surface of the GRS80 reference ellipsoid computed using Equation 4 – note the latitudinal variation; (**d**) the magnitude of the free-air correction that is subtracted from normal gravity to account for the reduction in normal gravity above the reference ellipsoid; (**e**) the Bouguer correction that represents the gravity effect of the mass of rock between the surface and the reference ellipsoid; (**f**) model value for gravity at the surface that is subtracted from the observed gravity shown in b to produce the Bouguer anomaly; (**g**) the free-air anomaly; (**h**) the Bouguer anomaly (map in **b** minus map in **f**); (**i**) isostatic residual anomaly, δg_I, computed by subtracting from the Bouguer anomaly an isostatic regional field (inset between **h** and **i**) computed by assuming regional isostatic compensation of Andean topography with an elastic thickness of 20 km. The triangles in (**i**) show active volcanoes (Trumbull et al., 2006). Note that the Bouguer anomaly in (**h**) does not incorporate a terrain correction, but when plotted at this scale, the effect of the terrain correction is not evident.

Summary

From a geophysical perspective, a gravity anomaly is best thought of as the difference between measured gravity and a model value, at the same point, that accounts for large-scale, non-geological effects that mask anomalies related to subsurface geology. The model gravity value is derived by correcting normal gravity for the effects of height above the reference ellipsoid (free-air correction) and the mass of rock between the reference ellipsoid and the measurement point (Bouguer correction). Additional corrections (terrain and atmospheric corrections, indirect effect) improve on the relatively straightforward free-air and Bouguer corrections. Since the earliest use of gravity data, our ability to compute the corrections to normal gravity has improved to the point where simplistic application of the corrections is giving way to procedures that consider global influences.

Depending on the application, long-wavelength field components remaining in the Bouguer anomaly may also be removed (isostatic correction or regional–residual field separation). Gravity anomalies can then be used to constrain geological interpretations of relatively shallow crustal features (e.g., basins, ore bodies, or even archeological sites), or be used to examine crustal structure and isostatic state.

Bibliography

Argast, D., Bacchin, M., and Tracey, R., 2009. An extension of the closed-form solution for the gravity curvature (Bullard B) correction in the marine and airborne cases. *ASEG Extended Abstracts*, **2009**(1), 6, doi:10.1071/ASEG2009ab129.

Bullard, E. C., 1936. Gravity measurements in East Africa. *Philosophical Transactions of the Royal Society of London*, **235**, 486–497.

Featherstone, W. E., and Dentith, M. C., 1997. A geodetic approach to gravity data reduction. *Computers and Geosciences*, **23**(10), 1063–1070.

Hackney, R. I., and Featherstone, W. E., 2003. Geodetic versus geophysical perspectives of the "gravity anomaly". *Geophysical Journal International*, **154**, 35–43. see also: Erratum, 2003. Geophysical Journal International, 154: 596; Corrigendum, 2006. Geophysical Journal International, 167: 585.

Hammer, S., 1939. Terrain corrections for gravimeter stations. *Geophysics*, **4**, 184–194.

Hinze, W. J., 2003. Bouguer reduction density, why 2.67? *Geophysics*, **68**(5), 1559–1560.

Hinze, W. J., Aiken, C., Brozena, J., Coakley, B., Dater, D., Flanagan, G., Forsberg, R., Hildenbrand, T., Keller, G. R., Kellogg, J., Kucks, R., Li, X., Mainville, A., Morin, R., Pilkington, M., Plouff, D., Ravat, D., Roman, D., Urrutia-Fucugauchi, J., Véronneau, M., Webring, M., and Winester, D., 2005. New standards for reducing gravity data: the North American gravity database. *Geophysics*, **70**(4), J25–J32, doi:10.1190/1.1988183.

Hofmann-Wellenhof, B., and Moritz, H., 2006. *Physical Geodesy*. Vienna: Springer.

Kuhn, M., Featherstone, W. E., and Kirby, J. F., 2009. Complete spherical Bouguer anomalies over Australia. *Australian Journal of Earth Sciences*, **56**(2), 213–223, doi:10.1080/08120090802547041.

LaFehr, T. R., 1991a. Standardization in gravity reduction. *Geophysics*, **56**(8), 1170–1178.

LaFehr, T. R., 1991b. An exact solution for the gravity curvature (Bullard B) correction. *Geophysics*, **56**(8), 1179–1184.

Li, X., and Götze, H.-J., 2001. Ellipsoid, geoid, gravity, geodesy, and geophysics. *Geophysics*, **66**(6), 1660–1668.

Mikuška, J., Pašteka, R., and Marušiak, I., 2006. Estimation of distant relief effect in gravimetry. *Geophysics*, **71**(6), J59–J69, doi:10.1190/1.2338333.

Morelli, C., Ganter, C., Hankasalo, T., McConnell, R. K., Tanner, J. B., Szabo, B., Uotila, U., and Whalen, C. T., 1974. The International Gravity Standardization net 1971. *International Association of Geodesy*. Special publication number, 4, International Association of Geodesy, Paris.

Moritz, H., 2000. Geodetic Reference System 1980. *Journal of Geodesy*, **74**(1), 128–162, doi:10.1007/s001900050278.

Nabighian, M. N., Ander, M. E., Grauch, V. J. S., Hansen, R. O., LaFehr, T. R., Li, Y., Pearson, W. C., Peirce, J. W., Phillips, J. D., and Rude, M. E., 2005. Historical development of the gravity method in exploration. *Geophysics*, **70**(6), 63ND–89ND, doi:10.1190/1.2133785.

Nowell, D. A. G., 1999. Gravity terrain corrections – an overview. *Journal of Applied Geophysics*, **42**, 117–134.

Schmidt, S., and Götze, H.-J., 2006. Bouguer and isostatic maps of the Central Andes. In Oncken, O., Chong, G., Franz, G., Giese, P., Götzte, H.-J., Ramos, V. A., Strecker, M. R., and Wigger, P. (eds.), *The Andes: Active Subduction Orogeny*. Berlin/Heidelberg: Springer. Frontiers in Earth Science, Vol. 1, pp. 559–562.

Tenzer, R., Hamayun, K., and Vajda, P., 2009. Global maps of the CRUST 2.0 crustal components stripped gravity disturbances. *Journal of Geophysical Research*, **114**, B05408, doi:10.1029/2008JB006016.

Tracey, R., Bacchin, M., and Wynne, P., 2007. AAGD07: a new absolute gravity datum for Australian gravity and new standards for the Australian National Gravity Database. *ASEG Extended Abstracts*, **2007**(1), 3, doi:10.1071/ASEG2007ab149.

Trumbull, R. B., Riller, U., Oncken, O., Scheuber, E., Munier, K., and Hongn, F., 2006. The time – space distribution of Cenozoic volcanism in the Central Andes: a new data compilation and some tectonic implications. In Oncken, O., Chong, G., Franz, G., Giese, P., Götzte, H.-J., Ramos, V. A., Strecker, M. R., and Wigger, P. (eds.), *The Andes: Active Subduction Orogeny*. Berlin/Heidelberg: Springer. Frontiers in Earth Science, Vol. 1, pp. 29–43.

Wenzel, H., 1985. *Hochauflosende Kugelfunktionsmodelle fur des Gravitationspotential der Erde [1]*. Wissenschaftliche arbeiten der Fachrichtung Vermessungswesen der Universitat Hannover, 137.

Cross-references

Earth Rotation
Earth Tides
Earth, Density Distribution
Geodesy, Figure of the Earth
Geodesy, Ground Positioning and Leveling
Geodesy, Physical
Geoid Determination, Theory and Principles
Geoid, Computational Method
GPS, Data Acquisition and Analysis
Gravimeters
Gravity Anomalies, Interpretation
Gravity Data, Advanced Processing
Gravity Data, Regional – Residual Separation
Gravity Field of the Earth
Gravity Measurements, Absolute
Gravity Method, Airborne
Gravity Method, Principles
Gravity Method, Satellite

Gravity Method, Surface
Gravity Modeling, Theory and Computation
Gravity, Global Models
Gravity, Gradiometry
International Gravity Formula
Isostasy
Seismic Tomography
Spherical Harmonic Analysis Applied to Potential Fields

GRAVITY, GLOBAL MODELS*

Nikolaos K. Pavlis
Geodesy and Geophysics Basic and Applied Research, National Geospatial-Intelligence Agency (NGA), Reston, VA, USA

Synonyms

Geopotential models; Global gravitational models; Potential coefficient models

Definition

A *Global Gravitational Model* (GGM) is a mathematical approximation to the external gravitational potential of an attracting body.
Gravitation is the term used to describe the potential generated by the masses of an attracting body like the Earth.
Gravity is the term used for the potential of gravitation plus the centrifugal potential due to the rotation of the attracting body.

Global gravitational models

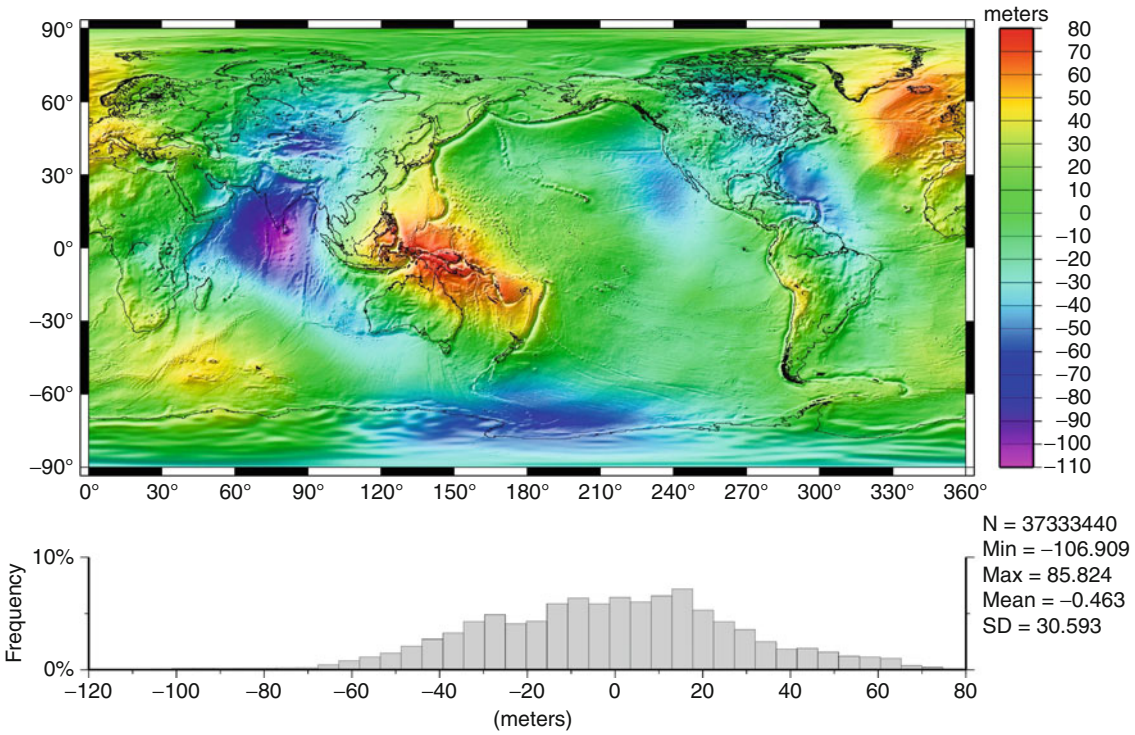
Introduction

A GGM is a mathematical approximation to the external gravitational potential of an attracting body. We will focus here on the case where the attracting body is the Earth, although many of the concepts that we discuss apply equally well to other planets and celestial bodies. In its most complete and general form, a GGM consists of a set of numerical values for certain parameters, the statistics of the errors associated with these values (as described, e.g., by their error variance-covariance matrix), and a collection of mathematical expressions and algorithms that allow a user to perform:

1. *Synthesis*, that is, computation of the numerical values of quantities related to the gravitational potential (*functionals* of the gravitational field), given the position of the evaluation point.
2. *Error Propagation*, that is, computation of the expected errors of the computed functionals, as implied by the propagation of the errors of the parameters defining the GGM.

A GGM must be able to support such computations at arbitrary points, located on or above the Earth's surface, in a fashion that is both rigorous and efficient. In addition, a GGM should fulfill certain conditions stemming from the underlying physics. Namely, it should represent a scalar function of position that is harmonic outside the attracting masses and vanishes at infinity as the reciprocal of the distance between attracted point and attracting mass element. Moreover, the GGM should permit the computation of any functional of the field in a way that guarantees *self-consistency*. This means that the model should preserve the relationships (differential or integral) between the various functionals. A GGM has numerous uses, both operational and scientific (see also [Tscherning, 1983]), including:

1. Orbit determination applications necessary for space surveillance (the detection, tracking, and orbit prediction of Earth-orbiting objects).
2. Inertial navigation applications for trajectory determination of airplanes and missiles.
3. *Geoid* determination. The geoid is one of the infinitely many *level* (or *equipotential*) surfaces of the gravity field of the Earth, that is, the gravity potential is constant at every point on the geoid. The geoid is the specific equipotential surface that would coincide exactly with the mean ocean surface of the Earth, if the oceans were in equilibrium, at rest, and extended through the continents (such as with very narrow canals). Figure 1 shows the geoid surface that is implied by the EGM2008 GGM (Pavlis et al., 2008). Geoid determination allows one to transform a *geometric* (ellipsoidal) height to an *orthometric* one, that is, to an elevation referenced to an equipotential surface, as those elevations that appear on topographic maps. This application has attracted great interest in recent years, because GPS positioning and geoid determination offer the possibility of determining orthometric heights and height differences without the need for the expensive and laborious spirit leveling (Schwarz et al., 1987).
4. Oceanographic applications that require the estimation of the Dynamic Ocean Topography (DOT), an estimate of which is shown in Figure 2, and of its slopes. These quantities are directly related to ocean circulation. This application puts very stringent accuracy and resolution requirements on GGMs (Ganachaud et al., 1997).
5. An accurate and high-resolution GGM may be used to provide a reference surface – a global, high-resolution geoid surface – suitable for the realization of a global vertical datum (Rapp and Balasubramania, 1992).
6. Geophysical prospecting applications where, in combination with other information (e.g., seismic data), a GGM may provide important constraints that aid the determination of the subsurface density distribution.
7. Comparisons of the decay of the spectrum of the gravitational potential to the decay of the spectrum of the Earth's topography (Rapp, 1989).



Gravity, Global Models, Figure 1 Geoid undulations (in meters), implied by the EGM2008 Global Gravitational Model, to degree 2,160, with respect to the WGS 84 reference ellipsoid.

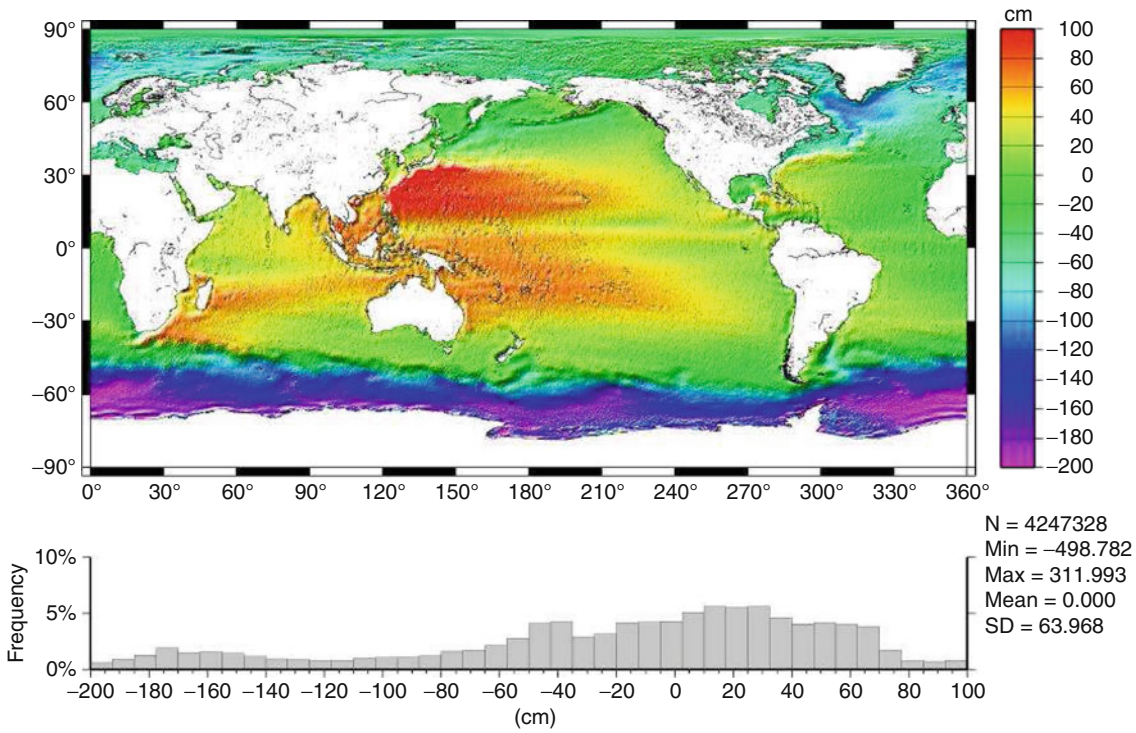
These and other applications represent integral parts of various civilian and military activities. Each of these applications has (in general) different accuracy and resolution requirements, as far as the supporting GGM is concerned. For example, due to the attenuation of the gravitational field with increasing altitude, a relatively low-resolution GGM (e.g., a *Spherical Harmonic Analysis Applied to Potential Fields* to maximum degree and order 70 or 90) is currently adequate for the precise orbit determination of most Earth-orbiting satellites. In contrast, accurate determination of the slopes of the equipotential surface (deflections of the vertical) demands a GGM of much higher resolution.

Geodesists have at various times developed "special purpose" GGMs that optimize performance for a particular application (e.g., orbit determination of a particular satellite, or geoid undulation computation over a specific geographic region). Although such "tailored" models have found some uses in the past, the ultimate goal has always been the development of a *single,* general purpose, GGM that addresses the different and diverse applications in an optimal manner, without over-performing in one application at the expense of its performance in others.

The development of a high-resolution GGM is a task that involves the optimal combination of a variety of data (satellite, land, marine, airborne). This is because a single data type with both global coverage and with uniformly high accuracy and high *spectral* sensitivity does not (yet) exist. The aforementioned data are of complementary character (in terms of spectral sensitivity and/or of geographic coverage), so that their optimal combination enables a GGM to satisfy the variety of applications described before. Comprehensive GGM solutions that combine different types of data may involve the simultaneous estimation of the parameters that describe the gravitational potential, along with parameters that describe the dynamic ocean topography, solid Earth and ocean tides, Earth orientation and tracking station position parameters, as well as a plethora of "nuisance" parameters necessary to model completely the content of certain data types (e.g., biases and delays associated with certain satellite tracking data). The result of a successful GGM development effort is a model that can be used as a *standard* for numerous applications, over a substantial period of time.

Local and regional gravimetric models

The accuracy and the resolving power of the data that were used in its development dictate the accuracy and resolution of a GGM. Geopolitical and/or proprietary data issues many times prevent the individual or the team developing a GGM from having access to *all* the existing data.



Gravity, Global Models, Figure 2 Dynamic Ocean Topography (DOT) (in centimeters), obtained by subtracting the EGM2008 geoid undulations from a Mean Sea Surface grid computed at the Danish National Space Center (DNSC). Notice the low associated with the Antarctic Circumpolar Current (ACC) and the highs associated with the Pacific Ocean gyres.

However, over some regions, data of higher accuracy and/or resolving power (e.g., geographically dense sets of gravity and elevation data) may be available to some individual(s) or may become available *after* a GGM has been developed. These data may be used in combination with the existing GGM to improve the accuracy and/or resolution of the determination of *one or more specific functionals of the field*, over the region where the detailed data became available. This local or regional "densification" can produce a specific local or regional gravimetric product or model.

Such densification has been among the favorite geodetic activities over many decades now, and represents the geodesist's way of creating a *multi-resolution* gravitational model resembling a "quilt." In effect, patches of fine detail (the *Local Gravimetric Models* – LGMs) are sewn on top of a more or less homogeneous piece of fabric (the reference GGM). The GGM is the foundation upon which any regional or local gravimetric approximation study and application is built. Geodesists do not necessarily have to reevaluate the reference GGM every time a new set of data becomes available locally. Such reevaluation is mostly warranted if new and improved satellite data become available, spanning a sufficiently long time period, and/or if new terrestrial data (of higher accuracy and/or resolution) become available over areas with substantial geographic extent.

Signal representation and data characteristics

Although geodesists have variously considered and studied the representation of the gravitational potential using *point masses* (Sünkel, 1981, 1983), *finite element* methods (Meissl, 1981; Baker, 1988), and *splines* (Sünkel, 1984; Jekeli, 2005), these approaches have seen only limited application in the representation of (especially) the "static" (i.e., the time-averaged) gravitational field of the Earth. *Spherical harmonic functions* have prevailed as the standard form used for the representation of the gravitational potential globally, from the very early days of global determinations to the present. Indeed, the set of coefficients of a *Spherical Harmonic Analysis Applied to Potential Fields* of the gravitational potential has become pretty much synonymous with a GGM.

Nerem, Jekeli, and Kaula (1995) provide a review of the gravity field determination process, including a review of global-scale modeling. Bouman (1997) provides a comprehensive "survey" of global gravity field models starting from the 1970s. Rapp (1998) provides a review of the geopotential modeling developments of

the twentieth century, which includes an extensive list of references.

The Earth's external gravitational potential, V, at a point P defined by its geocentric distance (r_P), geocentric colatitude (θ_P) (defined as 90° minus the geocentric latitude), and longitude (λ_P), can be expressed as:

$$V(r_P, \theta_P, \lambda_P) = \frac{GM}{r_P}\left[1 + \sum_{n=2}^{\infty}\left(\frac{a}{r_P}\right)^n \sum_{m=-n}^{n} C_{nm} Y_{nm}(\theta_P, \lambda_P)\right]. \quad (1)$$

GM is the geocentric gravitational constant (the product of the universal gravitational constant, G, times the mass of the Earth including its atmosphere, M) and a is a scaling factor associated with the fully normalized, unitless, spherical harmonic coefficients C_{nm} (a is usually numerically equal to the equatorial radius of an adopted mean-Earth ellipsoid). The absence of first-degree terms in Equation 1 is a consequence of placing the origin of the coordinate system at the Earth's center of mass. The *Spherical Harmonic Analysis Applied to Potential Fields* are defined as (Heiskanen and Moritz, 1967, Sect. 1–14):

$$Y_{nm}(\theta_P, \lambda_P) = \overline{P}_{n|m|}(\cos\theta_P) \cdot \begin{Bmatrix} \cos m\lambda_P \\ \sin|m|\lambda_P \end{Bmatrix} \begin{matrix} \text{if } m \geq 0 \cdot \\ \text{if } m < 0 \end{matrix} \quad (2)$$

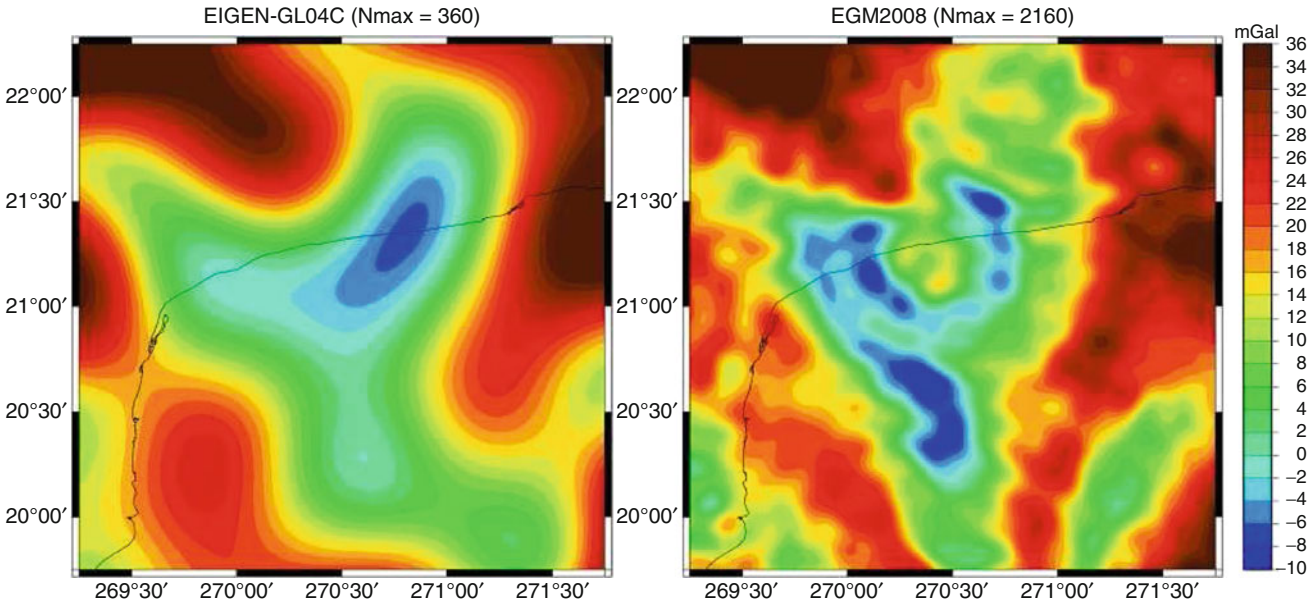
$\overline{P}_{n|m|}(\cos\theta_P)$ is the fully-normalized associated legendre function of the first kind, of degree n and order $|m|$. The normalization of the surface spherical harmonic functions Y_{nm} is such that:

$$\frac{1}{4\pi}\iint_{\sigma} Y_{nm}^2 \, d\sigma = 1, \quad (3)$$

where the integral is taken over the unit sphere σ, whose area element is $d\sigma$. In practice, the degree summation in Equation 1 is truncated to some finite degree N, which defines (approximately) the resolution of the GGM. A useful rule of thumb is that the half-wavelength linear resolution ($\lambda/2$), on the Equator, expressed in kilometers, is related to N by:

$$\lambda/2 = \frac{180^\circ}{N} \cdot 111 \quad (km). \quad (4)$$

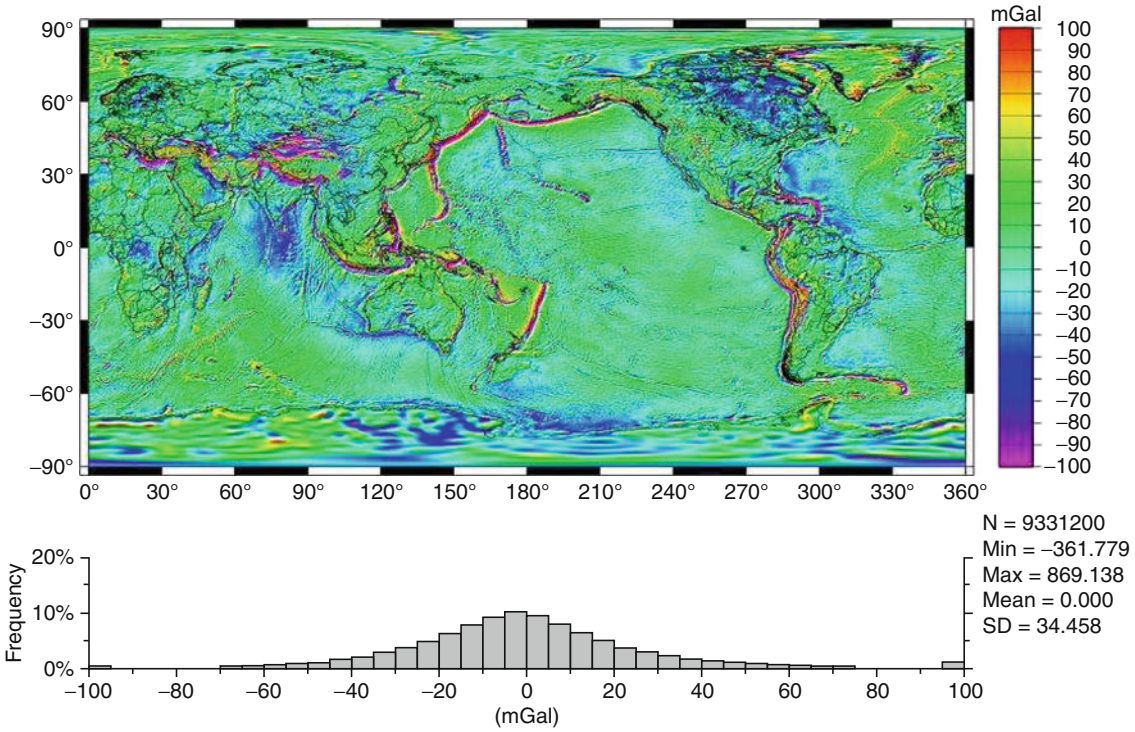
Figure 3 shows a representative example of the fine detail (high resolution) that can be captured with an expansion to degree 2,160, compared to the lower-resolution information that is retrievable using an expansion to degree 360. A spherical harmonic expansion complete to degree and order N involves $(N + 1)^2$ spherical harmonic coefficients, a number that grows rapidly as the maximum degree, and therefore the resolution of the model, increases. For example, an expansion to degree 360, as the EGM96 model involves $O(130\,\text{k})$ coefficients, while EGM2008, which extends to degree 2,160, involves



Gravity, Global Models, Figure 3 Free-air gravity anomalies (in mGal, where 1 mGal = 10^{-5} m/s^2) implied by two GGMs of different maximum degrees (different resolving powers), over the Yucatán Peninsula. EIGEN-GL04C extends to degree 360, while EGM2008 extends to degree 2,160. The Chicxulub impact crater "ring" is clearly visible in the EGM2008 gravity anomalies, but cannot be discerned in the EIGEN-GL04C gravity anomalies due to the lower resolution of this model.

approximately 4.7 million coefficients. The availability of efficient numerical algorithms for the estimation and for the subsequent use of these spherical harmonic expansions is obviously of critical importance. The goal of global high-resolution gravitational modeling is to estimate, as accurately as possible, the coefficients C_{nm}, to as high degree as possible, through the optimal combination of gravitational information from a variety of data sources. Of equal importance is the estimation of reliable error estimates for the C_{nm} values. The estimated C_{nm} values can then be used to compute functionals of the field (e.g., Geoid *undulations*, *gravity anomalies,* deflections of the vertical) while their associated errors (and error correlation when available) can be propagated to yield the errors of the derived functional(s). *Before* the dawn of the new millennium and the availability of data from the satellite missions CHAMP, GRACE, and GOCE, four kinds of gravitational information were commonly available for the development of high-degree combination gravitational models like EGM96 (Lemoine et al., 1998):

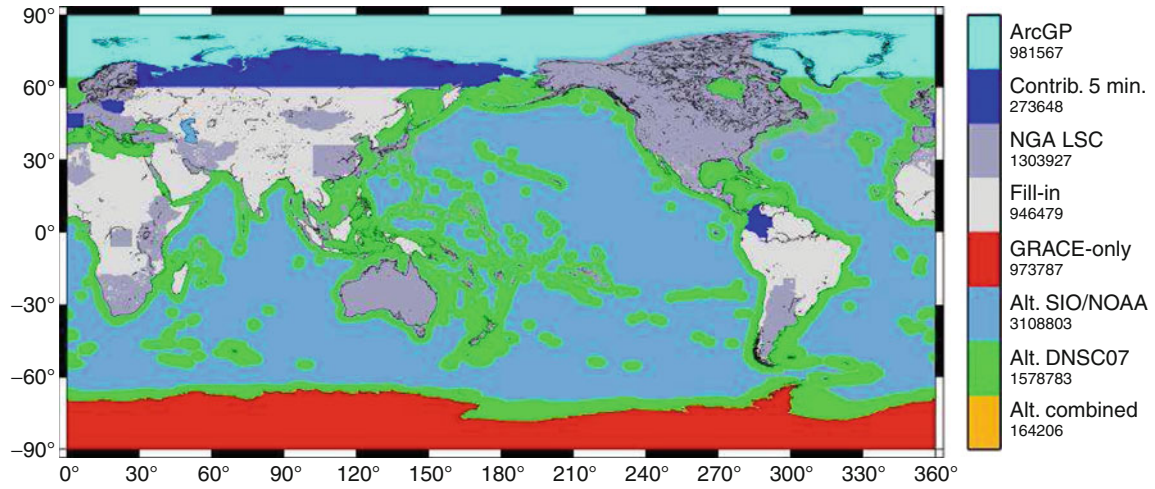
1. Gravitational information obtained from the analysis of satellite orbit perturbations that are deduced from tracking data. This is of critical importance for the accurate determination of the low degree part of the model. *Satellite-only* models have progressed from solutions to degree 4 in the early 1960s, to models complete to degree 70 or 90 available by the end of the twentieth century. These advances were made through the availability of ever more accurate tracking data acquired over a continuously expanding constellation of Earth orbiters. Tracking data from approximately 40 satellites have been used in the development of the satellite-only solution supporting EGM96 (denoted EGM96S) (Lemoine et al., 1998). These data include optical, radio Doppler and radio interferometric observations, Satellite Laser Ranging (SLR), Doppler Orbit determination and Radiopositioning Integrated on Satellite (DORIS), and Satellite-to-Satellite Tracking (SST) data from the Global Positioning System (GPS) and Tracking and Data Relay Satellite System (TDRSS) constellations to lower Earth orbiters. Despite these advances, these tracking data types are incapable of resolving the fine structure of the field due to the attenuation of the gravitational signal with altitude. Moreover, the available satellites do not sample uniformly the range of orbital inclinations and altitudes, which is a necessary condition for the de-correlation of the harmonic coefficients estimated from satellite tracking data only. This causes strong correlation especially among coefficients of higher degrees within satellite-only solutions, and necessitates the use of a priori constraints in the development of satellite-only models (Lerch et al., 1979).
2. Surface (land, marine, and airborne) gravimetric data, which are (essentially) measurements of the magnitude of the acceleration of gravity. These data are in principle capable of resolving both long and short wavelength features of the gravity field. This, however, requires uniform global coverage with dense gravity data of uniformly high accuracy. The best available data sets ca. 1996 (Kenyon and Pavlis, 1996) represent information derived from over 4,000 sources of detailed gravity data collected over several decades. The accuracy and density of point data vary substantially with geographic region, with extended regions (e.g., Antarctica) being practically void of gravity measurements. *Gravity Anomalies, Interpretation* values are the remainders after subtraction from the gravity measurements of the gravity acceleration implied by an ellipsoidal reference (normal) field (see also Figure 4). These gravity anomalies are susceptible to various systematic errors (Heck, 1990). These errors, in conjunction with the nonuniformity of coverage, degrade the long wavelength integrity of the gravitational information that can be extracted from surface gravimetry. Nevertheless, surface and airborne gravimetry presently provide the only data that can resolve short wavelength gravity features, especially over land areas. In addition, shipborne gravity measurements may aid the separation of the geoid from the DOT signal when used in combination with satellite altimetry. The processing of surface gravity data also requires accurate and high-resolution information of the Earth's topography. This information is necessary both for the preprocessing of the gravity data and for the computation of correction terms that are necessary in order to determine the geoid surface over the continents, from the estimated GGM coefficients (Rapp, 1997). It may also be used to supplement low-resolution gravitational information that may be available over certain geographic regions, with high-resolution information implied by the gravitational effects of the Earth's topography. This approach was used effectively for the development of EGM2008, as it is discussed by Pavlis et al. (2007).
3. Satellite altimeter data have enabled an unsurpassed mapping of the field over the oceans, in terms of both accuracy and resolution. TOPEX/Poseidon (T/P) (Fu et al., 1994) was the first satellite altimetry mission that routinely provided measurements of the Sea Surface Height (SSH), which were not significantly contaminated by radial orbit error (RMS radial orbit error at the ± 2 cm level). However, altimetric measurements are confined over the ocean areas bounded by the satellite's inclination, and furthermore provide a mapping of the sum of the geoid undulation (the height of the geoid surface with respect to the surface of an ellipsoid of revolution) plus the DOT. These aspects weaken somewhat the contribution of altimeter data in the determination of the long wavelength gravitational field and necessitate the appropriate modeling and estimation of the DOT when altimeter SSH data are used in combination solutions. There is, however, another way of incorporating altimeter data into a high-degree GGM, which is discussed next.



Gravity, Global Models, Figure 4 Free-air gravity anomalies (in mGal) implied by EGM2008 to degree 2,160, averaged over 5′ × 5′ cells on the surface of the Earth. These values reveal numerous geophysical features, such as oceanic trenches, ridges, subduction and fracture zones, and seamount chains. The lack of detailed data over Antarctica is also clearly evident from the smoothness of the values over this region (for the development of EGM2008, only GRACE satellite information was available over Antarctica).

4. The combination of altimeter data from multiple missions (spanning several years), some of which have produced very closely spaced ground tracks, has provided a dense sampling of most of the ocean's surface. These data, in the form of SSH and/or SSH slopes, can be used to estimate ocean-wide sets of gravity anomalies, at very fine resolution (e.g., 2′ × 2′ and even 1′ × 1′). Spatial averages of these values can be "merged" with corresponding land and airborne gravity data and gravity information obtained from models of the potential generated by the topography and its *isostatic* compensation (Pavlis and Rapp, 1990) to produce a complete global equiangular grid of gravity anomaly values. Figure 5 shows the identification of the data sources, which were used to compile the "merged" 5′ × 5′ free-air gravity anomaly data file that supported the development of EGM2008. The geometry of such a grid, in conjunction with the symmetry properties of the Legendre and the trigonometric functions, allows the applicability of very efficient harmonic analysis and *synthesis* methods (Rizos, 1979; Colombo, 1981a). These methods have revolutionized the development and use of very high-degree spherical harmonic expansions. These approaches also allow efficient combination with *satellite-only* information, as was done, for example, by Rapp and Pavlis (1990). However, incorporation of altimeter data into a GGM in this fashion requires some a priori knowledge of the DOT, so that the altimetry-derived gravity anomaly values are estimated from appropriately corrected SSH.

Satellite tracking, altimetric, and surface gravimetric data are of complimentary character both in a spectral as well as in a geographic sense. Their combination enables the determination of the gravitational field, over a wider band of its spectrum, with improved accuracy, than can be obtained by using any of the three data types alone. The particular means of combining these data, in order to develop a high-degree GGM, constitutes a solution strategy. A critical consideration in the design of a solution strategy is the treatment of altimeter data (Rapp, 1993), that is, if these data will be incorporated as in (3) or as in (4) above. OSU91A (Rapp, Wang and Pavlis, 1991) and EGM96 (Lemoine et al., 1998) represent the result of implementing a particular solution strategy, whereby altimeter data were used as in (3) for the determination of the low degree part of these models (maximum degree 50 and 70, respectively), and as in (4) for the higher degree part. The OSU89A/B models (Rapp and Pavlis, 1990) and the EGM2008 model (Pavlis et al., 2008) incorporated satellite altimeter data as in (4) above. Certain characteristics of the above data types that are particularly important for their effective combination are discussed next.



Gravity, Global Models, Figure 5 The major sources of free-air gravity anomaly data used to compile the "merged" file that supported the development of EGM2008. ArcGP stands for "Arctic Gravity Project," SIO/NOAA for "Scripps Institute of Oceanography/National Oceanic and Atmospheric Administration," which provided the major portion of the altimetry-derived gravity anomalies. Fill-in values were computed by augmenting low-resolution gravitational information, by the high-resolution gravitational information implied by the topography.

- *Information Content.* The observables within the above four categories contain information not only about the gravitational field, but also about numerous other effects. Some of these effects are of interest in their own right (e.g., the DOT information contained within altimetric SSH), while others represent, at least as far as gravitational modeling is concerned, systematic noise (e.g., the nonconservative forces acting on a satellite). In either case, effective incorporation of a particular data type into the combination solution requires precise modeling and optimal estimation of all the effects and signals contained within the observable. Otherwise, the estimated gravitational model can be severely corrupted by the mis-modeled (or unmodeled) systematic effects.
- *Spectral Sensitivity Overlap.* The development of a GGM through a *least squares adjustment* procedure combining different data types is meaningful, provided that the data used in the adjustment share some common degree of sensitivity to the gravitational signal over a certain portion of its spectrum (a range of harmonic degrees). Otherwise, there is little "adjustment" being performed to data representing disjoint spectral bands. For example, pre-GRACE satellite-only models have a narrow spectral sensitivity overlap with models recovered from surface gravity data alone. This complicated considerably the problem of optimal combination of these two data types. On the other hand, this also means that setting up and inverting extremely large linear systems corresponding to very high degree models may not be necessary, if a single data type (e.g., a complete global equiangular grid of gravity anomalies) uniquely determines the higher-degree portion of such a GGM.
- *Relative Weighting.* The optimal estimation of a GGM depends critically on the optimality of the relative weights assigned to the different data types. Considering the numerous sources of data that may be involved, this may require the solution to a very large *component of variance* estimation problem, complicated further by the fact that the extraction of gravitational information from satellite orbit observations is a strongly nonlinear problem. Although approximate solutions to this relative weight estimation problem have been used with considerable success (Lerch, 1991), many times the experience and intuition of the model developer(s) guide the selection of appropriate data weights more than anything else.

The recent gravity-mapping satellite missions

The satellite data used for the development of all GGMs published by the end of the twentieth century represent tracking of "targets of opportunity," that is, of spacecraft designed and equipped with instrumentation for applications other than the mapping of the gravitational field from space. As a result of three satellite missions, this situation has changed dramatically during the last decade. These three missions are CHAMP (Reigber et al., 1996), GRACE (GRACE, 1998), and GOCE (ESA, 1999). Table 1 summarizes the main characteristics of these missions.

A comprehensive discussion regarding the concepts involved in these three mission scenarios can be found in (ESA, 1999, Sect. 2.3). Mapping of the gravitational field from space requires missions that adhere as much as possible to the following fundamental design constraints:

Gravity, Global Models, Table 1 Main characteristics of three gravity-mapping satellite missions

Mission	Status	Orbit	Mission objective	Instrumentation, tracking, and comments
CHAMP	Launched on 7/15/2000 active	Alt. = 450 km e ≈ 0.004 i = 87°	Gravity and magnetic fields Atmospheric limb Sounding Ionosphere sounding	3-axis STAR accelerometer GPS and SLR altitude will decay from 450 km (BOL) to 300 km (EOL)
GRACE	Launched on 3/17/2002 active	Alt. = 485 km e ≈ 0.001 i = 89°	Gravity field and its temporal variation	3-axis accelerometers (1 per s/c) GPS and SLR K-band inter-satellite ranging between the 2 s/c
GOCE	Launched on 3/17/2009 active	Alt. ≈ 250 km i = 96.7° Sun-synchronous	Gravity field (Especially static)	Six 3-axis accelerometers forming the gradiometer GPS/GLONASS and SLR

- Uninterrupted tracking in three spatial dimensions.
- Measurement or compensation of the effects of nongravitational forces.
- Orbital altitude as low as possible, to enhance sensitivity to the gravitational signal, and inclination as high as possible, to permit (near) global coverage.
- Counteraction of the field's attenuation at altitude through the measurement of derivatives of the potential.

All three missions have in common the high–low Satellite-to-Satellite Tracking component (SST-*hl*) from the GPS (and GLONASS in the case of GOCE) constellation, and the measurement of nongravitational forces by the onboard accelerometers. These data permit highly accurate orbit determination for all three missions, and in addition may enhance the gravitational field determination at very long wavelengths (very low degrees). In addition to that, GRACE involves the continuous measurement of the range between two identical satellites that "chase each other," which constitutes a low-low SST formation (SST-*ll*). GOCE's accelerometer array on the other hand provides the measurements necessary to determine the gravitational *tensor* (i.e., the 3 × 3 matrix of second-order spatial derivatives of the gravitational potential) at altitude. GOCE is unique in the sense that it will provide boundary data at altitude covering the entire Earth, *except* for two polar cap areas of ∼6.7° radius (due to the satellite's inclination). The data acquired by each of these three missions result in different sensitivities to the gravitational spectrum. Simulation studies examining the geopotential recovery attainable from these (and other) mission scenarios have been reported, for example, by Sneeuw and Ilk (1997). When considering the data from these satellite missions two main questions arise:

1. What is the optimal way of analyzing their data?
2. What is the optimal way of combining their data with existing data, for example, from surface gravimetry and from satellite altimetry, in order to develop high-degree combination gravitational models?
    1. *Data Analysis.* In the case of CHAMP, the gravitational information is extracted from the analysis of the perturbations of a low Earth orbiter, in a fashion similar to other preexisting satellite missions. However, CHAMP's low orbit, in conjunction with the accelerometer data and with the availability of nearly global tracking data coverage, enabled *for the first time* the determination of an accurate long wavelength global gravitational model from a *single* satellite's data. Indicative of this "new state of affairs" was the fact that a very preliminary solution (complete to degree and order 91) was already developed based on a *single* month's worth of CHAMP data *only* and was presented during the 2001 meeting of the International Association of Geodesy (IAG) by Reigber et al. (2001). Although significantly better models that include GRACE data have by now surpassed considerably this preliminary solution, it served as a good example of the improvements that were to follow.
    Compared to CHAMP, GRACE added the SST-*ll* component, which permitted higher-resolution gravitational information to be extracted from the analysis of the orbital perturbation *differences* along the line of sight of the two low orbiting satellites. One can use traditional orbit perturbation analysis methods to process the GRACE data and derive a GGM, for example, in spherical harmonics. GGM01S (Tapley et al., 2004) and GGM02S (Tapley et al., 2005) were estimated following such a procedure. This analysis scenario, albeit costly, is within current computational capabilities for models extending to degree and order 180 or so. Geodesists have also considered alternative analysis methods for GRACE-type missions (e.g., Wolff, 1969; Colombo, 1981b; Jekeli, 1999a; Rowlands et al., 2002). Such methods provide higher computational efficiency at the cost of introducing certain approximations. Setting aside for a moment the details of optimal GRACE data processing, it is important to recognize here the quantum leap that has been accomplished with the GRACE mission. Approximately 14 months of GRACE data *alone* have been used to develop GGM02S, whose cumulative global Root Mean Square (RMS) geoid undulation error to degree 70 is less than ±1 cm (Tapley et al., 2005, Figure 2). By comparison, the cumulative geoid undulation global RMS error to degree 70 for EGM96, which required the combination of data from tens of satellites, along with surface gravity

and satellite altimetry, was ±19 cm (Lemoine et al., 1998, Table 10.3.2-1).
As far as GOCE is concerned, numerous investigations of the various aspects of its data analysis and of the development of a GGM from them can be found in (ESA, 2000). One particular issue that receives increased attention relates to the polar gaps and their impact on analysis schemes that exploit regularity and completeness in the data coverage.

2. *Data Combination.* While the existing satellite gravity-mapping missions (CHAMP and especially GRACE) have already delivered (or promise to deliver as in the case of GOCE) quantum leaps in the accuracy and resolution of the *satellite-only* gravitational models, there is still a need to combine that information with terrestrial gravity and satellite altimetry data in an optimal fashion. This is required so that a "seamless" extension of the gravitational spectrum can be achieved, taking advantage of the rich high frequency content of the surface and altimetric data. The higher resolution of GRACE-based and GOCE-based *satellite-only* models will significantly increase the spectral overlap with surface gravity and altimetry. This will enable for the first time the estimation of high-resolution models of the DOT as well as the significant reduction of systematic errors present in surface gravity data (see also [Pavlis, 2000]).

Beyond the sensitivity of satellite data

Due to the attenuation of the gravitational signal with increasing altitude, there is obviously a limit to the gravitational information that can be extracted from spaceborne sensors. This implies a limit to the resolution (maximum degree) of satellite-only models. Spectral estimation indicates that for GRACE, the signal-to-noise ratio becomes one around degree 170, while for GOCE the corresponding degree is not expected to exceed 300. Observations made on the Earth's surface (or on airplanes flying at low altitudes) can extend the resolution of gravitational models considerably. Surface and airborne data like *gravity anomalies*, *gravity disturbances*, etc. are therefore capable of supporting the development of much higher resolution gravitational models than those developed based on satellite data only. One way of developing such high-degree models involves first the formation of a regular grid of spatially averaged values of some functional of the field that completely covers the globe. These values represent data derived on the basis of other primary observables, using techniques like *Least Squares Collocation* (LSC) (Moritz, 1980). Surface gravity anomalies are a suitable choice for such a functional, both from a spectral sensitivity and from an availability viewpoint. Very high-degree and order spherical harmonic models are then developed from the analysis of such global anomaly grids, using efficient harmonic analysis techniques like those put forward by Colombo (1981a). Wenzel (1998, 1999) reported such expansions complete to degree and order 1800. Pavlis et al. (2005) reported the development and evaluation of the PGM2004A preliminary gravitational model, which combined GRACE information with a global set of $5' \times 5'$ gravity anomalies, and extended to maximum degree and order 2,160. The final product of that modeling effort was EGM2008 which also extends to degree 2,160 (Pavlis et al., 2008). EGM2008 was completed and released to the public in April 2008.

State-of-the-art global gravitational modeling

We briefly outline the main aspects of the development of two global gravitational models, representative of the state-of-the-art, given the data that were available, at the time of each model's development: EGM96 (Lemoine et al., 1998), which represents the state-of-the-art before the availability of data from the CHAMP and GRACE missions, and EGM2008 (Pavlis et al., 2008), which represents currently (March 2010) the model with the highest resolution and accuracy, prior to the anticipated availability of data from GOCE. A comparison of the approaches that were followed for the development of these two models reveals the critical changes in model development, which were brought about by the availability of GRACE data on one hand, and of high quality $5' \times 5'$ area-mean gravity anomalies (from the combination of terrestrial and altimetry-derived data sources) on the other.

In theory, the estimation of a GGM complete to some (arbitrary) high degree and order, from the combination of various (heterogeneous) data, could be carried out as follows:

(a) Form separate normal equations from each individual data type, to a maximum degree and order that corresponds to the resolution of the available data and their sensitivity to the gravitational signal.
(b) Treat satellite altimeter data as "direct tracking" observations, that is, ranges from the spacecraft to the ocean surface whose upper endpoint senses (through the orbit dynamics) attenuated gravitational signals (static and time varying), while their lower end point senses the combined effects of geoid undulation, DOT as well as tides and other time varying effects, without any attenuation. In this manner, altimeter data contribute to the estimation of the satellite's orbit, as well as the estimation of the DOT and of the potential coefficients.
(c) Combine the various normal equations (with appropriate relative weights) and invert the resulting system, to estimate the combination solution to its high degree, along with its *full* error covariance matrix.

Such an "ideal" approach would permit the most rigorous modeling of the observables and would allow the greatest flexibility in terms of data weighting. A combination solution to degree 360, if performed as outlined above, would require the formation of full (symmetric) normal equation matrices (from satellite altimetry and surface gravimetry) for approximately 130,000

parameters (considering *only* the gravitational potential coefficients). For maximum degree 2,160, there would be approximately 4.7 million such parameters involved. Such computational tasks are beyond our present computational capabilities. Therefore, at present, one may choose between, or combine, two main solution strategies to attack the problem:

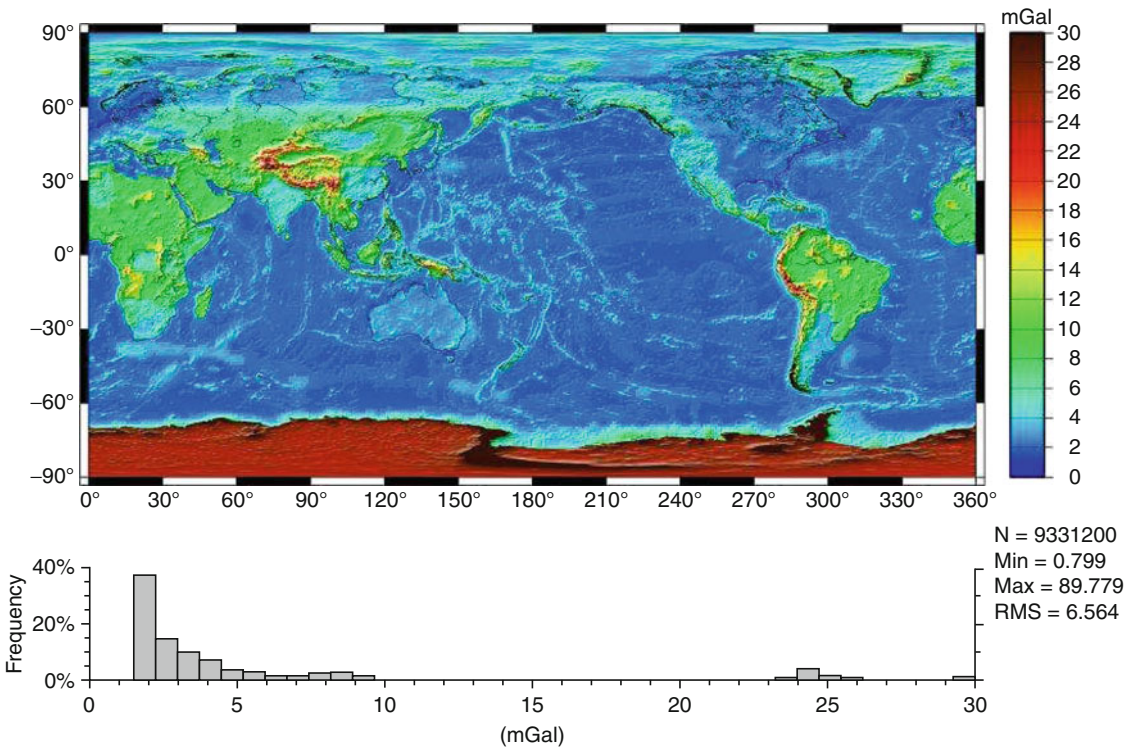
- *Solution Strategy (A)* Apply the "ideal" estimation strategy outlined above, to obtain a combination solution for the lower degree part of the field, up to a maximum degree that is computationally manageable. Apart from reasons of computational feasibility, this maximum degree should enable the appropriate modeling of the gravitational signal contained in the currently available satellite tracking data. Furthermore, since the DOT signal is of long wavelength nature, the benefits of "direct" altimetry are almost entirely retained here. To avoid aliasing effects however, the contribution to the altimetry and surface gravity data from the coefficients beyond the solved-for degree has to be filtered out of the data prior to the normal equation formation. This may be done using a preexisting high-degree solution (Pavlis, 1988; Denker and Rapp, 1990). Herein, we will refer to this type of solution as the *comprehensive* low-degree combination model. The obvious shortcomings of this approach are the relatively low maximum attainable degree (approximately 250 at present) and its computational demands. Some models developed using this approach (or similar ones) include JGM-1 and JGM-2 (Nerem et al., 1994) and JGM-3 (Tapley et al., 1996), the part of EGM96 up to degree and order 70 (see Lemoine et al., 1998, Chapt. 7), and EIGEN-GL04C (Förste et al., 2008).
- *Solution Strategy (B)* Consider that one is willing to make the following two approximations:
  1. The orbits of the altimeter satellites, whose data are included in the combination solution, are perfectly known (at least radially). This approximation is justifiable if one is working with altimeter satellites supported by T/P-class precise orbit determination. Moreover, after the availability of GRACE-based gravitational models for precise orbit determination of altimeter satellites, errors arising from gravitational model inaccuracies do not dominate the orbit error budget of these satellites. Errors due to, for example, mis-modeling of nonconservative forces acting on the spacecraft are likely to be more significant at present. In this regard, to allow the orbits of altimeter satellites to contribute (through their dynamics) to the determination of gravitational parameters within a combination solution may not be a desirable approach presently, because the effects of orbit errors of nongravitational origin could corrupt the solved-for gravitational parameters.
  2. The DOT is known a priori, for example, from an Ocean Circulation Model (OCM) or from a previous low-degree comprehensive combination solution.

The implication of (1) is that satellite altimetry does not have to be treated as "direct" tracking anymore, which simplifies the problem considerably, since now orbit dynamics are not involved in the altimeter data processing. One is left with a "surface" problem, where the geoid (N) and the DOT (ζ) signals have to be separated, given the "observed" SSH (h), which is their sum. If in addition, the approximation (2) is introduced, then altimetry contributes to the combination solution "observed" geoid heights (N) over (parts of) the ocean.

In addition to the above two approximations, a key issue here is that altimetric information may also be provided in a *gridded* form. This is possible through the use of a Mean Sea Surface (MSS), obtained from multiple altimetric missions. The success of T/P has significantly improved the accuracy of MSS data sets (especially at long wavelengths). This is accomplished by adjusting the SSH data from other altimetric missions (e.g., ERS-1, ERS-2, GEOSAT, GEOSAT-Follow-On), to the surface defined by T/P, using crossover minimization techniques. Such MSS data sets have been developed by, for example, Yi (1995); Kim et al. (1995); Anzenhofer et al. (1996); Wang (2001), and more recently by Andersen and Knudsen (2009). One may also have available in *gridded* form, altimetry-derived gravity anomaly values. Such values have been estimated using various techniques, on an ocean-wide basis by, for example, Rapp and Basic (1992); Andersen and Knudsen (1998), Andersen, Knudsen and Berry (2010), Trimmer and Manning (1996) and Sandwell and Smith (1997, 2009) among others.

The two simplifying approximations discussed above and particularly the availability of altimetric information in gridded form (especially in the form of gravity anomalies) make applicable an alternative class of high-degree combination solution techniques. These techniques combine the satellite-only information, with potential coefficient information obtained from the analysis of *complete*, regular grids of functional(s) of the disturbing potential (e.g., N, Δg), and are based on the highly efficient harmonic analysis algorithms originally studied and put forward by Colombo (1981a). These algorithms exploit the regularity of the data grids and the symmetry properties of Legendre and trigonometric (sine/cosine) functions. Inherent in these algorithms is the use of Fast Fourier Transform (FFT) techniques to process data arrays residing over latitude bands that are symmetric with respect to the equator, in a highly efficient manner. Estimators of this type are the (simple) Numerical Quadrature (NQ), the Block-Diagonal (BD) least-squares adjustment, and the Optimal Estimation (OE) technique. Models developed using the NQ approach include OSU86E/F (Rapp and Cruz, 1986a) and OSU89A/B (Rapp and Pavlis, 1990). BD techniques of varying sophistication have been used to develop GPM2 (Wenzel, 1985), DGFI92A (Gruber and Bosch, 1992), GFZ95A (Gruber et al., 1996), and EGM2008 (Pavlis et al., 2008). OE was used to develop the OSU86C/D models (Rapp and Cruz, 1986b).

The two solution strategies (A) and (B) discussed above have their respective advantages and disadvantages. Due to computational limitations, the development of a GGM to spherical harmonic degree 360 or higher, currently requires the implementation of either a combination of the two strategies (as it was done for EGM96 and EIGEN-GL04C), or the implementation of strategy (B), as it was done for OSU89A/B and EGM2008. EGM96 (Lemoine et al., 1998) employed a comprehensive solution to degree 70, augmented by a BD solution from degree $n = 71$ to degree $n = 359$, while the $n = 360$ coefficients (which cannot be estimated using the BD technique) were obtained from an NQ model. The main reason for the choice of this solution strategy was the fact that the satellite-only component of EGM96 (denoted EGM96S), was accompanied by a variance-covariance matrix that was fully occupied. This was due to the fact that the heterogeneous tracking data from the (approximately) 40 satellites that were used to derive EGM96S to degree and order 70, were incapable of de-correlating adequately the spherical harmonic coefficients up to this degree and order. Therefore, in order to preserve the integrity of the least-squares adjustment used to combine EGM96S with the surface gravity and altimetry data, one had to consider the satellite-only normal equations in their complete (fully occupied) form. Any block-diagonal approximation of these normal equations would result in estimation errors that could not be tolerated. This situation changed dramatically with GRACE, where the corresponding normal equation matrix, due to the global coverage and uniform GRACE data accuracy could be safely approximated by a block-diagonal matrix, without significant loss of accuracy. In addition, after the availability of satellite-only models from GRACE, there is really no need to incorporate altimeter data into the combination solution in the form of "direct" tracking. Instead, a preliminary model from GRACE and an MSS can be used to estimate a preliminary model of the DOT. This could then be used to correct the SSH of the MSS and estimate an ocean-wide set of altimetry-derived gravity anomalies. These can be "merged" with terrestrial and airborne data to form a complete global grid (see also Figure 5). The gravitational information implied by these gridded data could then be combined (in a least-squares adjustment) with the satellite-only model from GRACE, to derive the combination solution, up to the high degree (2,160), in a single step. The entire process may be iterated, using the high-degree combination solution to derive the next estimate of the DOT, and so on. This is essentially the approach that was used to develop EGM2008 (Pavlis et al., 2008). Two iterations of the estimation of the altimetry-derived gravity anomalies were performed, which was sufficient for the process to converge. Despite its iterative nature, this approach permits the development of very high-degree combination solutions in an efficient manner, and moreover



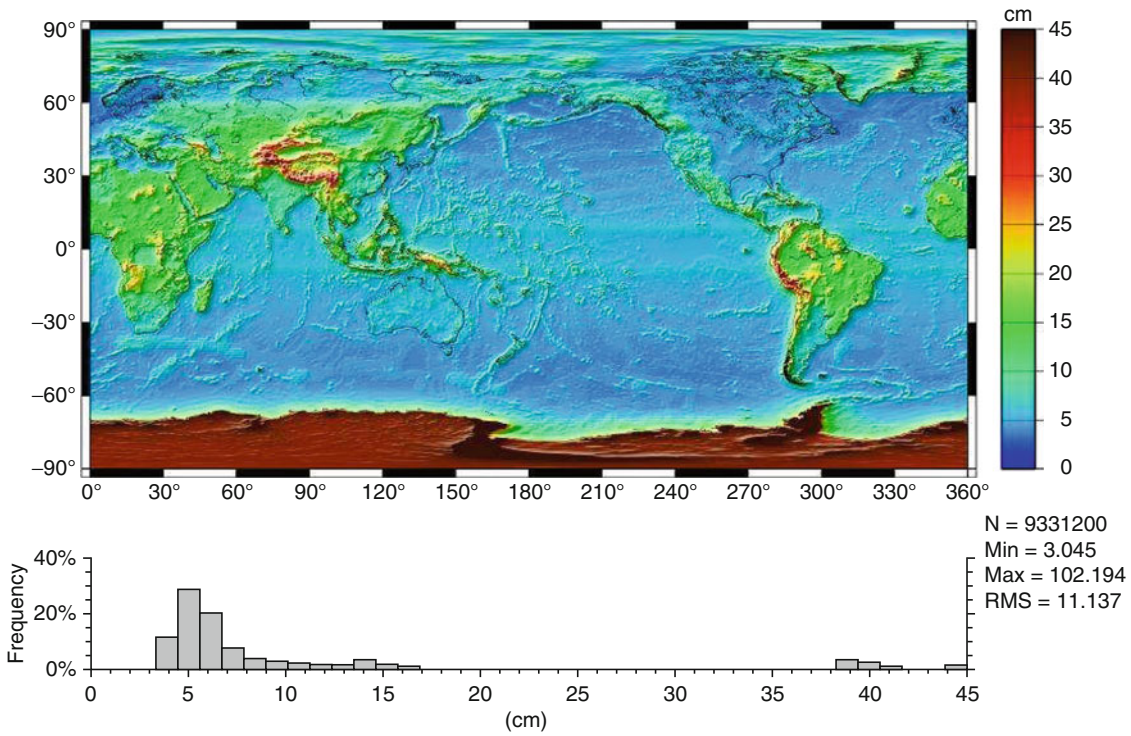
Gravity, Global Models, Figure 6 Propagated error estimates (in mGal) of the free-air gravity anomalies computed using the EGM2008 model. Notice the large error estimates over Antarctica and over mountainous areas (Himalaya, Andes), due to the lack of accurate and detailed data.

in a single adjustment step, thereby avoiding the "piecewise" nature of models like EGM96.

Model evaluation and accuracy assessment

Of central importance to the development of a GGM are the evaluation of its performance, and the estimation of reliable error measures to quantify its accuracy. The former requires the comparison of model-derived quantities to corresponding values obtained *independently*. Several different data types, withheld from a model's development, are used for this purpose. These independent data have different spectral sensitivities and/or occupy different geographic regions. Tests that are usually employed here include satellite orbit determination and comparisons with tracking data, comparisons with geoid undulations obtained from GPS and leveling data, comparisons with deflections of the vertical obtained from astrogeodetic techniques (Jekeli, 1999b), comparisons employing altimetry and general ocean circulation models, etc. A useful practice, introduced during the development of EGM96, and used also during the development of EGM2008, is to invite a voluntary evaluation working group, independent of a model's developers, that evaluates and provides feedback to the model's developers regarding candidate preliminary solutions, as well as the final outcome from a modeling effort, in a manner as objective as possible. These groups work under the auspices of the International Association of Geodesy (IAG) and upon completion of a certain evaluation effort they report their findings in IAG-sponsored publications, which can be accessed freely by the public. In the case of EGM2008, the results from such an evaluation by 25 different international investigating teams are reported in *Newton's Bulletin No. 4*, which is jointly published by the Bureau Gravimétrique International (BGI) and the International Geoid Service (IGeS).

The estimation of reliable errors that should accompany a model's derived quantities is also very important. These errors are geographically dependent, reflecting the different quality of the data used to develop the model, as a function of geographic area. For models like EGM2008, which involve a very large number of parameters (approximately 4.7 million), one cannot perform such error propagation using variance-covariance matrices and linear algebra methods, due to the excessively large sizes of the matrices involved. Instead, Pavlis and Saleh (2005) introduced and implemented an alternative approach that employs global convolutions with band-limited kernels. This technique is capable of providing very reliable error estimates, with geographic specificity, in a very efficient manner. Figures 6 and 7 illustrate



Gravity, Global Models, Figure 7 Propagated error estimates (in centimeters) of the geoid undulations computed using the EGM2008 model. The larger values around the equator are due to the lower density of GRACE data at low latitudes.

geographically the propagated errors in free-air gravity anomalies and in geoid undulations, implied by EGM2008 to degree 2,160.

Summary

Accurate knowledge of the gravitational potential of the Earth, on a global scale and at high resolution, is a fundamental prerequisite for various geodetic, geophysical, and oceanographic investigations and applications. Over the last 50 years, continuing improvements and refinements to the basic theory have been paralleled by the availability of more accurate and complete data and by dramatic improvements in the computational resources available for numerical modeling studies. These advances have brought the state-of-the-art from the early spherical harmonic models of degree 8 (Zhongolovich, 1952) to solutions like EGM2008 that extends to degree 2,160. The availability of satellite missions like CHAMP, GRACE, and GOCE, dedicated to the mapping of the gravitational field from space, is enabling the determination of the long wavelength features of the field with unprecedented accuracy. Nevertheless, the determination of the fine details of the field still relies on the availability of terrestrial and airborne data, of high accuracy and sufficient geographic coverage and density. In this regard, several areas of the Earth, most notably Antarctica, await more comprehensive gravimetric surveying.

Bibliography

Andersen, O. B., and Knudsen, P., 1998. Global marine gravity field from the ERS-1 and GEOSAT geodetic mission altimetry. *Journal of Geophysical Research*, **103**(C4), 8129–8137.

Andersen, O. B., and Knudsen, P., 2009. DNSC08 mean sea surface and mean dynamic topography models. *Journal of Geophysical Research*, **114**(C11), C11001.

Andersen, O. B., Knudsen, P., and Berry, P. A. M., 2010. The DNSC08GRA global marine gravity field from double retracked satellite altimetry. *Journal of Geodesy*, **84**(3), 191–199, doi:10.1007/s00190-009-0355-9.

Anzenhofer, M., Gruber, T., and Rentsch, M., 1996. Global high resolution mean sea surface based on ERS-1 35- and 168-day cycles and TOPEX data. In Rapp, R. H., Cazenave, A. A., and Nerem, R. S. (eds.), *Global Gravity Field and Its Temporal Variations*. Berlin: Springer-Verlag. IAG Symposia, Heidelberg Vol. 116.

Baker, E. M., 1988. *A finite element model of the Earth's anomalous gravitational potentia.*, Rep. 391, Department of Geodetic Science and Survey, Ohio State University. Columbus.

Bouman, J., 1997. *A survey of global gravity models, Delft Institute for Earth-Oriented Space Research. Rep. No. 97.1*, Delft Institute of Technology. The Netherlands.

Colombo, O. L., 1981a. *Numerical methods for harmonic analysis on the sphere.* Rep. 310, Department of Geodetic Science and Survey, Ohio State University, Columbus.

Colombo, O. L., 1981b. *Global geopotential modelling from satellite-to-satellite tracking.* Rep. 317, Department of Geodetic Science and Survey, Ohio State University, Columbus.

Denker, H., and Rapp, R. H., 1990. Geodetic and oceanographic results from the analysis of 1 year of geosat data. *Journal of Geophysical Research*, **95**(C8), 13151–13168.

ESA SP-1233 (1) 1999 – The Four Candidate Earth Explorer Core Missions – Gravity Field and Steady-State Ocean Circulation Mission, ESA Publications Division, ESTEC, Noordwijk, The Netherlands.

ESA 2000 – From Eötvös to mGal – Final report, ESA/ESTEC Contract No. 13392/98/NL/GD, Graz, Austria.

Förste, C., Schmidt, R., Stubenvoll, R., Flechtner, F., Meyer, U., König, R., Neumayer, H., Biancale, R., Lemoine, J.-M., Bruinsma, S., Loyer, S., Barthelmes, F., and Esselborn, S., 2008. The GeoForschungsZentrum Potsdam/Groupe de Recherche de Gèodésie Spatiale satellite-only and combined gravity field models: EIGEN-GL04S1 and EIGEN-GL04C. *Journal of Geodesy*, **82**(6), 331–346, doi:10.1007/s00190-007-0183-8.

Fu, L.-L., Christensen, E., Yamarone, C., Jr., Lefebvre, M., Ménard, Y., Dorrer, M., and Escudier, P., 1994. TOPEX/POSEIDON: Mission overview. *Journal of Geophysical Research*, **99**(C12), 24369–24381.

Ganachaud, A., Wunsch, C., Kim, M.-C., and Tapley, B., 1997. Combination of TOPEX/POSEIDON data with a hydrographic inversion for determination of the oceanic general circulation and its relation to geoid accuracy. *Geophysical Journal International*, **128**, 708–722.

GRACE 1998. – Gravity Recovery and Climate Experiment: Science and Mission Requirements Document, revision A, JPLD-15928, NASA's Earth System Science Pathfinder Program.

Gruber, T., and Bosch, W., 1992. *A New 360 Gravity Field Model, Presented at the XVII General Assembly of the European Geophysical Society.* Scotland: Edinburgh.

Gruber, T., Anzenhofer, M., and Rentsch, M., 1996. The 1995 GFZ high resolution gravity model. In Rapp, R. H., Cazenave, A. A., and Nerem, R. S. (eds.), *Global Gravity Field and Its Temporal Variations*. Berlin: Springer-Verlag. IAG Symposia, Vol. 116.

Heck, B., 1990. An evaluation of some systematic error sources affecting terrestrial gravity anomalies. *Bulletin Géodésique*, **64**, 88–108.

Heiskanen, W. A., and Moritz, H., 1967. *Physical Geodesy.* San Francisco: WH Freeman.

Jekeli, C., 1999a. The determination of gravitational potential differences from satellite-to-satellite tracking. *Celestial Mechanics & Dynamical Astronomy*, **75**(2), 85–100.

Jekeli, C., 1999b. An analysis of vertical deflections derived from high-degree spherical harmonic models. *Journal of Geodesy*, **73**(1), 10–22, doi:10.1007/s001900050213.

Jekeli, C., 2005. *Spline representations of functions on a sphere for geopotential modeling.* Rep. 475, Department of Civil and Environmental Engineering and Geodetic Science, Ohio State University, Columbus.

Kenyon, S. C., and Pavlis, N. K., 1996. The development of a global surface gravity data base to be used in the joint DMA/GSFC geopotential model. In Rapp, R. H., Cazenave, A. A., and Nerem, R. S. (eds.), *Global Gravity Field and Its Temporal Variations*. Berlin: Springer-Verlag. IAG Symposia, Vol. 116.

Kim, M. -C., Tapley, B. D., Shum, C. -K., and Ries, J. C., 1995. Center for Space Research mean sea surface model, presented at the TOPEX/POSEIDON Working Team Meeting, Pasadena, California.

Lemoine, F. G., Kenyon, S. C., Factor, J. K., Trimmer, R. G., Pavlis, N. K., Chinn, D. S., Cox, C. M., Klosko, S. M., Luthcke, S. B., Torrence, M. H., Wang, Y. M., Williamson, R. G., Pavlis, E. C., Rapp, R. H., and Olson, T. R., 1998. *The development of the joint NASA GSFC and the National Imagery and Mapping Agency (NIMA) geopotential model EGM96*, NASA Tech. Publ. TP-1998-206861, pp 575.

Lerch, F. J., 1991. Optimum data weighting and error calibration for estimation of gravitational parameters. *Bulletin Géodésique*, **65**, 44–52.

Lerch, F. J., Klosko, S. M., Laubscher, R. E., and Wagner, C. A., 1979. Gravity model improvement using GEOS-3 (GEM9 and 10). *Journal of Geophysical Research*, **84**(B8), 3897–3916.

Meissl, P., 1981. *The use of finite elements in physical geodesy.* Rep. 313, Department of Geodetic Science and Survey, Ohio State University, Columbus.

Moritz, H., 1980. *Advanced Physical Geodesy.* Karlsruhe: Herbert Wichmann Verlag. Federal Republic of Germany.

Nerem, R. S., Lerch, F. J., Marshall, J. A., Pavlis, E. C., Putney, B. H., Tapley, B. D., Eanes, R. J., Ries, J. C., Schutz, B. E., Shum, C. K., Watkins, M. M., Klosko, S. M., Chan, J. C., Luthcke, S. B., Patel, G. B., Pavlis, N. K., Williamson, R. G., Rapp, R. H., Biancale, R., and Nouel, F., 1994. Gravity model development for TOPEX/POSEIDON: Joint gravity models 1 and 2. *Journal of Geophysical Research*, **99**(C12), 24421–24447.

Nerem, R. S., Jekeli, C., and Kaula, W. M., 1995. Gravity field determination and characteristics: Retrospective and prospective. *Journal of Geophysical Research*, **100**(B8), 15053–15074, doi:10.1029/94JB03257.

Pavlis, N. K., 1988. *Modeling and estimation of a low degree geopotential model from terrestrial gravity data*. Rep. 386, Department of Geodetic Science and Survey, Ohio State University, Columbus.

Pavlis, N. K., 2000. On the modeling of long wavelength systematic errors in surface gravimetric data. In Schwarz, K.-P. (ed.), *Geodesy Beyond 2000 – The Challenges of the First Decade*. Berlin: Springer-Verlag. IAG Symposia, Vol. 121.

Pavlis, N. K., and Rapp, R. H., 1990. The development of an isostatic gravitational model to degree 360 and its use in global gravity modelling. *Geophysical Journal International*, **100**, 369–378.

Pavlis, N. K., Holmes, S. A., Kenyon, S. C., Schmidt, D., and Trimmer, R., 2005. A preliminary gravitational model to degree 2160. In Jekeli, C., Bastos, L., and Fernandes, J. (eds.), *Gravity, Geoid and Space Missions*. Berlin: Springer-Verlag. IAG Symposia, Vol. 129.

Pavlis, N. K., and Saleh, J., 2005. Error propagation with geographic specificity for very high degree geopotential models. In Jekeli, C., Bastos, L., and Fernandes, J. (eds.), *Gravity, Geoid and Space Missions*. Berlin: Springer-Verlag. IAG Symposia, Vol. 129.

Pavlis, N. K., Factor, J. K., and Holmes, S. A., 2007. Terrain-related gravimetric quantities computed for the next EGM. In *Gravity Field of the Earth*, *Proceedings of the 1st International Symposium of the International Gravity Field Service (IGFS)*, Istanbul, Turkey, Harita Dergisi, Special Issue 18, pp 318–323.

Pavlis, N. K., Holmes, S. A., Kenyon, S. C., and Factor, J. K., 2008. An Earth Gravitational Model to Degree 2160: EGM2008, presented at the 2008 General Assembly of the European Geosciences Union, Vienna, Austria, April 13–18.

Rapp, R. H., 1989. The decay of the spectrum of the gravitational potential and the topography of the Earth. *Geophysical Journal International*, **99**, 449–455.

Rapp, R. H., 1993. Use of altimeter data in estimating global gravity models. In Rummel, R., and Sansò, F. (eds.), *Satellite Altimetry in Geodesy and Oceanography.* Berlin: Springer-Verlag, pp. 373–417.

Rapp, R. H., 1997. Use of potential coefficient models for geoid undulation determinations using a spherical harmonic representation of the height anomaly/geoid undulation difference. *Journal of Geodesy*, **71**, 282–289.

Rapp, R. H., 1998. Past and future developments in geopotential modeling. In Forsberg, R., Feissel, M., and Dietrich, R. (eds.), *Geodesy on the Move – Gravity, Geoid, Geodynamics and Antarctica*. Berlin: Springer-Verlag. IAG Symposia, Vol. 119.

Rapp, R. H., and Cruz, J. Y., 1986a. *Spherical harmonic expansions of the Earth's gravitational potential to degree 360 using 30′ mean anomalies.* Rep. 376, Department of Geodetic Science and Survey, Ohio State University, Columbus.

Rapp, R. H., and Cruz, J. Y., 1986b. *The representation of the Earth's gravitational potential in a spherical harmonic expansion to degree 250.* Rep. 372, Department of Geodetic Science and Survey, Ohio State University, Columbus.

Rapp, R. H., and Pavlis, N. K., 1990. The development and analysis of geopotential coefficient models to spherical harmonic degree 360. *Journal of Geophysical Research*, **95** (B13), 21885–21911.

Rapp, R. H., Wang, Y. M., and Pavlis, N. K., 1991. *The Ohio State 1991 geopotential and sea surface topography harmonic coefficient models.* Rep. 410, Department of Geodetic Science and Survey, Ohio State University, Columbus.

Rapp, R. H., and Balasubramania, N., 1992. *A conceptual formulation of a world height system.* Rep. 421, Department of Geodetic Science and Survey, Ohio State University, Columbus.

Rapp, R. H., and Basic, T., 1992. Oceanwide gravity anomalies from GEOS-3, SEASAT and GEOSAT altimeter data. *Geophysical Research Letters*, **19**(19), 1979–1982.

Reigber, C., Bock, R., Förste, C., Grunwaldt, L., Jakowski, N., Lühr, H., Schwintzer, P., and Tilgner, C., 1996. CHAMP Phase-B Executive Summary, GFZ, STR96/13.

Reigber, C., et al., 2001. Global gravity field recovery with CHAMP, presented at the 2001 IAG Scientific Assembly, Budapest, Hungary, September 2–7.

Rizos, C., 1979. An efficient computer technique for the evaluation of geopotential from spherical harmonic models. *Australian Journal of Geodesy, Photogrammetry, and Cartography*, **31**, 161–170.

Rowlands, D. D., Ray, R. D., Chinn, D. S., and Lemoine, F. G., 2002. Short-arc analysis of intersatellite tracking data in a gravity mapping mission. *Journal of Geodesy*, **76**(6–7), 307–316, doi:10.1007/s00190-002-0255-8.

Sandwell, D. T., and Smith, W. H. F., 1997. Marine gravity anomaly from Geosat and ERS 1 satellite altimetry. *Journal of Geophysical Research*, **102**(B5), 10039–10054.

Sandwell, D. T., and Smith, W. H. F., 2009. Global marine gravity from retracked Geosat and ERS-1 altimetry: ridge segmentation versus spreading rate. *Journal of Geophysical Research*, **114**, B01411, doi:10.1029/2008JB006008.

Schwarz, K.-P., Sideris, M. G., and Forsberg, R., 1987. Orthometric heights without leveling. *Journal of Surveying Engineering*, **113**(1), 28–40.

Sneeuw, N., and Ilk, K. H., 1997. The status of spaceborne gravity field mission concepts: a comparative simulation study. In Segawa, J., Fujimoto, H., and Okubo, S. (eds.), *Gravity, Geoid and Marine Geodesy.* Berlin: Springer-Verlag. IAG Symposia, Vol. 117.

Sünkel, H., 1981. *Point mass models and the anomalous gravitational field.* Rep. 328, Department of Geodetic Science and Survey, Ohio State University, Columbus.

Sünkel, H., 1983. *The generation of a mass point model from surface gravity data.* Rep. 353, Department of Geodetic Science and Survey, Ohio State University, Columbus.

Sünkel, H., 1984. *Splines: their equivalence to collocation.* Rep. 357, Department of Geodetic Science and Survey, Ohio State University, Columbus.

Tapley, B. D., Watkins, M. M., Ries, J. C., Davis, G. W., Eanes, R. J., Poole, S. R., Rim, H. J., Schutz, B. E., Shum, C. K., Nerem, R. S., Lerch, F. J., Marshall, J. A., Klosko, S. M., Pavlis, N. K., and Williamson, R. G., 1996. The joint gravity model–3. *Journal of Geophysical Research*, **101**(B12), 28029–28049.
Tapley, B. D., Bettadpur, S., Watkins, M., and Reigber, C., 2004. The gravity recovery and climate experiment: mission overview and early results. *Geophysical Research Letters*, **31**, L09607, doi:10.1029/2004GL019920.
Tapley, B., Ries, J., Bettadpur, S., Chambers, D., Cheng, M., Condi, F., Gunter, B., Kang, Z., Nagel, P., Pastor, R., Pekker, T., Poole, S., and Wang, F., 2005. GGM02 – An improved earth gravity field model from GRACE. *Journal of Geodesy*, **79**(8), 467–478, doi:10.1007/s00190-005-0480-z.
Trimmer, R. G., and Manning, D. M., 1996. The altimetry derived gravity anomalies to be used in computing the joint DMA/NASA earth gravity model. In Rapp, R. H., Cazenave, A. A., and Nerem, R. S. (eds.), *Global Gravity Field and Its Temporal Variations*. Berlin: Springer-Verlag. IAG Symposia, Vol. 116.
Tscherning, C. C., 1983. The role of high-degree spherical harmonic expansions in solving geodetic problems. In *Proceedings of the International Association of Geodesy Symposia, IUGG XVII General Assembly*, Department of Geodesy Science and Survey, Ohio State University, Columbus, Vol. 1, pp 431–441
Wang, Y. M., 2001. GSFC00 mean sea surface, gravity anomaly, and vertical gravity gradient from satellite altimeter data. *Journal of Geophysical Research*, **106**(C12), 31167–31174.
Wenzel, H -G., 1985. Hochauflösende Kugelfunktionsmodelle für das Gravitationspotential der Erde, *Wiss. Arb. 137*, Fachrichtung Vermess. der Univ. Hannover, Hannover, Federal Republic of Germany.
Wenzel, G., 1998. Ultra high degree geopotential models GPM98A, B and C to degree 1800, http://www.gik.uni-karlsruhe.de/∼wenzel/gpm98abc/gpm98abc.htm.
Wenzel, G., 1999. Schwerefeldmodellierung durch ultrahochauflösende Kugelfunktionsmodelle. *Zeitschrift für Vermessungswesen*, **124**(5), 144–154.
Wolff, M., 1969. Direct Measurements of the Earth's gravitational potential using a satellite pair. *Journal of Geophysical Research*, **74**(22), 5295–5300.
Yi, Y., 1995. *Determination of gridded mean sea surface from TOPEX, ERS-1 and GEOSAT altimeter data*. Rep. 434, Department of Geodetic Science and Survey, Ohio State University, Columbus.
Zhongolovich, I. D., 1952. The external gravitational field of the Earth and the fundamental constants related to it, Acad. Sci. Publ. Inst. Teor. Astron., Leningrad.

Cross-references

Geoid
Geoid Determination, Theory and Principles
Geoid Undulation, Interpretation
Geoid, Computational Method
Gravity Anomalies, Interpretation
Gravity Field of the Earth
Gravity Field, Temporal Variations from Space Techniques
Gravity Field, Time Variations from Surface Measurements
Gravity Measurements, Absolute
Gravity Method, Airborne
Gravity Method, Principles
Gravity Method, Satellite
Gravity Method, Surface
Gravity Modeling, Theory and Computation
Spherical Harmonic Analysis Applied to Potential Fields

GRAVITY, GRADIOMETRY

Christopher Jekeli
Division of Geodetic Science, School of Earth Sciences, Ohio State University, Columbus, OH, USA

Introduction

Gravity gradiometry is over 110 years old and it offers a rich perspective of the Earth's external gravitational field as generated by its internal mass density structure. Like the measurement of gravity, gradiometry is used to make both qualitative and quantitative interpretations of the Earth's near subsurface interior. The spatial gradients of the Earth's gravity acceleration vector form a tensor of nine elements, three components for each of the three vector components. They emphasize the fine structure and the short-wavelength characteristics of the field. They are largest over sharp subsurface geologic contrasts, such as faults, the edges of salt domes and igneous intrusions, and the flanks of synclines and anticlines. As the second derivatives of the gravity potential, the gravity gradients describe the curvature of the potential, and values of the differential curvature help to indicate directions of principal axes of source bodies. Gravity gradiometry also plays a significant role in understanding the fundamental nature of gravitation from a cosmological viewpoint.

While popular in the early part of the twentieth century to aid in the exploration for oil and gas, static gravity gradiometry has largely been replaced by scalar gravimetry (measurements of the magnitude of the gravity vector), both in the static mode and on airborne platforms, due to its greater efficiency and yet equal accuracy. Gradiometry has resurfaced recently because for moving-base systems (e.g., aircraft and satellites), it is arguably the preferred system of measurement since it does not require an independent system that determines kinematic accelerations of the vehicle with equal precision, as does the gravimeter due to the principle of equivalence between inertial and gravitational masses.

The static or moving-base measurement of gradients requires exquisitely sensitive instrumentation, which over the last 100 years has improved less in accuracy than in resolution (integration time) for the surface and airborne systems. Typical airborne systems today advertise the same precision that the static instruments had at the beginning of the twentieth century. However, they have a much higher temporal resolution (of the order of a few seconds) compared to the many hours required for the early static instruments. On the other hand, the recent launch of a satellite-borne gradiometer represents a major advance in global gradiometry with an anticipated accuracy improvement of several orders of magnitude, but at the expense of the relatively far removed (from the sources), but dynamically quiet environment of space. Because of

the relatively great expense of current gradiometer technology, to date only a few terrestrial (airborne and shipborne) systems are in operation.

The following sections cover the basic theory of gravitational gradients, the history of the sensor technology from the first torsion balance to modern-day gradiometers, the forward modeling of gravitational gradients, and applications in geodesy, geophysics, and physics.

Mathematical foundations

Gravitational gradients

Newtonian gravitational field theory adequately serves both to develop the necessary field equations, their solutions, and the derivation and use of gravitational gradients. No attempt is made or needed for present purposes to discuss the gradients in terms of Einstein's theory of general relativity. We start with Newton's law of gravitation that posits the force between two point masses as proportional to each of the masses, inversely proportional to the square of the distance between them, and directed along the connecting line. An equivalent concept is the Newtonian gravitational field potential that quantifies the work (energy) needed to bring a unit test mass from a position, $\boldsymbol{x}$, to infinity for a given attracting point mass, m, located at $\boldsymbol{x}'$:

$$V(\boldsymbol{x}) = \frac{Gm}{|\boldsymbol{x} - \boldsymbol{x}'|}, \tag{1}$$

where G is Newton's gravitational constant. We use the geophysical convention of positive potential. By the law of superposition, the potential may be generalized by summation to a discrete set of mass points; and, ultimately, using the concept of mass density, ρ, to continuous distributions of mass:

$$V(\boldsymbol{x}) = G \int_v \frac{\rho(\boldsymbol{x}')}{|\boldsymbol{x} - \boldsymbol{x}'|} \mathrm{d}v, \tag{2}$$

where v is the region in space occupied by the mass density. The potential is a continuous function for all points inside and outside v and on its boundary.

The spatial gradient of the gravitational potential is the gravitational acceleration (gravitation) experienced by the test mass at $\boldsymbol{x}$:

$$\mathbf{g}(\boldsymbol{x}) = \nabla V(\boldsymbol{x}). \tag{3}$$

The spatial gradient of the components of the gravitation vector constitutes the elements of the gravitational gradient tensor:

$$\Gamma(\boldsymbol{x}) = \nabla\nabla^{\mathrm{T}} V(\boldsymbol{x}) = \nabla \boldsymbol{g}^{\mathrm{T}}(\boldsymbol{x}), \tag{4}$$

where "T" denotes the transpose of a vector (or matrix). The gradient tensor is a continuous function where the density is continuous.

In Cartesian coordinates, $\boldsymbol{x} = (x_1, x_2, x_3)^{\mathrm{T}}$, the gravitational gradient tensor is given by

$$\Gamma = \begin{pmatrix} \frac{\partial^2 V}{\partial x_1^2} & \frac{\partial^2 V}{\partial x_1 \partial x_2} & \frac{\partial^2 V}{\partial x_1 \partial x_3} \\ \frac{\partial^2 V}{\partial x_2 \partial x_1} & \frac{\partial^2 V}{\partial x_2^2} & \frac{\partial^2 V}{\partial x_2 \partial x_3} \\ \frac{\partial^2 V}{\partial x_3 \partial x_1} & \frac{\partial^2 V}{\partial x_3 \partial x_2} & \frac{\partial^2 V}{\partial x_3^2} \end{pmatrix} = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_2}{\partial x_1} & \frac{\partial g_3}{\partial x_1} \\ \frac{\partial g_1}{\partial x_2} & \frac{\partial g_2}{\partial x_2} & \frac{\partial g_3}{\partial x_2} \\ \frac{\partial g_1}{\partial x_3} & \frac{\partial g_2}{\partial x_3} & \frac{\partial g_3}{\partial x_3} \end{pmatrix} = \begin{pmatrix} \Gamma_{11} & \Gamma_{12} & \Gamma_{13} \\ \Gamma_{21} & \Gamma_{22} & \Gamma_{23} \\ \Gamma_{31} & \Gamma_{23} & \Gamma_{33} \end{pmatrix}. \tag{5}$$

The tensor is symmetric by the continuity of the derivatives of the potential. Poisson's field equation characterizes a Newtonian gravitational field through its second-order gradients, or the trace of the gravitational gradient tensor:

$$\nabla \bullet \nabla V(\boldsymbol{x}) = \nabla^2 V(\boldsymbol{x}) = -4\pi G \rho(\boldsymbol{x}), \tag{6}$$

where the Laplacian operator in Cartesian coordinates is

$$\nabla^2 = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2}. \tag{7}$$

Therefore, the gravitational gradient tensor has only five independent elements for points where the mass density distribution is known. For later reference, the potential in free space satisfies Laplace's equation:

$$\nabla^2 V(\boldsymbol{x}) = 0. \tag{8}$$

If the local coordinate system has its third axis aligned with the vertical, positive downward, and the first axis points northward, then the horizontal rate of change of the vertical component of gravitation, g_3, is called the horizontal gradient, with magnitude and direction of greatest variation (from north) given by

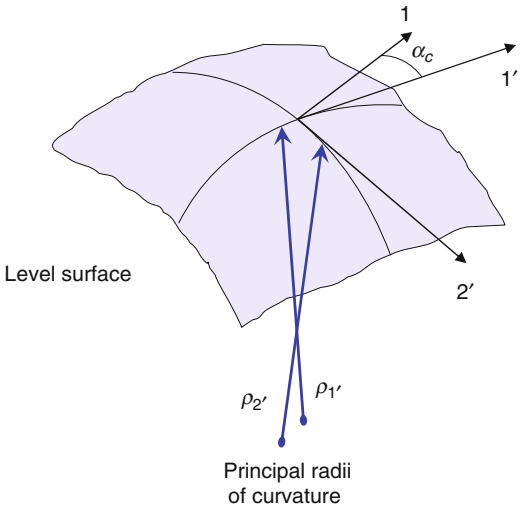
$$\Gamma_H = \sqrt{\Gamma_{13}^2 + \Gamma_{23}^2}, \quad \alpha_H = \tan^{-1} \frac{\Gamma_{23}}{\Gamma_{13}}. \tag{9}$$

Gravitational gradients also describe the curvature of a surface of constant gravitational potential (a level surface) (Figure 1). Let $\rho_{1'}$ and $\rho_{2'}$ be the two principal radii of curvature at a point on the surface, representing the minimum and maximum curvatures along corresponding (orthogonal) axes, $x_{1'}$ and $x_{2'}$. It can be shown (Heiland, 1940; Nettleton, 1976) that the differential curvature, $\Delta\kappa = g_3(1/\rho_{2'} - 1/\rho_{1'})$, is given by

$$\Delta\kappa = \sqrt{(2\Gamma_{12})^2 + (\Gamma_{22} - \Gamma_{11})^2}. \tag{10}$$

The direction of minimum curvature is

$$\alpha_c = \frac{1}{2} \tan^{-1} \frac{-2\Gamma_{12}}{\Gamma_{22} - \Gamma_{11}}. \tag{11}$$



Gravity, Gradiometry, Figure 1 Principal radii of curvature at a point on an equipotential surface.

If $a_c = 0$, then the coordinate axes, x_1 and x_2, coincide with the directions of minimum and maximum curvatures; and $\Gamma_{12} = 0$. In general, we also have

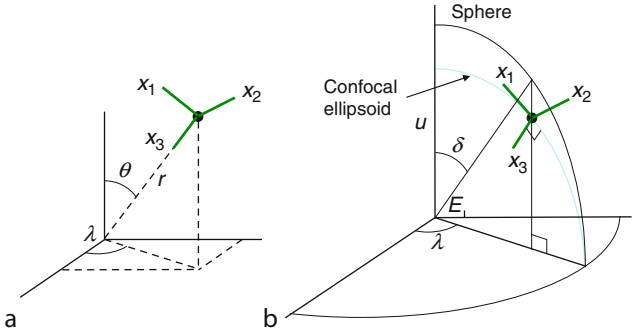
$$\Gamma_{22} - \Gamma_{11} = g_3 \left(\frac{1}{\rho_{2'}} - \frac{1}{\rho_{1'}} \right) \cos 2\alpha_c,$$
$$\Gamma_{12} = -\frac{1}{2} g_3 \left(\frac{1}{\rho_{2'}} - \frac{1}{\rho_{1'}} \right) \sin 2\alpha_c. \tag{12}$$

Other coordinate systems

For a local Cartesian coordinate system embedded in a spherical polar coordinate system as shown in Figure 2a, the first- and second-order derivatives are related to corresponding derivatives with respect to the spherical coordinates, as follows:

$$\frac{\partial}{\partial x_1} = -\frac{1}{r}\frac{\partial}{\partial \theta}, \quad \frac{\partial}{\partial x_2} = \frac{1}{r \sin\theta}\frac{\partial}{\partial \lambda}, \quad \frac{\partial}{\partial x_3} = -\frac{\partial}{\partial r}; \tag{13}$$

$$\begin{aligned}
\frac{\partial^2}{\partial x_1^2} &= \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2} \\
\frac{\partial^2}{\partial x_1 \partial x_2} &= \frac{\cot\theta}{r^2 \sin\theta}\frac{\partial}{\partial \lambda} - \frac{1}{r^2 \sin\theta}\frac{\partial^2}{\partial \theta \partial \lambda} \\
\frac{\partial^2}{\partial x_1 \partial x_3} &= -\frac{1}{r^2}\frac{\partial}{\partial \theta} + \frac{1}{r}\frac{\partial^2}{\partial \theta \partial r} \\
\frac{\partial^2}{\partial x_2^2} &= \frac{\cot\theta}{r^2}\frac{\partial}{\partial \theta} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2 \sin^2\theta}\frac{\partial^2}{\partial \lambda^2} \\
\frac{\partial^2}{\partial x_2 \partial x_3} &= \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \lambda} - \frac{1}{r \sin\theta}\frac{\partial^2}{\partial \lambda \partial r} \\
\frac{\partial^2}{\partial x_3^2} &= \frac{\partial^2}{\partial r^2}
\end{aligned} \tag{14}$$



Gravity, Gradiometry, Figure 2 Local Cartesian coordinates relative to (**a**) spherical polar coordinates and (**b**) ellipsoidal coordinates.

with corresponding Laplacian operator

$$\nabla^2 = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2} = \frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2} + \frac{\cot\theta}{r^2}\frac{\partial}{\partial \theta} + \frac{1}{r^2 \sin^2\theta}\frac{\partial^2}{\partial \lambda^2}. \tag{15}$$

Usually the Earth is approximated by a rotational ellipsoid, generated by rotating an ellipse with linear eccentricity, E, about its semi-minor axis, identified as the polar axis. Corresponding ellipsoidal (spheroidal) coordinates, (u, δ, λ), are shown in Figure 2b, where surfaces of constant u are confocal ellipsoids and surfaces of constant δ are hyperboloids of revolution of one sheet. As for spherical coordinates, surfaces of constant longitude, λ, are half-planes containing the polar axis. The local first- and second-order derivatives with respect to Cartesian coordinates (where the x_3-axis is orthogonal to the confocal ellipsoid that passes through the origin of the local system) are related to the corresponding derivatives with respect to the ellipsoidal coordinates, as follows:

$$\frac{\partial}{\partial x_1} = -\frac{1}{w}\frac{\partial}{\partial \delta}, \quad \frac{\partial}{\partial x_2} = \frac{1}{v \sin\delta}\frac{\partial}{\partial \lambda}, \quad \frac{\partial}{\partial x_3} = -\frac{w}{v}\frac{\partial}{\partial u}; \tag{16}$$

$$\begin{aligned}
\frac{\partial^2}{\partial x_1^2} &= \frac{1}{w^2}\left(\frac{uv^2}{w^2}\frac{\partial}{\partial u} + \frac{E^2 \sin 2\delta}{2w^2}\frac{\partial}{\partial \delta} + \frac{\partial^2}{\partial \delta^2} \right) \\
\frac{\partial^2}{\partial x_1 \partial x_2} &= \frac{1}{wv \sin\delta}\left(\cot\delta \frac{\partial}{\partial \lambda} - \frac{\partial^2}{\partial \delta \partial \lambda} \right) \\
\frac{\partial^2}{\partial x_1 \partial x_3} &= \frac{v}{w^2}\left(-\frac{u}{w^2}\frac{\partial}{\partial \delta} + \frac{E^2 \sin 2\delta}{2w^2}\frac{\partial}{\partial u} + \frac{\partial^2}{\partial \delta \partial r} \right) \\
\frac{\partial^2}{\partial x_2^2} &= \frac{\cot\delta}{w^2}\frac{\partial}{\partial \delta} + \frac{u}{w^2}\frac{\partial}{\partial u} + \frac{1}{v^2 \sin^2\delta}\frac{\partial^2}{\partial \lambda^2} \\
\frac{\partial^2}{\partial x_2 \partial x_3} &= \frac{1}{w \sin\delta}\left(\frac{u}{v^2}\frac{\partial}{\partial \lambda} - \frac{\partial^2}{\partial \lambda \partial u} \right) \\
\frac{\partial^2}{\partial x_3^2} &= \frac{1}{w^2}\left(-\frac{E^2 \sin 2\delta}{2w^2}\frac{\partial}{\partial \delta} - \frac{uE^2 \sin^2\delta}{w^2}\frac{\partial}{\partial u} + v^2 \frac{\partial^2}{\partial u^2} \right)
\end{aligned} \tag{17}$$

where

$$v^2 = u^2 + E^2, \quad w^2 = u^2 + E^2\cos^2\delta. \tag{18}$$

The Laplacian operator is

$$\nabla^2 = \frac{1}{w^2}\left(v^2\frac{\partial^2}{\partial u^2} + 2u\frac{\partial}{\partial u} + \frac{\partial^2}{\partial \delta^2} + \cot\delta\frac{\partial}{\partial\delta} + \frac{w^2}{v^2\sin^2\delta}\frac{\partial^2}{\partial\lambda^2}\right). \tag{19}$$

Invariants

Being a second-order tensor, the gradients change under a rotation of Cartesian coordinates from system s to system t according to $\Gamma^t = C_s^t\Gamma^s C_t^s$, where C_s^t is the matrix of appropriate direction cosines. For any arbitrary Cartesian system, s, there exists a rotation that diagonalizes the tensor, Γ^s:

$$\begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} = \Gamma^t = C_s^t\Gamma^s C_t^s. \tag{20}$$

The diagonal elements, λ_j, are the eigenvalues of the gradient tensor and are called invariants (with respect to coordinate rotations). They are real numbers because the tensor is real and symmetric. All combinations of eigenvalues are also invariants.

Given the tensor, Γ, in some coordinate system, the eigenvalues are roots to the following characteristic equation (det = determinant):

$$\begin{aligned} \det(\lambda I - \Gamma) &= (\lambda - \lambda_1)(\lambda - \lambda_2)(\lambda - \lambda_3) \\ &= \lambda^3 - I_0\lambda^2 + I_1\lambda - I_2 = 0, \end{aligned} \tag{21}$$

from which it is easy to show that, with the gradient elements as designated in Equation 5,

$$I_0 = \lambda_1 + \lambda_2 + \lambda_3 = \Gamma_{11} + \Gamma_{22} + \Gamma_{33},$$

$$\begin{aligned} I_1 =& \lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_3\lambda_1 = \Gamma_{11}\Gamma_{22} \\ &+ \Gamma_{22}\Gamma_{33} + \Gamma_{11}\Gamma_{33} - \Gamma_{12}^2 - \Gamma_{23}^2 - \Gamma_{31}^2, \end{aligned} \tag{22}$$

$$I_2 = \lambda_1\lambda_2\lambda_3 = \det\Gamma; \tag{23}$$

and, therefore, these quantities are invariants with respect to rotation. For example, Poisson's Equation 6 implies $I_0 = -4\pi G\rho$.

From Equations 22, we have $I_0^2 = \lambda_1^2 + \lambda_2^2 + \lambda_3^2 + 2I_1$. If $I_0 = 0$ (e.g., above the Earth's surface), then $I_1 < 0$. Pedersen and Rasmussen (1990) show that

$$0 \le I \equiv -\left(\frac{I_2}{2}\right)^2 \Big/ \left(\frac{I_1}{3}\right)^3 \le 1; \tag{24}$$

and call I an indicator of two dimensionality. $I = 1$ implies that the source of the gradients is like a monopole and $I = 0$ implies that the generated field is invariant in some direction, which means that the source is two dimensional and can be described by a cross section.

Finally, we note that both Γ_H (Equation 9) and $\Delta\kappa$ (Equation 10) are horizontal invariants, that is, independent of the azimuthal orientation of the local frame.

Models and modeling

As will be quantified later in this section, the gravitational gradients, being second-order derivatives of the potential, reflect primarily the short-wavelength, local features of the Earth's crust. For this reason, the forward modeling of gradients due to sources has largely concentrated on formulating the gradient signals of polyhedra of constant density, which by suitable assemblage can simulate the local crust, specifically the local topography. Given a regular distribution of elevation data for a particular region, the topography may be modeled as a collection of right-rectangular prisms whose vertical dimension is defined by an elevation data point and whose base has dimensions equal to the horizontal resolution of the data. Other options to model the topography include triangular prisms (using a triangulation of the elevation data) and any of several numerical integration methods applied to (from Equation 2)

$$\Gamma_{jk}(\boldsymbol{x}) = G\int_v \rho(\boldsymbol{x}')\frac{\partial^2}{\partial x_j\partial x_k}\frac{1}{|\boldsymbol{x} - \boldsymbol{x}'|}dv. \tag{25}$$

The purposes of modeling gravitational gradients range from validation of gradiometer accuracy to introducing constraints on the geophysical inverse problem. As new gradiometer systems are developed, particularly airborne systems, determining their accuracy is aided by flying over regions with well-known topographic elevations and density contrasts. Comparing the measurements to gradients forward modeled from the topography calibrates the noise level and enables tuning of appropriate data filters. Conversely, detecting subsurface geologic features, such as kimberlites (for diamond prospecting), from gradiometric surveys is facilitated by knowing what to look for in terms of a particular gradient signature. Thus, even in the inversion of gradiometric data to infer sources, forward modeling of structures plays an important role.

The gradients generated by constant-density, right-rectangular prisms follow from Equation 25. If the prism has dimensions bounded by Cartesian coordinates, $a_1 \le x_1' \le a_2, b_1 \le x_2' \le b_2, c_1 \le x_3' \le c_2$, the gradients for points outside the prism are given by

$$\Gamma_{jk}(\boldsymbol{x}) = -G\rho\ln(x_\ell - x_\ell' + r)\Big|_{x_1'=a_1}^{a_2}\Big|_{x_2'=b_1}^{b_2}\Big|_{x_3'=c_1}^{c_2}, \tag{26}$$

$$\Gamma_{jj}(\boldsymbol{x}) = G\rho\tan^{-1}\frac{(x_k - x_k')(x_\ell - x_\ell')}{(x_j - x_j')r}\Big|_{x_1'=a_1}^{a_2}\Big|_{x_2'=b_1}^{b_2}\Big|_{x_3'=c_1}^{c_2}, \tag{27}$$

where $r^2 = (x_1 - x_1')^2 + (x_2 - x_2')^2 + (x_3 - x_3')^2$ and where (j,k,ℓ) is a cyclic permutation of $(1,2,3)$. Using these formulas to model the gradients associated with the local topography is analogous to a numerical integration of Equation 25 using the rectangular rule.

To take advantage of various other numerical integration techniques based on a discretization of the topographic surface, $h(x_1', x_2')$, over an area, A, one first integrates Equation 25 with respect to the vertical, x_3':

$$\begin{aligned} \Gamma_{jk}(\boldsymbol{x}) &= G\rho \iint_A T_{jk}(\boldsymbol{x}; x_1', x_2')\, dA \\ &\approx G\rho \sum_{i_1} \sum_{i_2} T_{jk}\left(\boldsymbol{x}; (x_1)_{i_1}, (x_2)_{i_2}\right) \Delta A_{i_1,i_2}, \end{aligned} \tag{28}$$

where the functions,

$$T_{jk}(\boldsymbol{x}; x_1', x_2') = \int_0^{h(x_1', x_2')} \frac{\partial^2}{\partial x_j \partial x_k}\left(\frac{1}{r}\right) dx_3', \tag{29}$$

have well-defined analytic expressions (Jekeli and Zhu, 2006). The discretization indicated by Equation 28 then may take any number of forms.

For example, it may also be desirable to express T_{jk} as a Taylor series in h, in which case Equation 28 becomes a series of two-dimensional convolutions on the plane:

$$\Gamma_{jk}(\boldsymbol{x}) = G\rho \iint_A \left(\sum_n \frac{1}{n!} \frac{\partial^n}{\partial h^n} T_{jk}\big|_{h=0} h^n \right) \mathrm{d}x_1' \mathrm{d}x_2' \,, \tag{30}$$

where it is noted that T_{jk} is a function of $x_1 - x_1'$ and $x_2 - x_2'$ (besides x_3 and h). Discretized, such convolutions can be evaluated rapidly using fast algorithms for the discrete Fourier transform (Forsberg, 1985), provided the variable, x_3, remains constant (which is a limitation of the method). Similar Fourier transform methods have also been developed by (Parker, 1972), but found to be less suitable (Jekeli and Zhu, 2006).

Other fast methods transform the volume integral (Equation 25) to surface and line integrals along the theory of Gauss (divergence theorem, etc.); Petrovic (1996) and Tsoulis and Petrovic (2001) have developed corresponding formulas for arbitrary polyhedra. In all these cases, the underlying coordinate system is Cartesian. Modeling gravitational gradients in spherical coordinates that account for the convergence of coordinate lines of a data grid, as well as the convergence of radial (height) lines has not received much attention (see Smith et al., 2001) since the gradients are influenced primarily by very local topography.

Spectral models of the gravitational gradients can be derived directly from corresponding models of the gravitational potential, which exist in abundance for the global field. The Legendre spectrum, $\{(GM/R)C_{nm}\}$, of the potential, V, on a sphere of radius, R, represents its eigenvalues with respect to surface spherical harmonic functions, $\bar{Y}_{nm}(\theta,\lambda)$; and because the potential satisfies Laplace's Equation 8 in free space, the potential for $r \geq R$ is given by

$$V(r,\theta,\lambda) = \frac{GM}{R} \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \left(\frac{R}{r}\right)^{n+1} C_{nm} \bar{Y}_{nm}(\theta,\lambda), \tag{31}$$

assuming no masses exist outside the sphere of radius, R, and total mass, M. The spherical harmonic functions are defined by

$$\bar{Y}_{n,m}(\theta,\lambda) = \bar{P}_{n,|m|}(\cos\theta) \begin{cases} \cos m\lambda, & m \geq 0 \\ \sin|m|\lambda, & m < 0 \end{cases} \tag{32}$$

where the functions, $\bar{P}_{n,m}$, are associated Legendre functions of the first kind, fully normalized so that

$$\frac{1}{4\pi} \iint_\sigma \bar{Y}_{n',m'}(\theta,\lambda) \bar{Y}_{n,m}(\theta,\lambda) d\sigma = \begin{cases} 1, & n' = n \text{ and } m' = m \\ 0, & n' \neq n \text{ or } m' \neq m \end{cases} \tag{33}$$

Typically, a spherical harmonic model is given with a limited spectrum up to maximum degree, $n_{\max}$. The gradients, Γ_{jk}, corresponding to this model in a local coordinate system, as shown in Figure 2a, can then be determined using Equation 14. However, the spectrum with respect to the surface spherical harmonics does not exist for all the gradients. Specifically, any of the derivatives in latitude destroy the orthogonality of the basis (eigen-) functions on the sphere. It is easy to see that the radial derivatives possess Legendre spectra.

Similarly, for a given ellipsoidal harmonic model with coefficients, C_{nm}^e (Jekeli, 1988a),

$$V(u,\delta,\lambda) = \frac{GM}{R} \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \frac{\bar{Q}_{n|m|}(iu/E)}{\bar{Q}_{n|m|}(ib/E)} C_{nm}^e \bar{Y}_{nm}(\delta,\lambda), \tag{34}$$

where $\bar{Q}_{nm}$ is the fully normalized associated Legendre function of the second kind, the gradients in a local coordinate system are obtained using the derivatives in Equation 17. None of the derivatives of the potential have a Legendre spectrum with respect to the eigenfunctions, $\bar{Y}_{nm}(\delta,\lambda)$ because the geometry of the ellipsoid does not admit to a decomposition formula for the Legendre polynomials of $\cos\delta$.

Usually, a local Cartesian coordinate system suffices to model the spectrum of the gradients. In this case, the spectra of all gradients may be derived from the spectrum of the potential. Starting with the representation of the potential in terms of its Fourier spectrum, $\upsilon(f_1,f_2)$, on a plane $(x_3 = 0)$ and including the harmonic extension into free space $(x_3 > 0)$, we have

$$V(x_1,x_2,x_3) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \upsilon(f_1,f_2)e^{-2\pi f x_3}e^{i2\pi(f_1x_1+f_2x_2)}df_1df_2, \tag{35}$$

where $f = \sqrt{f_1^2+f_2^2}$ and f_1, f_2 are spatial, cyclical frequencies. The Fourier spectrum, $\mathcal{G}$, of each gravitational gradient on the x_3-plane is obtained by applying the appropriate derivative. We have

$$\begin{aligned}
\mathcal{G}_{11}(f_1,f_2;x_3) &= -4\pi^2 f_1^2 \upsilon(f_1,f_2)e^{-2\pi f x_3},\\
\mathcal{G}_{12}(f_1,f_2;x_3) &= -4\pi^2 f_1 f_2\, \upsilon(f_1,f_2)e^{-2\pi f x_3},\\
\mathcal{G}_{13}(f_1,f_2;x_3) &= -i4\pi^2 f_1 f\, \upsilon(f_1,f_2)e^{-2\pi f x_3},\\
\mathcal{G}_{22}(f_1,f_2;x_3) &= -4\pi^2 f_2^2\, \upsilon(f_1,f_2)e^{-2\pi f x_3},\\
\mathcal{G}_{23}(f_1,f_2;x_3) &= -i4\pi^2 f_2 f\, \upsilon(f_1,f_2)e^{-2\pi f x_3},\\
\mathcal{G}_{23}(f_1,f_2;x_3) &= 4\pi^2 f^2\, \upsilon(f_1,f_2)e^{-2\pi f x_3}.
\end{aligned} \tag{36}$$

These relationships can be used to determine any gradient from any other gradient or from any derivative of the potential. For example, the gradient, Γ_{13}, at altitude, x_3, is the inverse Fourier transform of $2\pi f\gamma(f_1,f_2)e^{-2\pi f x_3}$, where $\gamma(f_1,f_2) = -2\pi f\, \upsilon(f_1,f_2)$ is the Fourier transform of the vertical derivative of V (the gravitational acceleration) at ground level. In such transformations, due care must be exercised when calculating the higher frequencies of the gradients if the original data have noise and are limited in resolution due to finite data spacing.

Reference field and power spectral density

The first-order approximation of the Earth's gravity potential is the "normal potential," which includes the centrifugal component due to Earth's rotation and whose defining property is that it is constant on a co-rotating ellipsoidal model for the Earth. It may be expressed by the spherical harmonic expansion,

$$\begin{aligned}
U(r,\theta,\lambda) = &\frac{GM}{r}\left(1-\sum_{n=1}^{\infty}J_{2n}\left(\frac{a}{r}\right)^{2n}P_{2n}(\cos\theta)\right)\\
&+\frac{1}{2}\omega_e^2 r^2\sin^2\theta,
\end{aligned} \tag{37}$$

where a is the semi-major axis of the ellipsoid, ω_e is Earth's rotation rate, and the coefficients are given by

$$J_{2n} = (-1)^{n+1}\frac{3e^{2n}}{(2n+1)(2n+3)}\left(1-n+5n\frac{J_2}{e^2}\right), \tag{38}$$

where e is the first eccentricity of the ellipsoid, and J_2 is the dynamical form factor related to the difference in principal moments of inertia of the Earth, or also the polar flattening of the ellipsoid (Heiskanen and Moritz, 1967; Hofmann-Wellenhof and Moritz, 2005). The series (37) is usually truncated at $n = 5$ with entirely sufficient accuracy. For values of the defining parameters, GM, a, J_2, ω_e, of the current and past reference fields, see entry on *Gravity Field of the Earth*.

The normal gravity potential, U, does not depend on longitude due to the rotational symmetry of the ellipsoid; and in view of Equation 14, only the local gravity gradients, Γ_{11}, Γ_{13}, Γ_{22}, and Γ_{33} include a normal component. A standard approximation for the normal gradient, $\Gamma_{33}^{(U)}$ is $\Gamma_{33}^{(U)} \approx 3086\times10^{-9}\mathrm{s}^{-2}$, based on just the first term in Equation 37. The common unit of measure for gravity gradients is the Eötvös ($1\ \mathrm{E} = 10^{-9}\mathrm{s}^{-2}$), in honor of the Hungarian geophysicist, Loránd Eötvös, who invented the torsion balance (see *Gravity, Gradiometry*).

The residual potential, called the disturbing potential, then gives rise to disturbing gradients, which may in some instances attain an order of magnitude similar to the reference part, especially in very mountainous areas (e.g., Hein, 1977).

Globally, the power spectral density (psd) of the potential, defined on the sphere of radius, R, is given in terms of its spherical spectrum by (Jekeli, 2010)

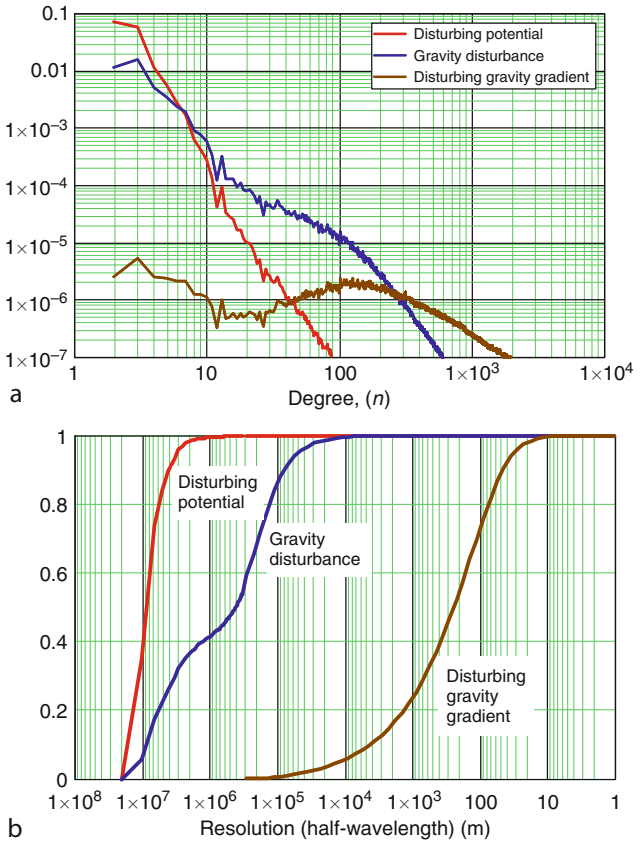
$$\Phi_n = \frac{1}{2n+1}\left(\frac{GM}{R}\right)^2\sum_{m=-n}^{n}C_{nm}^2. \tag{39}$$

Corresponding radial derivatives have psd's scaled by fixed powers of the degree, meaning that they emphasize the higher degrees or shorter wavelengths of the field. Figure 3a shows the psd's of the disturbing potential and its first two radial derivatives (relative to the normal field) according to the Earth Gravitational Model 2008 (EGM08, Pavlis et al., 2008).

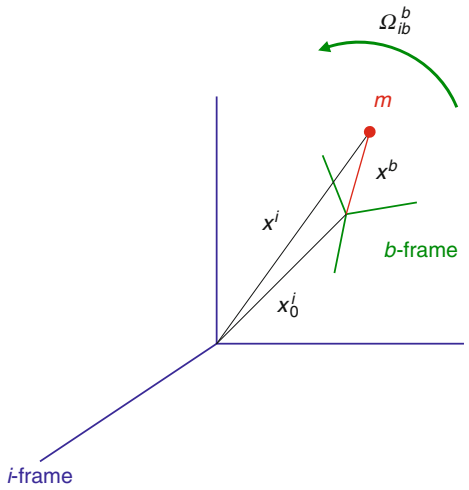
Also shown (Figure 3b) is the cumulative contribution of the disturbing field and its radial derivatives up to a given degree, n, or resolution, where spatial resolution (half-wavelength) in kilometers is given by $\Delta s = 6,371\pi/n$, 6,371 km being the mean Earth radius. While about 99.3% of the disturbing potential is defined by the field with wavelengths longer than 2,000 km, the gravity disturbance (first radial derivative) comprises only about 41.4% up to these wavelengths, and the disturbing gradient barely has any (relative) power at wavelengths longer than 200 km. Here, it is noted that the cumulative contribution of the disturbing gradient is highly dependent on the local field and is only an illustrative example. As another example, Jordan (1978) provided a psd model wherein 99.995% of the power of the disturbing gradient is due to the density variations in the upper 150 m of the crust, while this regime contributes 89% to the gravity disturbance, and only 7% to the disturbing potential.

Measurement error analysis

All practical gravity gradiometers measure "gradients" by differencing in one form or another the inertial accelerations sensed by linear or angular accelerometers. In the



Gravity, Gradiometry, Figure 3 (**a**) Power spectral densities of radial derivatives of the Earth's disturbing potential normalized by the respective total variance. (**b**) Cumulative relative power of each radial derivative as function of spatial resolution of the field.



Gravity, Gradiometry, Figure 4 The geometry for the lever-arm effect on the acceleration of m in the b-frame rotating with respect to the i-frame.

limit of infinitesimal separation, the model with linear accelerometers can be used to derive the theoretical equation for gravity gradiometers in this case, which also offers a general interpretation of measured gradients on a moving platform. We start with the lever-arm equation (Jekeli, 2000) for the output of an accelerometer (Figure 4)

$$\boldsymbol{a}^b = \boldsymbol{a}^b_0 + \left(\boldsymbol{g}^b_0 - \boldsymbol{g}^b\right) + \dot{\Omega}^b_{ib}\boldsymbol{x}^b + \Omega^b_{ib}\Omega^b_{ib}\boldsymbol{x}^b, \tag{40}$$

where $\mathbf{a^b}$ is the inertial acceleration of a test mass, m, in a frame (the b-frame) that rotates with respect to the inertial frame (i-frame) with angular rate given by the vector, $\boldsymbol{\omega}^b_{ib} = (\omega_1\ \omega_2\ \omega_3)^{\mathrm{T}}$, expressed in terms of a cross product as

$$\left[\boldsymbol{\omega}^b_{ib}\times\right] = \Omega^b_{ib} = \begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix}. \tag{41}$$

The acceleration of the b-frame in the i-frame is $\boldsymbol{a}^i_0$ and the location of the test mass (the lever-arm) in the b-frame is $\boldsymbol{x}^b$. The gravitational acceleration of the b-frame (in b-frame coordinates) is $\boldsymbol{g}^b_0$ and that of the test mass is $\boldsymbol{g}^b$. The last two terms of Equation 40 represent Coriolis and centrifugal accelerations due to the frame rotation.

Taking the spatial derivative of Equation 40, analogous to differencing the inertial accelerations of two test masses in the b-frame, we obtain

$$\frac{\partial \boldsymbol{a}^b}{\partial \boldsymbol{x}^b} = -\frac{\partial \boldsymbol{g}^b}{\partial \boldsymbol{x}^b} + \dot{\Omega}^b_{ib} + \Omega^b_{ib}\Omega^b_{ib}. \tag{42}$$

The measured acceleration gradient is the combination of the gravitational gradient and angular accelerations. Thus, a gravity gradiometer must either be stabilized against rotations with respect to the inertial frame or some other realizable frame, or these rotations must be measured independently with gyroscopes. An analysis of the required accuracy of these measurements (or, equivalently, of the stabilization) was conducted by Jekeli (2006) and is summarized in Table 1. Each row of this table corresponds approximately to commensurate sensor errors (gradiometer and gyroscope). For typical airborne gradiometer systems with noise levels in the range $1-10\,\mathrm{E}/\sqrt{\mathrm{Hz}}$, conventional high-accuracy gyroscopes are sufficient to separate the gravitational gradients from the non-gravitational components.

Due to the symmetry of the gravitational tensor, $\Gamma^b = \left(\partial \boldsymbol{g}^b/\partial \boldsymbol{x}^b\right)$, and the anti-symmetry of Ω^b_{ib}, the measurement of the full matrix, $\left(\partial \boldsymbol{a}^b_m/\partial \boldsymbol{x}^b\right)$, permits some simplifications, such as

$$\Gamma^b = -\frac{1}{2}\left(\frac{\partial \boldsymbol{a}^b}{\partial \boldsymbol{x}^b} + \left(\frac{\partial \boldsymbol{a}^b}{\partial \boldsymbol{x}^b}\right)^{\mathrm{T}}\right) + \Omega^b_{ib}\Omega^b_{ib}. \tag{43}$$

The next section offers further details on the measurement of gravity gradients.

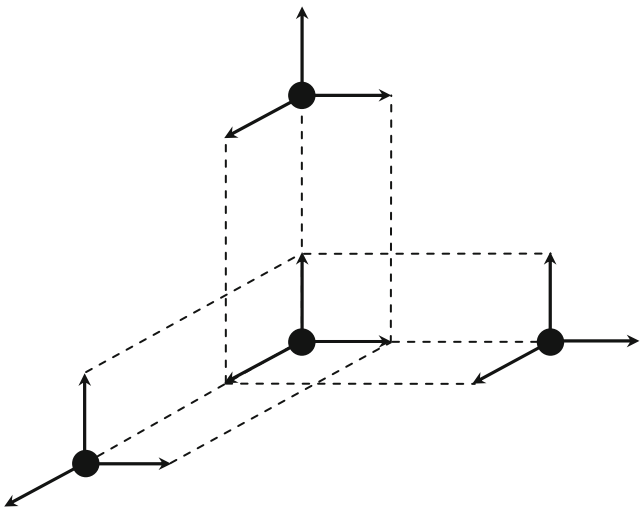
Gradiometry

Whereas the gravity meter, or any other accelerometer, senses the inertial reaction of a single test mass to an applied force, the (gravitational) gradiometer senses the difference of reactions from two test masses. In general terms, one may think of a gradiometer as the juxtaposition of two accelerometers and the sensed gradient as the ratio of the acceleration difference to the baseline length between them. As such, the gradiometer measures only an approximation of the gradient at a point, which is adequate for sufficiently small baselines. Measuring the entire tensor of gradients (a "full-tensor gradiometer") requires a minimum of 12 accelerometers as shown in Figure 5, although because some gradients share an accelerometer in this configuration, the errors between them may be correlated. Completely independent measurements of all nine tensor elements, therefore, require 18 accelerometers. On the other hand, taking advantage of the tensor symmetry and Laplace's field Equation 8 brings the minimum to five accelerometer pairs.

Most gradiometers actually are designed to sense components of the differential curvature of the field (Equation 12), and thus, they rely on Laplace's equation to determine the individual diagonal components. Specifically, imagine that a gradiometer senses $\Gamma_{22} - \Gamma_{11}, \Gamma_{12}$ in the instrument frame. Combining three such devices with mutually orthogonal orientation within the instrument system (Figure 6) thus yields in addition to $\Gamma_{12}, \Gamma_{23}, \Gamma_{31}$, the components $\Gamma_{22} - \Gamma_{11}$, $\Gamma_{33} - \Gamma_{22}$, and $\Gamma_{11} - \Gamma_{33}$. The latter always sum to zero (except for measurement errors) and cannot by themselves separate the diagonal elements of the tensor. Nevertheless, it is common to refer to this as a full-tensor gradiometer (FTG), even though effectively only five of the nine tensor elements are sensed independently.

Gravity, Gradiometry, Table 1 Roughly commensurate (per row) sensor errors in mobile gravity gradiometry. Emphasized entries represent typical calibrated error levels for high-accuracy components

Gradiometer white noise	Orientation error	Gyro bias	Gyro white noise
$30\,\mathrm{E}/\sqrt{\mathrm{Hz}}$	20°	0.5°/h	$3°/\mathrm{h}/\sqrt{\mathrm{Hz}}$
$\mathbf{10\,E}/\sqrt{\mathbf{Hz}}$	6°	0.15°/h	$1°/\mathrm{h}/\sqrt{\mathrm{Hz}}$
$1\,\mathrm{E}/\sqrt{\mathrm{Hz}}$	0.6°	**0.015°/h**	$\mathbf{0.1°/h/\sqrt{Hz}}$
$0.1\,\mathrm{E}/\sqrt{\mathrm{Hz}}$	**0.06°**	**0.0015°/h**	$\mathbf{0.01°/h/\sqrt{Hz}}$
$0.01\,\mathrm{E}/\sqrt{\mathrm{Hz}}$	**0.006°**	0.00015°/h	$0.001°/\mathrm{h}/\sqrt{\mathrm{Hz}}$



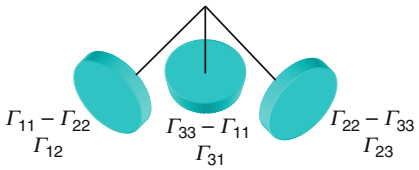
Gravity, Gradiometry, Figure 5 Conceptual minimal configuration for a full-tensor gradiometer.

The first instrument specifically designed to measure the gravity gradient was the torsion balance, invented by the Hungarian geophysicist, Lorand von Eötvös (1848–1919), at the end of the nineteenth century. Prolific in many endeavors of exploration geophysics, one of his passionate goals was to design an instrument that could sense all gravity gradients (he left the vertical-vertical gradient to others, since it would not be possible for his apparatus) (Szabó, 1998). (We make the distinction in geodesy between "gravitation" and "gravity," where the latter refers to measurements on the Earth's surface and is the vector sum of acceleration due to mass attraction (gravitation) and the centrifugal acceleration associated with Earth's rotation.)

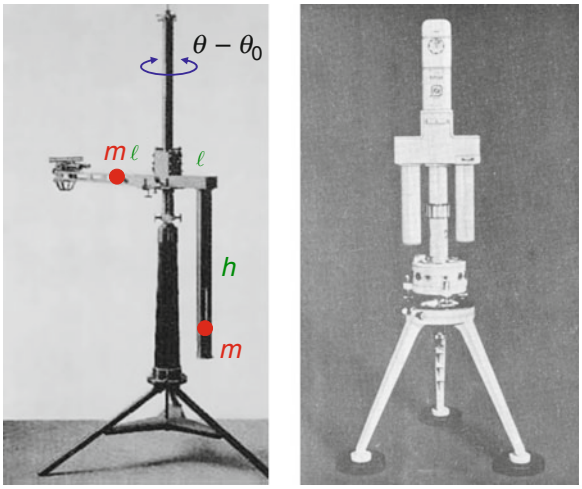
Eötvös adapted the concept of Coulomb's torsion balance (designed to sense electrostatic forces) to measure gravitational gradients. After several initial prototype instruments, including the curvature variometer, which was sensitive to the differential curvature components (Equation 12), and the horizontal variometer, which measured the horizontal gradients (Γ_{13}, Γ_{23}), the final product combined these into one instrument, shown in Figure 7. Two equal masses, m, are attached to the ends of a beam of length, 2ℓ, which is suspended by a platinum-iridium (or, tungsten) wire that may twist about the vertical axis in response to the curvature of the gravitational field. If both masses were at the same height ($h = 0$, in Figure 7), the instrument would be the curvature variometer. With the vertical offset of one mass, the instrument is also sensitive to the horizontal gradients.

Assuming the beam has negligible mass, the torque on the beam is

$$\boldsymbol{L} = \boldsymbol{r}_A \times m\boldsymbol{g}_A + \boldsymbol{r}_B \times m\boldsymbol{g}_B, \tag{44}$$



Gravity, Gradiometry, Figure 6 Mutually orthogonal gradiometer sensors in the "umbrella configuration."



Gravity, Gradiometry, Figure 7 Eötvös's original torsion balance (*Left*, from Szabó, 1998, reprinted with permission, Eötvös Lorand Geophysical Institute of Hungary.) and the more modern version manufactured by the Eötvös Loránd Geophysical Institute of Hungary, Budapest in the 1950s. (*Right*, from Badekas, 1967.)

where $\boldsymbol{r}_A = (-x, -y, 0)$ and $\boldsymbol{r}_B = (x, y, h)$ are the position vectors of the masses in a north-east-down coordinate system with origin at the instrument center, and $\mathbf{g}_A$, $\mathbf{g}_B$ are the gravity vectors at these points. The torque component about the vertical is

$$L_3 = mx(g_2(x,y,h) - g_2(-x,-y,0)) + my(g_1(-x,-y,0) - g_1(x,y,h)). \tag{45}$$

To first-order approximation, the gravity components are

$$g_1 = \Gamma_{11}x + \Gamma_{12}y + \Gamma_{13}z, \tag{46}$$

$$g_2 = \Gamma_{21}x + \Gamma_{22}y + \Gamma_{23}z, \tag{47}$$

which yields

$$L_3 = 2m(x^2 - y^2)\Gamma_{12} + 2mxy(\Gamma_{22} - \Gamma_{11}) + mh(x\Gamma_{23} - y\Gamma_{13}). \tag{48}$$

Here, the Γ_{jk} are gravity gradients as opposed to the gravitational gradients of the previous sections, as noted above.

We may introduce the azimuth, α, of the instrument with respect to north, such that $x = \ell\cos\alpha$ and $y = \ell\sin\alpha$. Then using the angular form of Hooke's law, $L_3 = -\tau(\theta_\alpha - \theta_0)$, Equation 48 becomes

$$-\tau(\theta_\alpha - \theta_0) = m\ell^2((\Gamma_{22} - \Gamma_{11})\sin 2\alpha + 2\Gamma_{12}\cos 2\alpha) + m\ell h(\Gamma_{23}\cos\alpha - \Gamma_{13}\sin\alpha), \tag{49}$$

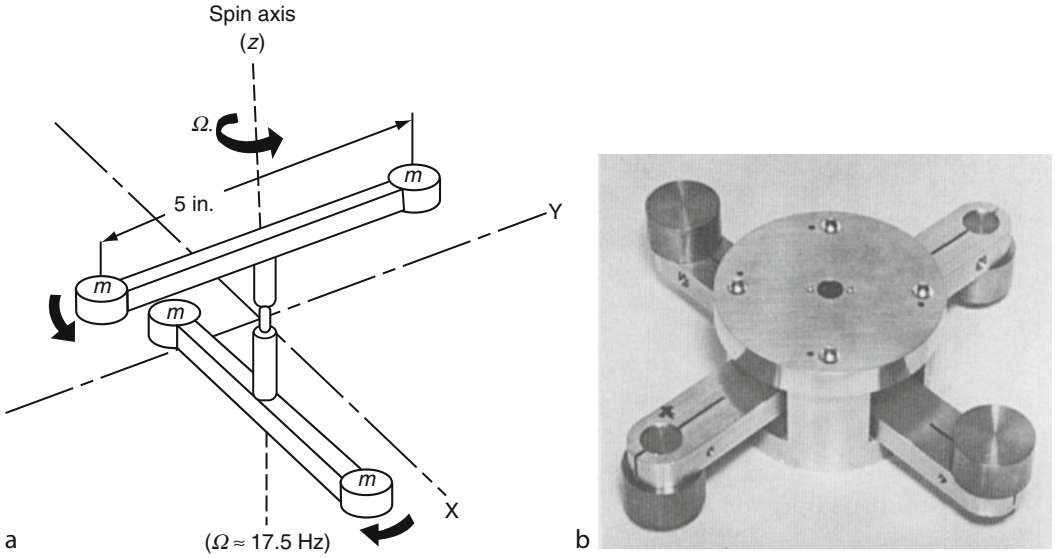
where τ is the torsion coefficient of the wire; θ_0 is the torsion-free, zero-direction of the beam; and θ_α is the angle of the beam relative to the zero-direction when the instrument (casing) is placed at azimuth, α. Measurements at several azimuths are required to determine all unknown quantities: θ_0, $\Gamma_{22} - \Gamma_{11}$, Γ_{12}, Γ_{13}, Γ_{23}.

Eötvös designed several other versions, such as the double torsion balance (two opposing beams) that he used to study the equivalence of inertial and gravitational mass (which Einstein postulated as intrinsically true for his general theory of relativity; see Applications). Though quite laborious and extremely sensitive to nearby masses in the field, the torsion balance (several versions built by various manufacturers) was widely used for geophysical prospecting of oil and gas in the early twentieth century both in Europe and in the USA. Ultimately, it gave way to the more efficient and equally accurate land gravity meter (Nettleton, 1976; Chapin, 1998).

Gravity gradiometry did not regain significant attention until the dawn of the space age since only gradiometers on planetary probes can make in-situ measurements of the gravitational field. Early developments of moving-base gradiometers mimicked the torsional sensing apparatus, similar, in principle, to Eötvös's horizontal variometer, but in a smaller package (thus sensing only $\Gamma_{22} - \Gamma_{11}$ and Γ_{12} in a local frame; Equation 49 with $h = 0$). In addition, by rotating the device about the axis perpendicular to the mass baseline, the sensed gradients are modulated by twice the rotation rate due to the symmetry of the mass-baseline configuration. In this way, some of the errors (such as accelerometer scale mismatches and misalignments) that modulate at once the rotation rate, Ω, can be separated in the frequency domain from the gradient signal that is modulated at twice this rate. That is, demodulation of the output signal at frequency, 2Ω, effectively filters these errors.

One of the earliest modern gradiometers was developed by Hughes Aircraft Research Laboratories for a lunar orbiter mission, the Hughes Rotating Gravity Gradiometer (RGG) (Forward, 1981). Figure 8 shows a schematic of the device that senses the relative torques on the two arms that are coupled to the base by torsion pivots. Two dumbbells are required to distinguish the gravitational gradients from torques on the case in a dynamic environment. Furthermore, the rotation rate of the device is designed to be half its mechanical resonant frequency so as to amplify sensed gradients. A demonstration model of the RGG (which was never put into operation) had a diameter of 14 cm and an accuracy of 1 E for a 10 s integration time (Forward, 1981).

The Charles Stark Draper Laboratory of Cambridge, Massachusetts, using its leading expertise in gyroscope technology, also devised several gradiometer concepts for moving-base applications. The "floated gradiometer," pictured in Figure 9a, was derived from the floated gyroscope whose buoyant/viscous/magnetic support system is virtually insensitive to high levels of acceleration and vibration. In this case, the device, like Eötvös's torsion balance, had two anomalous masses at different heights (skewed mass configuration). If, for example, $\alpha = -90^\circ$



Gravity, Gradiometry, Figure 8 Concept (From Heller, 1977) and breadboard model (From Forward, 1981) of the Hughes Aircraft Rotating Gravity Gradiometer.

is substituted in Equation 49 (maintaining the north-east-down coordinate system and indicating that the first axis of the device points west, toward $-x_2$), then the measurement is the applied restoring moment, M, in response to a deflection due to the gradients:

$$M = \tau(\theta_\alpha - \theta_0) = 2m\ell^2\Gamma_{12} - m\ell h\Gamma_{13}, \tag{50}$$

where ℓ is the radius of the device.

On a stabilized platform with axes constrained to the local north-east-down directions, a cluster of three such devices (Figure 9b) yields measured moments:

$$\begin{aligned} M_1 &= 2m\ell^2\Gamma_{12} - m\ell h\Gamma_{13} \\ M_2 &= 2m\ell^2\Gamma_{12} + m\ell h\Gamma_{13} \\ M_3 &= -2m\ell^2\Gamma_{12} + m\ell h\Gamma_{23} \end{aligned} \tag{51}$$

which can be combined to yield the horizontal gradients, Γ_{13} and Γ_{23}. Derived from the differences of corresponding torques (as shown), these measured gradients are insensitive to angular accelerations about the vertical axis, the direction about which a north stabilized platform is least accurate. Later designs used electrostatically suspended spheres, with anomalous masses, that in an appropriate triad configuration would constitute a full-tensor gradiometer (with Laplace's constraint) and be less susceptible to platform jitter effects (Wells, 1984). Like the Hughes Aircraft instruments, the Draper Lab gradiometers did not reach the production stage, but did demonstrate very good performance in the laboratory.

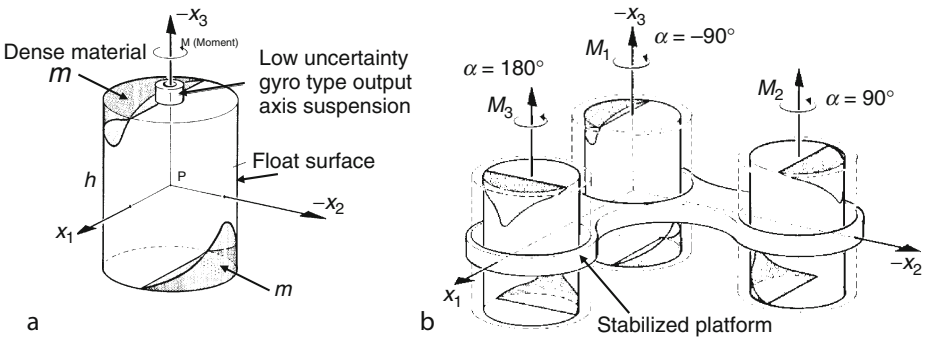
Another concept that utilized device rotation to separate the gradient signal from noise sources in the frequency domain was the gravity gradiometer instrument (GGI) of Bell Aerospace. The instrument consisted of pairs of accelerometers mounted on a rotating disk as shown in Figure 10, left. The input axes are perpendicular to the radius of the disk and thus, in analogy to the Hughes Aircraft RGG, their measurements may be thought of as proportional to torques about the spin axis. Substituting $\alpha = \Omega t$ and $\alpha = \Omega t + \pi/2$ for the respective pairs of accelerometers and $M = mr(a_1 + a_2)$ and $M = mr(a_3 + a_4)$ for the respective torques into Equation 49 (with $h = 0$), the measurement equation, therefore, is given by

$$\begin{aligned} &a_1(t) + a_2(t) - (a_3(t) + a_4(t)) \\ &\quad = 2r(\Gamma_{22} - \Gamma_{11})\sin 2\Omega t + 4r\Gamma_{12}\cos 2\Omega t, \end{aligned} \tag{52}$$

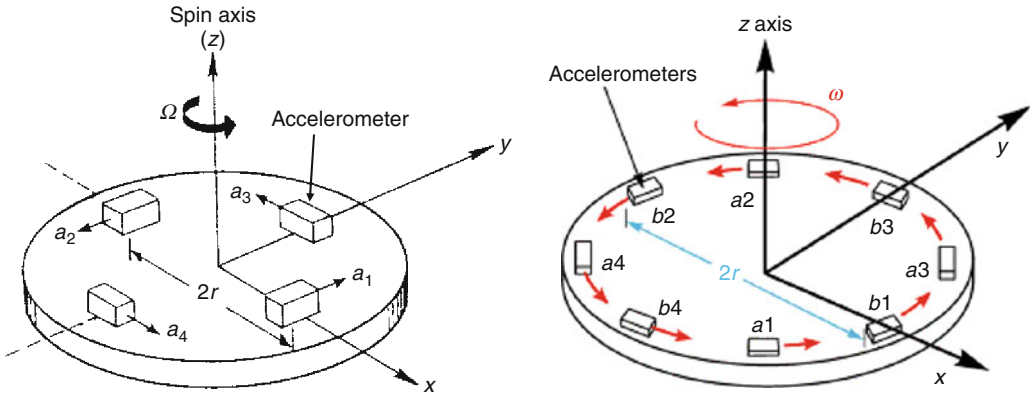
where Ω is the spin rate of the disk and t is time. The factor of 2 on the right side comes from the dual pairs of accelerometers. This instrument has been used for military and commercial applications and today's version (acquired by Lockheed Martin, Figure 10, right) has four pairs of accelerometers per disk, thus, in essence, further doubling the signal amplitude in the output (Schweitzer et al., 2000). However, three GGIs are still needed (e.g., in the configuration of Figure 6) to obtain all elements of the gradient tensor.

The Bell Aerospace (Lockheed Martin) gradiometer is currently operated for various geophysical and geodetic applications by Fugro (the erstwhile BHP Falcon single-disk system, Figure 11) and by Bell Geospace (full-tensor gradiometer, with the original four-accelerometer disks). The Falcon system is used primarily for airborne surveys, while the Bell Geospace system is deployed on aircraft and ocean vessels.

Satellite in-situ gravitational field measurements are limited fundamentally to gradiometry. Aside from small atmospheric drag forces and solar radiation (and Earth



Gravity, Gradiometry, Figure 9 Floated gradiometer geometry of the Charles Stark Draper Laboratory; (**a**) single unit, (**b**) cluster of three units on a stabilized platform (Both diagrams after Trageser, 1970, reprinted with permission, Draper Laboratory.)



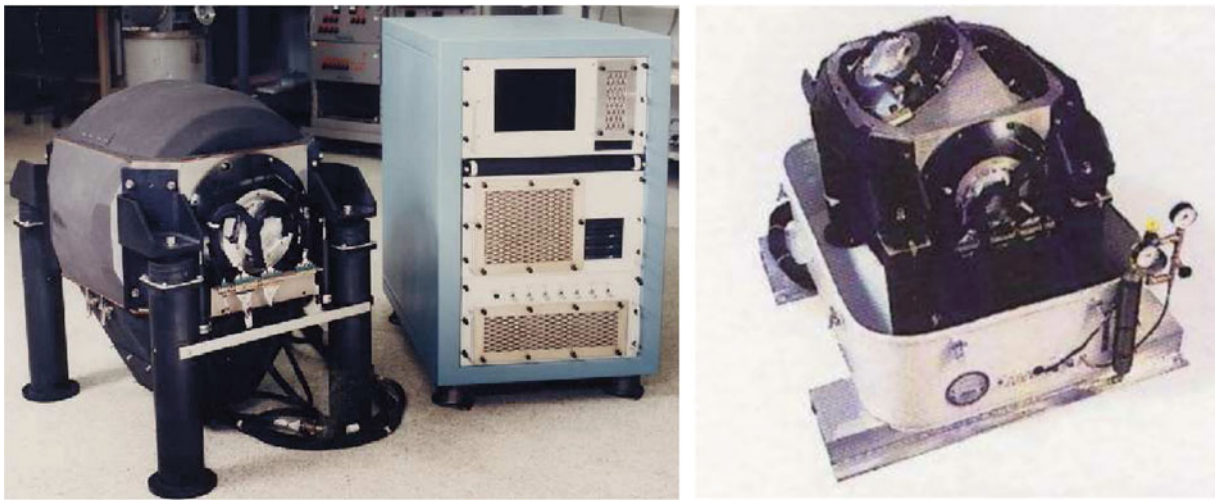
Gravity, Gradiometry, Figure 10 *Left*, Gravity Gradiometer Instrument (GGI), Bell Aerospace; *right*, current version, Lockheed Martin.

albedo) pressure, an Earth-orbiting satellite is in free fall. Accelerometers on a satellite sense only the non-gravitational forces, but a gradiometer, as on any moving platform, senses the gravitational gradient (Equation 42). Satellite gradiometry has a long, interesting, and ultimately successful history. Besides the Hughes RGG (which did not fly on a satellite), a major effort was undertaken by NASA to put a superconducting gravity gradiometer (SGG) into low Earth orbit in the 1980s. The SGG was developed by the University of Maryland (Moody et al., 1986; Paik et al., 1988) as an outgrowth of the technology of highly sensitive gravity wave detectors. Based on accelerometers that use the magnetic flux exclusion principle of superconducting bodies, the SGG was designed primarily to sense the in-line gradients. Referring to Figure 12, left, as the gap between the superconducting proof mass and the sensing coil changes due to acceleration, the magnetic flux is forced out of the sensing coil and into the input coil of a superconducting quantum interference device (SQUID), which is able to sense extremely small changes in the flux. The system was not flown on a satellite but did contribute to physics experiments involving the search for non-Newtonian gravitational effects (see Applications).

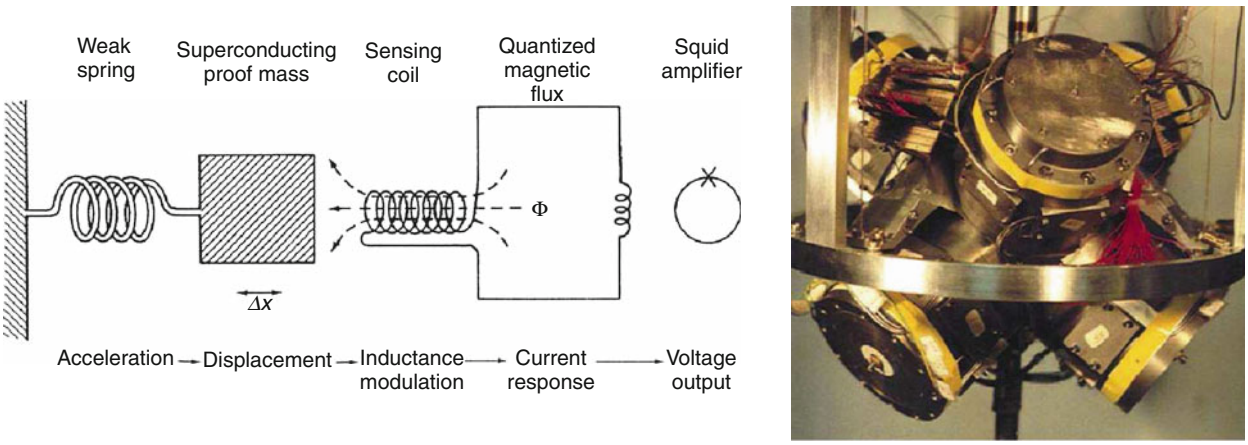
In March 2009, the first satellite-borne gravity gradiometer was launched by the European Space Agency (ESA) into low Earth orbit (altitude = 255 km) for a 20-month mission to map the Earth's gravitational field globally to a resolution of 100 km, or better (Drinkwater et al., 2003). The *Gravity field and steady-state Ocean Circulation Explorer* (GOCE) gradiometer is a true full-tensor device (Figure 13), based on three pairs of three-axis accelerometers, separated pair-wise by 0.5 m, thus in principle measuring all nine elements of the tensor, $\partial \boldsymbol{a}^b / \partial \boldsymbol{x}^b$ (Equation 42).

However, each accelerometer has one weakly calibrated axis, due to the earthly 1-g laboratory environment. As a consequence, the in-line components, $\partial a_1/\partial x_1$, $\partial a_2/\partial x_2$, $\partial a_3/\partial x_3$, and the cross components, $\partial a_1/\partial x_3$, $\partial a_3/\partial x_1$ are most accurate. The latter allow an accurate determination of the angular velocity, ω_2, about the normal to the orbital plane, which is greatest in magnitude (Earth-pointing stabilization), since from Equation 42 and the symmetry of the gravity gradient tensor, we have

$$\frac{\partial a_1}{\partial x_3} - \frac{\partial a_3}{\partial x_1} = 2\dot{\omega}_2. \tag{53}$$



Gravity, Gradiometry, Figure 11 Bell Aerospace (Lockheed Martin) GGI used on current terrestrial gradiometer systems. *Left*, the Fugro Falcon single-axis system (Previously BHP Billiton, from van Leeuwen, 2000, reprinted with permission, Society of Exploration Geophysics.); *right*, the full-tensor 3 GGI system of Bell Geospace (Murphy, 2004; reproduced with permission, Bell Geospace). Both systems are shown on inertially stabilized platforms.

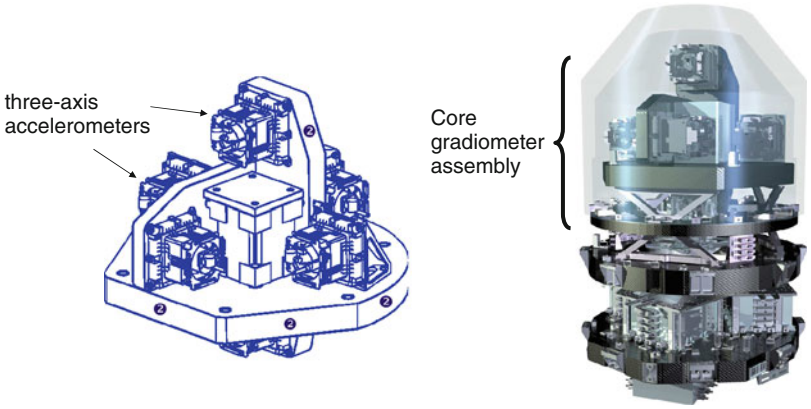


Gravity, Gradiometry, Figure 12 *Left*, the accelerometer concept of the superconducting gravity gradiometer of the University of Maryland (Reprinted with permission from Moody et al., 1986). Copyright (1986), American Institute of Physics). *Right*, a three-axis system that senses the three in-line gradients. (Paik, 2004; reprinted with permission of author.)

As a consequence and since ω_1 and ω_3 are comparatively small, the most accurate gravity gradients obtained by the GOCE gradiometer are the in-line gradients and the cross gradient, Γ_{13}.

Other types of gradiometers may be found discussed in the literature and are at various stages in their development. For example, a cryogenic gradiometer based on the same principle as the Hughes Aircraft RGG was constructed and tested by the Canadian company, Gedex (Tryggvason et al., 2004). Stanford University has investigated a gradiometer using dual cold-atom interferometer accelerometers (McGuirk et al., 2002). And, MEMS (micro-electro-mechanical systems) technology is being applied to the manufacture of miniature gradiometers (Flokstra et al., 2009). The reader is referred also to the review by Difrancesco (2007).

Designing and using a gravity gradiometer system requires consideration of several significant sources of error. Gradiometers on a moving platform must detect a very small differential gravitational signal buried in a large-amplitude acceleration environment. An accuracy of 1 E over an accelerometer baseline of 0.1 m implies better than 7×10^{-11} m/s^2 accuracy in the sensed accelerations, a resolving power that is challenging for most accelerometers. Rotation of the gradient sensor has been a key development to separate signal from noise in the



Gravity, Gradiometry, Figure 13 GOCE core gradiometer assembly (*left*) and total instrument (*right*).

frequency domain. Cryogenics is another solution, as is the dynamically quiet environment of a space vehicle.

Essential error sources for accelerometer-based gradiometers include the scale factor and bias stability of the accelerometers, as well as the degree to which their scale factors are matched and their sensitive axes are aligned. These all determine how well the common linear acceleration is rejected in the difference that ideally leaves only the gravitational gradient (with proper stabilization). Often, specialized electronic or mechanical devices or procedures (rotation and induced shaking) are used to eliminate or calibrate these erroneous effects. A particularly important calibration is the self-gradient of the vehicle that carries the gradiometer. Since the platform is stabilized (e.g., in the local level frame), the rotation of the vehicle about the platform creates a varying gradient field due to the changing attitude of the vehicle masses. This field must be calibrated, often using elaborate procedures that place the vehicle in varying attitudes while the gravitational field is constant (i.e., the vehicle is stationary) – see (Jekeli, 1988b).

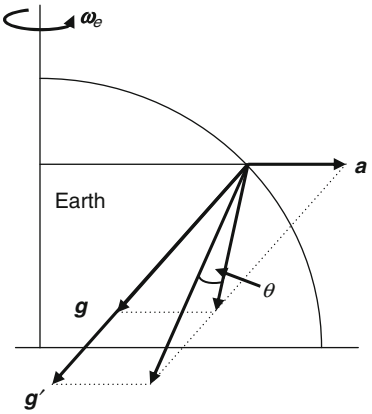
In summary, the development of instruments that measure gravitational gradients spans over 100 years; and, despite a variety of system advancements, mostly to make gradiometry mobile, many modern sensor concepts have strong analogies to the original Eötvös curvature variometer – they measure differential curvature parameters. In-line differential accelerometers are only just being deployed on a satellite system (GOCE) or are in development for future systems. The 1 E accuracy of the Eötvös torsion balance rivals or still bests the terrestrial mobile systems of today, but the latter have improved the integration time vastly from many hours to tens of seconds, or better. In principle, gradiometry relies primarily on the technology of inertial sensors, and the theoretical limit in measurement precision in many cases has not yet been reached at the operational level. It is clear that gravity gradiometry is the holy grail of precision gravimetry on moving platforms – the future will see many more ideas and prototype sensors added to the existing repertoire of instruments.

Applications

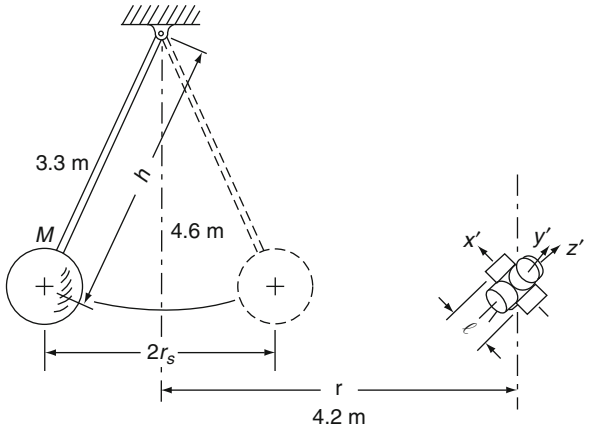
One of the more interesting applications of the torsion balance was Eötvös's experimental validation of the equivalence of inertial and gravitational mass, assumed by Einstein in his theory of general relativity. Inertial mass is the proportionality factor between velocity and linear momentum (or, between acceleration and *applied* force); while, gravitational mass is the proportionality factor between gravitational acceleration and gravitational force. Conceptually, the acceleration caused by the gravitational attraction of a mass body need not be indistinguishable from the acceleration caused by a force, such as propulsion, friction, or similar physical action. Yet the assumption of their equivalence led Einstein to his very successful theory.

Eötvös experimented already in the 1890s with the question of dependence of gravitational attraction on the composition of mass (continuing with Galileo's experiments that showed no such dependence). The idea of his experiment is illustrated in Figure 14 showing that different types of test masses on the beam of the torsion balance might yield different gravitational accelerations that in combination with the unchanging centrifugal acceleration due to Earth's rotation would yield a measurable deflection. Eötvös et al. (1922) found no deviation at the 5 part-per-billion level. If indeed, such a deflection is detected, it could also be interpreted as evidence of a non-constant gravitational constant, or an additional non-Newtonian force (Fischbach et al., 1986; Fischbach and Talmadge, 1992).

Another experimental test of Newton's law of gravitation involving gravitational gradients considers the sum of in-line gradients, which for a Newtonian field must sum to zero (Figure 15) in a vacuum (Laplace's Equation 8). Moody and Paik (1993) used a swinging pendulum in the laboratory to induce modulated in-line gradients that were measured with their superconducting gradiometer (see *Gravity, Gradiometry*). The sum of the measured gradients was zero with a statistical variation ($\sigma = 3 \times 10^{-4}$ E) that placed further limits on any local departure from Newton's law of gravitation.



Gravity, Gradiometry, Figure 14 Eötvös's experiment to determine if different types of mass experience different gravitational accelerations (**g** versus **g**′) that could be measured as a deflection, θ, with respect to the constant centrifugal acceleration, **a**.



Gravity, Gradiometry, Figure 15 Experimental setup to test Laplaces's equation by measuring the sum of in-line gradients modulated by the changing gravitational field induced by a swinging pendulum. (From Paik, 2004; reprinted with permission of author.)

The principal utility of the torsion balance and later the gravity gradiometer was and is in geophysical exploration. The Eötvös torsion balance (and its successors) played a significant part, with over a hundred in operation worldwide, in the exploration for oil and gas reservoirs in the early twentieth century, 1918–1940. The achievable accuracy was as good as 1–3 E, but the measurements were arduous, requiring up to 6 h per station. Thus, the torsion balance quickly lost ground when the more portable and robust, as well as sufficiently accurate spring-type gravimeters were developed in the 1930s and 1940s.

Gravity gradiometry in modern times is conducted almost exclusively with mobile systems, in order to take advantage of the increased resolution capability and avoid the difficulty of separating the kinematic and inertial accelerations, required in mobile gravimetry systems (Jekeli, 2000). Submarine gradiometers are used to compensate inertial navigation systems for the effects of gravity, and shipborne gradiometers have been used for marine gravity surveys. A significant investment has been made by a number of commercial exploration companies to use airborne gradiometers to map the gravitational field at high resolution in search for minerals, such as diamonds found in kimberlite pipes (Liu et al., 2001).

A long-standing goal of the space exploration agencies (NASA, ESA, among others) for half a century has been the gravitational mapping of the Earth (and other bodies of the solar system) using a satellite-borne gradiometer. This was realized finally starting in 2009 with the launch of the GOCE satellite whose 2-year mission will provide the highest resolution (100 km) global gravitational model derived solely from a satellite (see *Gravity, Gradiometry*). Such a detailed model of the global field will advance the earth sciences in many fields, from oceanography to regional tectonics and the dynamics of the upper mantle and crust, and to polar ice sheets and glaciology.

Bibliography

Badekas, J., 1967. The horizontal gradients of gravity in southwest Ohio. Report 89, Department of Geodetic Science, Ohio State University, Columbus.

Chapin, D., 1998. Gravity instruments; past, present, future. *The Leading Edge*, **17**, 100–112.

Difrancesco, D., 2007. Advances and challenges in the development and deployment of gravity gradiometer systems. In Proceedings of the EGM 2007 International Workshop Innovation in EM, Grav and Mag Methods: a new Perspective for Exploration, Capri, Italy, April 15 – 18, 2007, http://www2.ogs.trieste.it/eageseg/egm2007/INDEX.pdf.

Drinkwater, M. R., Floberghagen, R., Haagmans, R., Muzi, D., and Popescu, A., 2003. GOCE: ESA's first Earth explorer core mission. In Beutler, G. B., Drinkwater, M., Rummel, R., and von Steiger, R. (eds.), *Earth Gravity Field from Space – from Sensors to Earth Sciences. In the Space Sciences Series of ISSI*. Dordrecht: Kluwer Academic, Vol. 18, pp. 419–432.

Eötvös, L., Pekár, D., and Fekete, E., 1922. Beitrage zum Gesetze der Proportionalität von Trägheit und Gravität. *Ann. Physik*, Series IV, **68**, 11. English translation in: Szabó, Z. (ed.), 1998. *Three Fundamental Papers of Loránd Eötvös*. Loránd Eötvös Geophysical Institute of Hungary, Budapest.

Fischbach, E., and Talmadge, C., 1992. Six years of the fifth force. *Nature*, **356**(6366), 207–215.

Fischbach, E., Sudarsky, D., Szafer, A., Talmadge, C., and Aronson, S. H., 1986. Reanalysis of the Eötvös experiment. *Physical Review Letters*, **56**(1), 3–6.

Flokstra, J., Cuperus, R., Wiegerink, R. J., and van Essen, M. C., 2009. A MEMS-based gravity gradiometer for future planetary missions. *Cryogenics*, **49**(11), 665–668.

Forsberg, R., 1985. Gravity field terrain effect computations by FFT. *Bulletin Goddsiqe*, **59**, 342–360.

Forward, R. L., 1981 Gravitational science from the polar orbiting lunar observatory. Presented at the Gravity Gradiometer Workshop, NASA Headquarters, Washington, DC, 14–15 January 1981.

Heiland, C. A., 1940. *Geophysical Exploration*. New York: Prentics-Hall.

Hein, G., 1977. Die Berechnung von anomalen vertikalen Schweregradienten hoher Präzision. Deutsche Geodätische Kommission, Reihe C, no. 234, München.
Heiskanen, W. A., and Moritz, H., 1967. *Physical Geodesy.* San Francisco: Freeman.
Heller, W. G., 1977. Error models for prototype moving-base gravity gradiometers. Report AFGL-TR-77-0131, by The Analytic Sciences Corporation for Air Force Geophysics Laboratory, Hanscom AFB, MA.
Hofmann-Wellenhof, B., and Moritz, H., 2005. *Physical Geodesy.* Wien: Springer.
Jekeli, C., 1988a. The exact transformation between ellipsoidal and spherical harmonic expansions. *Manuscripta Geodaetica*, **14**, 106–113.
Jekeli, C., 1988b. The gravity gradiometer survey system. *Eos, Transactions, American Geophysical Union*, **69**(8), 105–117.
Jekeli, C., 2000. *Inertial Navigation Systems with Geodetic Applications*. Berlin: Walter deGruyter.
Jekeli, C., 2006. Airborne gradiometry error analysis. *Surveys in Geophysics*, **27**(2), 257–275.
Jekeli, C., 2010. Correlation modeling of the geopotential field in classical geodesy. In Freeden, W. et al. (eds.), *Handbook of Geomathematics.* Berlin: Springer-Verlag, pp. 834–863.
Jekeli, C., and Zhu, L., 2006. Comparison of methods to model the gravitational gradients from topographic data bases. *Geophysical Journal International*, **166**, 999–1014, doi:10.1111/j.1365-246X.2006.03063.x.
Jordan, S. K., 1978. Statistical model for gravity, topography, and density contrasts in the Earth. *Journal of Geophysical Research*, **83**(B4), 186–1824.
Liu, G., Diorio, P., Stone, P., Lockhart, G., Christensen, A., Fitton, N., and Dransfield, M., 2001. Detecting kimberlite pipes at Ekati with airborne gravity gradiometry. In 15th Conference, Australian Society of Exploration Geophysicists, Extended Abstracts, doi:10.1071/ASEG2001ab073.
McGuirk, J. M., Foster, G. T., Fixler, J. B., Snadden, M. J., and Kasevich, M. A., 2002. Sensitive absolute-gravity gradiometry using atom interferometry. *Physical Review A*, **65**, 033608.
Moody, M. V., and Paik, H. J., 1993. Gauss's law test of gravity at short range. *Physical Review Letters*, **70**(9), 1195–1198.
Moody, M. V., Chan, H. A., and Paik, H. J., 1986. Superconducting gravity gradiometer for space and terrestrial applications. *Journal of Applied Physics*, **60**(12), 4308–4315.
Murphy, C. A., 2004. The Air-FTG airborne gravity gradiometer system. In Lane, R. J. L. (ed.), Airborne Gravity 2004 – Abstracts from the ASEG-PESA Airborne Gravity 2004 Workshop: Geoscience Australia Record 2004/18, pp. 7–14.
Nettleton, L. L., 1976. *Gravity and Magnetics in Oil Prospecting*. New York: McGraw-Hill.
Paik, H. J., 2004. Probing extra dimensions using a superconducting gravity gradiometer. Presented at Caltech-JPL Association for Gravitational Wave Research Seminar, 13 April 2004, http://cajagwr.caltech.edu/scripts/seminars.html
Paik, H. J., Leung, J. S., Morgan, S. H., and Parker, J., 1988. Global gravity survey by an orbiting gravity gradiometer. *Eos, Transactions, American Geophysical Union*, **69**(48), 1601–1611.
Parker, R. L., 1972. The rapid calculation of potential anomalies. *Geophysical Journal of the Royal Astronomical Society*, **31**, 447–455.
Pavlis, N. K., Holmes, S. A., Kenyon, S. C., and Factor, J. K., 2008. An Earth Gravitational Model to Degree 2160: EGM2008. Presented at the General Assembly of the European Geosciences Union, Vienna, Austria, April 13–18, 2008.
Pederson, L. B., and Rasmussen, T. M., 1990. The gradient tensor of potential field anomalies – some implications on data collection and data processing of maps. *Geophysics*, **55**(12), 1558–1566.
Petrovic, S., 1996. Determination of the potential of homogeneous polyhedral bodies using line integrals. *Journal of Geodesy*, **71**, 44–52.
Schweitzer, M., Feldman, W. K., Konig, W. F., DiFrancesco, D. J., Sieracki, D. L., and San Giovanni, C. P., 2000. System and process for optimizing gravity gradiometer measurements. U.S. Patent no.6125698. U.S. Patent and Trademark Office.
Smith, D., Robertson, D., and Milbert, D., 2001. Gravitational attraction of local crustal masses in spherical coordinates. *Journal of Geodesy*, **74**, 783–795.
Szabó, Z. (ed.), 1998. *Three Fundamental Papers of Loránd Eötvös*. Loránd Eötvös Geophysical Institute of Hungary, Budapest.
Trageser, M. B., 1970. *A Gradiometer System for Gravity Anomaly Surveying. Report R-588, Charles Stark Draper Laboratory.* Cambridge: MIT Press.
Tryggvason, B., Main, B., and French, B., 2004. A high resolution airborne gravimeter and airborne gravity gradiometer. In Lane, R. J. L. (ed.), Airborne Gravity 2004 – Abstracts from the ASEG-PESA Airborne Gravity 2004 Workshop: Geoscience Australia Record 2004/18, pp. 41–47.
Tsoulis, D., and Petrovic, S., 2001. On the singularities of the gravity field of a homogeneous polyhedral body. *Geophysics*, **66**(2), 535–539.
van Leeuwen, E. H., 2000. BHP develops airborne gravity gradiometer for mineral exploration. *The Leading Edge*, **19**(12), 1296–1297.
Wells, W. C. (ed.), 1984. Spaceborne gravity gradiometers. In *Proceedings of the Spaceborne Gravity Gradiometer Workshop*, NASA Goddard Space Flight Center, Greenbelt MD, 28 February – 2 March 1983, NASA Conference Publication 2305.

Cross-references

Gravity Field of the Earth
Gravity Method, Satellite
Gravity Modeling, Theory and Computation
Gravity, Global Models
Spherical Harmonic Analysis Applied to Potential Fields

GREAT EARTHQUAKES

Roger M. W. Musson
Earthquake Seismology, British Geological Survey, Edinburgh, Scotland, UK

Definition

Any earthquake with magnitude over 8 Mw. Alternatively, any earthquake famous for one reason or another.

Introduction

Under the technical definition, one can find some very obscure "great" earthquakes – a magnitude 8 earthquake in the Southern Ocean is likely to go wholly unremarked by the world's media. Also, given the inherent imprecision in magnitude scales, it is possible for an earthquake with a magnitude of 8.0 to be successively promoted to and demoted from the canon of greatness with each successive revision.

If one took a sample of educated laymen and asked them to name any earthquakes they could think of, one

could be sure to get a few of the most recent high-profile calamities; further back, the results might vary from country to country. In the central USA, probably many are aware of the 1811 New Madrid earthquakes; in central Europe, the name of the 1356 Basel earthquake might come up, and so on. But probably not many could name correctly the largest earthquake ever known, the 1960 Valdariva, Chile, earthquake, which reached 9.6 Mw in magnitude, a value many seismologists suspect to be a physical upper limit of earthquake size. This devastating earthquake also produced a huge tsunami, which, on the Chilean coast, pitched boats up to 3 km inland. (See box-out: Ten Largest Earthquakes).

The title "largest earthquake ever" gives the 1960 Valdariva event an indisputable claim to greatness by any reckoning. Other earthquakes are famous for a variety of reasons: their place in history, the lessons learned from them, their significance in the development of seismology, or their sheer excessive destructiveness. The earthquakes of 1356 and 1811 mentioned above are both important for the same two reasons – they demonstrate that significant earthquakes can still occur well away from plate boundaries, and also that the historical record can reveal potential threats in places not much remarked for seismicity today.

The 1556 Shaanxi earthquake bears the record for being the most deadly earthquake, with about 830,000 dead. The figure of one million dead in one earthquake has not yet been reached (contrary to some sources, the 20 May 1202 earthquake in the Levant did not kill this many; Ambraseys and Melville, 1988), but, with several current megacities dangerously exposed, it is feared that it is only a matter of time. (See box-out: Ten Deadliest Earthquakes).

There are probably three earthquakes that stand out as the most unquestionably great earthquakes of history. Firstly, the 1755 Lisbon earthquake: not only was this (with its accompanying tsunami) one of the most deadly and devastating earthquakes in European history (not forgetting its impact on North Africa), but it sent shockwaves throughout the whole of Western culture. Second, the 1906 San Francisco earthquake so ingrained itself into history that the very name of San Francisco conjures earthquakes to mind; to which must be added the important role this event played in the development of seismological theory. Lastly, the 2004 Sumatra earthquake, staggering in its immensity, changed both the popular and professional understanding of earthquakes.

1 November 1755: the advance of rationalism

Of all great earthquakes, the one with the most lasting impact is certainly the Lisbon earthquake of 1755. Not only did it destroy what was then one of the most important cities in the world, the capital of the Portuguese empire, it shook up the very development of Western philosophy. Furthermore, it continues to be controversial amongst seismologists to this very day.

The first of November is the feast of All Souls, one of the most solemn occasions in the calendar of the Catholic Church. It was on the morning of this day that a massive earthquake struck somewhere off the coast of Portugal. Damage occurred over a wide area across Portugal and western Spain; significant damage occurred even in Morocco. The shaking was perceptible throughout the Iberian Peninsula and into France. Even beyond France, long-period waves caused the agitation of water in ponds and lakes at great distances – as far away as Scotland (Loch Ness was disturbed) and Scandinavia.

The worst damage, however, occurred in Lisbon. The churches, many of which collapsed, were packed with worshippers when the quake struck, and casualties were heavy. The sea then mysteriously withdrew, leaving fishes floundering in the mud. Many ventured out from curiosity, or to collect some of the fish. Shortly afterward the tsunami roared in, surged up into the semi-ruined city, collapsing many buildings that had already been damaged by the earthquake.

All Europe was shocked, and not just by the scale of the destruction. It was axiomatic in Christian thinking that all prodigious events such as earthquakes were of divine origin. The milder ones were sent as a warning of God's displeasure, the more severe ones were a punishment to send mankind back on the path of righteousness. Furthermore, this was true whether or not it was appreciated that such events had natural causes, as many printed works of the previous century explicitly argued: God works through, and controls, the natural world. Whether or not earthquakes were due to natural causes (the dominant explanation at the time was the Aristotelian view that earthquakes were due to subterraneous winds), the timing and location was divinely ordained (Walsham, 1999).

But the injustice of the destruction of Lisbon strained this train of thought to breaking point. Not only did the earthquake occur on one of the holiest days of the Christian year – the godly citizens who were marking the occasion by attending mass were killed, whereas sinners who chose to spend the morning in playhouses or brothels by and large escaped. If this was God's judgment, it was a very strange one. The earthquake was seized upon by Voltaire as a means of attacking the optimistic school of philosophy characterized by Leibniz (Kendrick, 1956). In Lisbon, a decree was issued that anyone found preaching that the earthquake was sent by God, should be put to death.

For modern study of the 1755 earthquake, there are several sorts of information that can be used to try and understand which fault produced it. Firstly, there is the distribution of earthquake effects; for most historical earthquakes, analysis of the distribution of intensity is the predominant means of location, though this is markedly less easy for offshore events. Secondly, for tsunamigenic earthquakes like 1755, the time and run-up of the tsunami at different locations can be inverted to estimate the parameters of the engendering fault rupture. Thirdly, some use can be made of information from

aftershocks. Since these are generally much smaller than the main event, they are usually felt only at short distances. Lastly, in the case of a major earthquake, there should be some geological trace of the structure that produced it.

In terms of available geological structures, there are two main candidates, and a possible third. The first, and most obvious, is the Gorringe Bank, a segment of the offshore plate boundary between Europe and Africa. This was identified as the source of the 1755 earthquake by Johnston (1996); it was also the location of a large earthquake in 1969, and the intensity distribution of the 1969 event resembled to some extent a scaled-down version of the pattern observed from the 1755 event (Levret, 1991).

The second candidate is a more recent discovery – a fault off the coast of Portugal trending NNE-SSW showing signs of recent activity (Baptista et al., 1998). This was named the Marquês de Pombal Thrust in honor of the man who oversaw the reconstruction of Lisbon. This fault provides a good fit to the analyses of tsunami arrivals on the Portuguese coast (Baptista et al., 2003). However, it doesn't seem very likely that an earthquake on this fault would have caused damage in Morocco.

The third is a controversial suggestion by Gutscher et al. (2002) and Gutscher (2004) that the causative feature was an N-S aligned subduction zone in the Gulf of Cadiz, dipping eastward under Gibraltar. Apart from the fact that the very existence of such a feature is disputed, if this was the source, it is hard to understand why Lisbon was destroyed in place of Cadiz. The earthquake is well described in Gibraltar itself, and damage was insignificant.

There are still more problems though. Given that the magnitude of this great earthquake must have been around 8.7 (Johnston, 1996), none of the faults described seem long enough to host such a great earthquake. Recent suggestions to account for this propose multiple near-simultaneous earthquakes: Baptista et al. (2003) and Baptista and Miranda (2005) propose a composite source involving the Marquês de Pombal Thrust and a fault segment on the southern flank of the Guadalquivir Bank; while Vilanova et al. (2003) have the second earthquake on a fault in the Lower Tagus Valley, close to Lisbon. The issue is still unresolved and controversial; the source of this great earthquake is still an enigma.

18 April 1906: the archetype of earthquakes

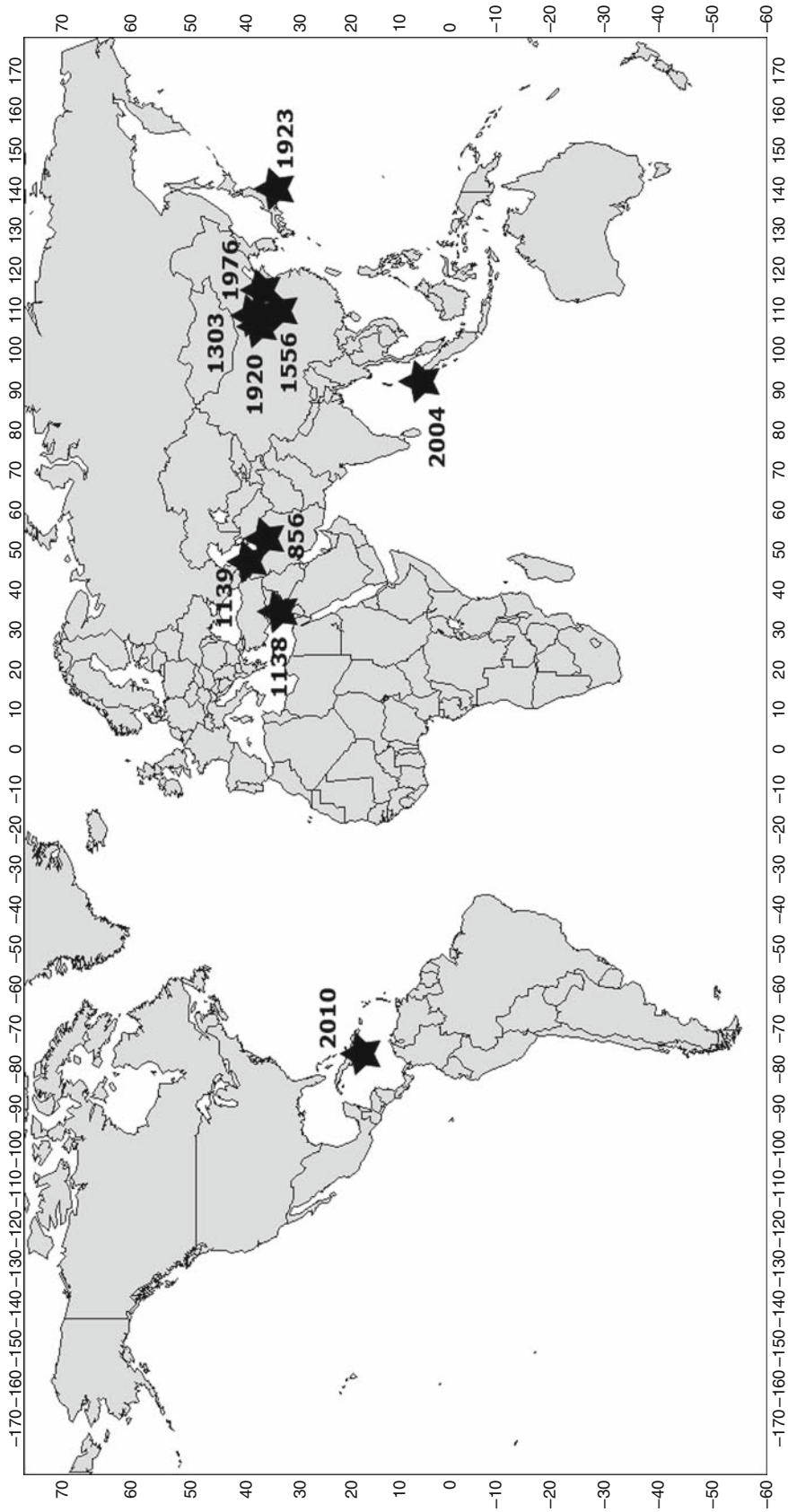
The San Francisco earthquake of 18 April 1906 is of all earthquakes, probably the most iconic in popular culture; the one earthquake that practically any layman can mention. As Davison (1927) remarked, "If the Californian earthquake were the only shock known to mankind, the attention paid to it could hardly have been more exclusive." Not only was it a major disaster in a romantic locale, it even had a celebrity cast – Caruso being in the city at the time. Thanks to the telegraph, the news spread quickly round the world, much faster than had been the case in 1755.

As with the Lisbon earthquake, fire was a contributor to the damage; but this time, the major contributor. Earthquakes are great fire starters, upsetting stoves, breaking gas pipes. But they also hinder fire-fighting efforts. As the fire gripped the city, much of which consisted of wooden buildings, it was found that the water supply system was also damaged by the quake. Unable to use hosepipes, the city authorities resorted to trying to control the fire with a system of firebreaks. Unfortunately, this was carried out somewhat ineptly. The black powder that was used often simply set fire to the building being blasted; and where it didn't, it blew houses into clouds of matchwood that easily ignited – the firebreaks being often created too close to the advancing flames. Furthermore, no attempt was made to check what the buildings marked for demolition actually contained. One such building, belonging to the Viavi pharmaceuticals company, contained a huge store of spirit alcohol. The resulting fireball so exacerbated the fire that probably another 50 blocks in the northern part of the city were destroyed as a result (Fradkin, 2006).

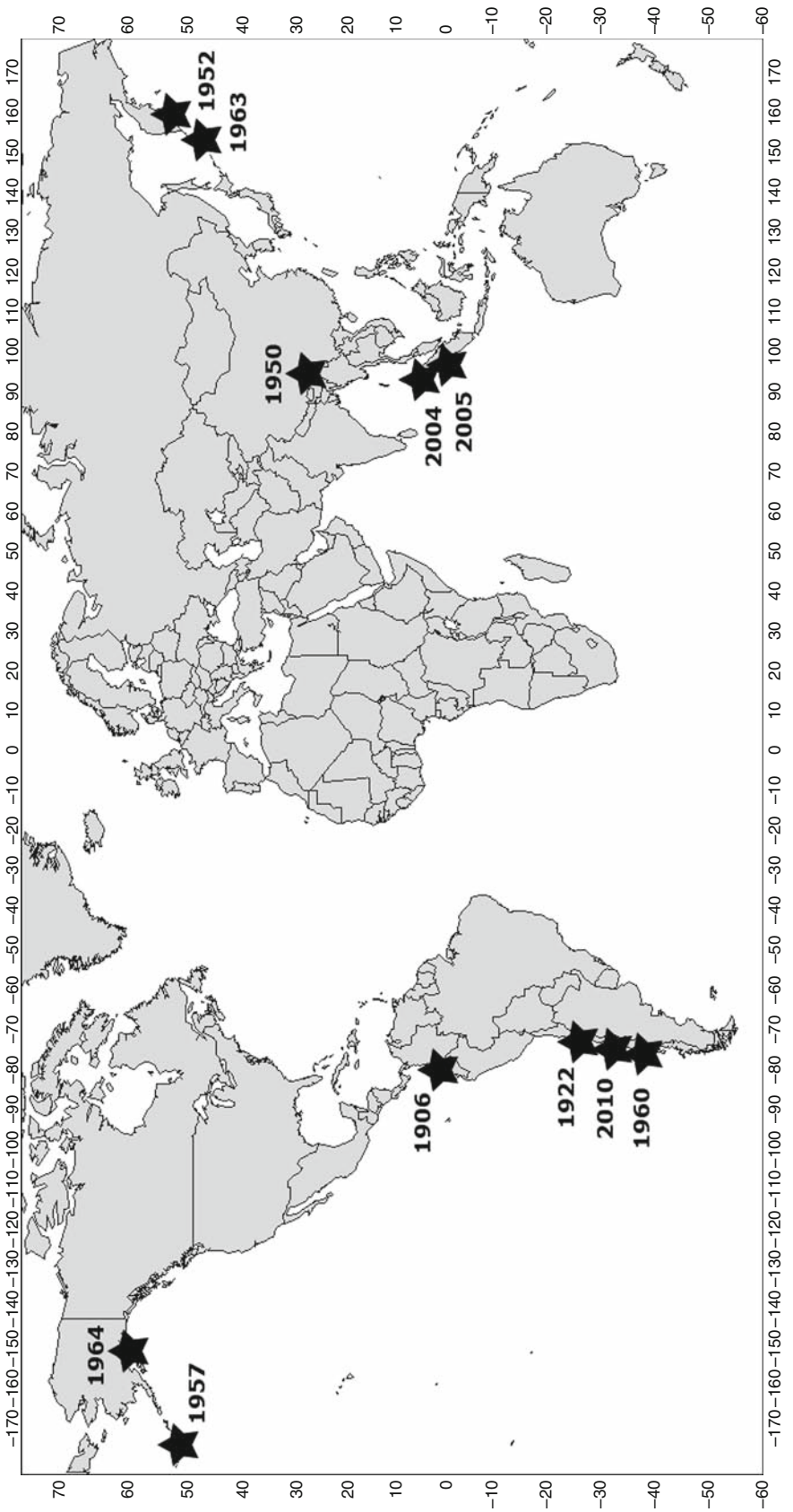
The final death toll is a matter of dispute; initial figures of around 700 are clearly underestimates, partly for propaganda purposes (reporters were encouraged to call it the Great San Francisco Fire and play down mention of the earthquake). Current estimates are about 3,000 dead (Stover and Coffman, 1993).

From a seismological point of view, the 1906 earthquake is rather more straightforward than the 1755 event. The fault it occurred on is not just known, it is perhaps the most famous fault in the world: the San Andreas, forming the plate boundary between the North American and Pacific plates. This is a right-lateral strike-slip fault; the fault rupture extended probably around 430 km including the offshore extension, with a slip of over 6 m in places (Lawson, 1908–1910; Segall and Liskowski, 1990; Stover and Coffman, 1993). Since strike-slip faults are limited in width to the thickness of the crust, this gives a magnitude of only around 7.8 Mw. The felt area was much less than that of the Lisbon earthquake, covering only California and extending partially into Nevada and Oregon (attenuation of seismic shaking is higher in California than in Iberia).

The dislocation of fence posts across the fault trace made it very obvious that the earthquake was due to fault displacement. It was studying the 1906 rupture that led Reid (1910) to develop the theory of elastic rebound, which still underpins basic seismological theory today – that a fault gradually deforms due to the build-up of crustal stresses, until a locking asperity breaks, releasing energy as the fault violently snaps. The connection between earthquakes and faulting was well-known at the time; Hoernes (1878) held that the majority of earthquakes were tectonic in origin, due to displacement along fractures related to mountain building. As early as 1855, Lyell identified repeated earthquake-related faulting in New Zealand (Lyell, 1856), but it is not clear that he understood that the fault produces the earthquake and not vice versa. The earliest statement that dislocation on a fault actually



Great Earthquakes, Figure 1 Map of the ten deadliest earthquakes in history.



Great Earthquakes, Figure 2 Map of the ten largest earthquakes since 1900 (actually 11 due to a tie for the tenth place).

causes earthquake shaking seems to be Mallet (1851), but he was ahead of his time.

However, Reid's fundamental understanding began a new era in seismology, opening the door to the interpretation of seismograms in terms of fault mechanics (Wald et al., 1993; Agnew, 2002).

26 December 2004: cataclysm

Although most of the world's subduction zones are around the Pacific, there are two in the Indian Ocean: on the Makran coast of Iran and along the south coast of Sumatra (extending to Burma in the north and bending round to Java in the east). While the historical seismicity of Sumatra (including a major event in 1833) was by no means unknown (Newcomb and McCann, 1987), it was not generally considered high on the list of potential localities for catastrophe before the earthquake of Boxing Day 2004. The epicenter was off the Sumatran coast at the northern end of the island, near Bandeh Aceh, and the rupture propagated almost entirely northward, for a staggering distance of 1,600 km (Nalbant et al., 2005), reaching up to the Andaman Islands. The northern part of the rupture zone, where the plate boundary is extremely oblique to the direction of convergence, was almost north–south in orientation. The resulting tsunami, instead of heading off into the southern Indian Ocean as occurred in 1833, headed due east and west, hitting crowded tourist beaches in Thailand about 15 min later (shortly after 8 a.m. local time) and India and Sri Lanka 2 h after that.

The effect was literally a cataclysm (Gr: down-washing), as the waves swept inland up to about the 30 m mark in some places, causing immense destruction. About 230,000 lost their lives, of which 167,000 were in Indonesia. It tends to be the case that most high-fatality earthquakes kill local people in substandard housing; this event was exceptional in killing large numbers of foreign tourists, which accounts in part for the great impact of the event on public consciousness in the West. (For a while afterward, journalists would enquire of seismologists "will there be a tsunami?" even in cases for earthquakes in central Russia.)

The result was a sudden impetus toward installing tsunami warning systems in different parts of the world, though perhaps not always with the understanding that technical solutions are of limited use without the practical means to evacuate people at risk in a timely fashion.

Given that no earthquake anything like as large had occurred anywhere in the world since the 1960s, in the following days, the analysis of the earthquake tested the methods of modern seismology in new ways. The shock waves circled the globe, being distinctly recognizable even on the third pass, and even altered the rotation of the planet itself, the length of day dropping by up to −5μs (Bizouard, 2005). The pole itself shifted by up to 2 mas (milli-arc seconds) toward 90°E (Bizouard, 2005; 1 mas equals an equatorial shift at the Earth's surface of 3.1 cm). The difficulty for analysts was caused by the excessive duration of the rupture; the unidirectional propagation of the rupture took about 10 min to reach its terminus. This complicated the seismic records as waves propagating from different parts of the rupture at different times over-wrote each other. The immediate impact was the underestimation of the earthquake's magnitude – immediate press releases giving values of 8.5 or 8.7. Initial attempts to model the slip distribution could only manage the southernmost part of the rupture (Ji, 2004). The first complete slip models were made by Ammon et al. (2005) from a combination of body waves, intermediate period waves and long period waves, and Vigny et al. (2005) using an inversion of GPS data (the two models disagree extensively).

Previously, it had been thought that such a large earthquake was impossible on this part of the Sunda thrust system, on tectonic grounds, and due to parts of the system having already broken in previous large earthquakes. The fact that the 2004 earthquake ruptured through segments that had already recently ruptured has caused seismologists to reassess the risk of great thrust earthquakes in other areas (e.g., Suárez and Albini, 2009).

It was speculated at the time that the movement due to the earthquake would increase the stress on the Great Sumatran Fault, a strike-slip fault running the length of the island, and bring this close to failure (Nalbant et al., 2005). As of the time of writing, the expected event on this fault has not yet occurred. Instead, a remarkable cascade of events occurred on segments of the thrust zone south of the 2004 rupture, including further great earthquakes on 25 March 2005 and 12 September 2007 with magnitudes $\geq$8.5 Mw. Current estimates (e.g., Megawati and Pan, 2009) are that there is still a potential slip in this part of the system to make further great earthquakes on the coast of Sumatra a continuing threat.

The ten deadliest earthquakes

A number of lists have appeared over the years purporting to give the ten earthquakes with the highest number of fatalities. Many of these are prone to easily avoidable errors, such as perpetuating fake earthquakes like the notorious 1737 Calcutta event (see Bilham 1994 for the refutation of this) or the 893 Ardabil (Iran) earthquake. However, compiling an accurate list is difficult, since for even relatively modern earthquakes, death tolls are often inaccurate, given the sheer difficulty of making a body count in the aftermath of a major disaster (as witness the difficulty of estimating the death toll in the 2010 Haiti earthquake). The problem is greatly multiplied for historical earthquakes, where sources are vague, unspecific, or exaggerated, and in most cases, there was no census to begin with. The earthquake generally agreed to be the deadliest of all time is the 1556 Shaanxi earthquake, with 830,000 deaths. This great earthquake (M ~ 8) affected an area where many people were living in loess caves, which collapsed, burying the inhabitants.

The second deadliest is probably the 1976 Tangshan disaster. The death toll from this event has been a matter of controversy; the earthquake occurred during the Cultural Revolution in China, when news about disasters was generally suppressed. Shortly after

the earthquake, a suspiciously accurate figure of 655,237 deaths appeared in the Hong Kong media (South China Morning Post, 5 January 1977). The official death toll later released was 242,419 (Chen, 1988), though Marza (2004) suggests the true figure may have been over a million.

There are a number of earthquakes with death tolls in the region of 200,000–300,000 that are hard to discriminate between. Willis (1928) attributes 260,000 deaths to the Syrian earthquake of 115, but a comprehensive review of the sources by Ambraseys (2009) shows that while the death toll must have been very high, the total figure is unknown. There are several candidates for third place. The 1920 Ningxia earthquake killed 235,502 according to Utsu (2002) and 160,000 according to Tiedemann (1992). The official death toll of the 26 December 2004 (9.2 Mw) Sumatran earthquake is 227,898, mostly from the tsunami. The 1139 earthquake in Ganjak, Azerbaijan is considered by Ambraseys and Adams (1989) to have killed 230,000, a figure rejected by Guidoboni and Comastri (2005) as exaggerated (considering the total population of the affected area). Finally, the 12 January 2010 Haiti earthquake, at the time of writing, is believed to have killed around 220,000–230,000.

Events with death tolls around 200,000 include another Chinese earthquake, in Shanxi province in 1303, with a death toll around 200,000 (Fu et al., 2005), and the 856 Qumis earthquake in Iran, believed to have killed around 200,000 (Ambraseys and Melville, 1982).

Next highest is undoubtedly the 1923 Kanto (Japan) earthquake with 143,000. The only other earthquake (leaving aside events with very doubtful high death tolls) with attributed deaths over 100,000 is the 1138 Aleppo earthquake, but this is another case where, according to Ambraseys (2009), the actual figure is quite unknown (Ambraseys, 1962 gives 130,000).

If one removed 1138 and 1139 from the list as too doubtful, the next two events would be the Sichuan earthquake of 2008 with 87,587, and the 1908 Messina earthquake in Southern Italy (82,000).

This gives us the following list (all magnitudes are moment magnitude, Mw):

1.	23 January 1556	~8	Shaanxi (830,000)
2.	27 July 1976	7.5	Tangshan (240,000)
3.	16 December 1920	7.8	Ningxia (235,000)
4.	26 December 2004	9.2	Sumatra (230,000)
5.	12 January 2010	7.0	Haiti (230,000)
6.	17 September 1303	~8	Shanxi (200,000)
7.	22 December 856	~7	Qumis (200,000)
8.	25 September 1139	~7	Azerbaijan (high)
9.	1 September 1923	7.9	Kanto (143,000)
10.	15 October 1138	~7	Aleppo (130,000)

Numbers of fatalities are rounded to the nearest 1,000; the precise ranking is obviously conjectural, as the figures are really only indicative of general loss levels. The distribution is shown in Figure 1.

The ten largest earthquakes

Listing the ten largest earthquakes presents different problems. For the period since 1900, it is possible, at least in theory, to provide indisputable magnitude values from instrumental data, although figures are not precise for the very largest events. For the pre-instrumental era, the situation is much harder. Early attempts to convert historical data into estimates of magnitude tended to rely on maximum intensity; however, this rapidly saturates in large earthquakes. More sophisticated methods based on intensity data fields are available (e.g., Sibol et al., 1987) but these are unlikely to be accurate for very large events (the lack of calibration is one problem).

Earthquakes around the 9.0 Mw mark are necessarily subduction events – only large subduction thrust planes can provide large enough rupture areas to accommodate the vast energies required. Estimation of magnitude from noninstrumental data in such cases cannot rely on intensity distribution, which is liable to be incomplete, but really needs to be based on estimating rupture area and displacement. It was possible, for instance, on the morning of 26 December 2004, to see at once that the great Sumatran earthquake was at least 9 Mw just by looking at the huge extent of the aftershock distribution, which stretched for over 1,000 km.

Since 1900, one can at least be certain that the catalog for very large earthquakes is complete. Taking the Centennial Catalogue of Engdahl and Villaseñor (2002) as a basis, one can list the ten largest earthquakes since 1900 as follows (all magnitudes are moment magnitude, Mw):

1.	22 May 1960	9.6	Chile
2=.	28 March 1964	9.2	Alaska
2=.	26 December 2004	9.2	Sumatra
4.	9 March 1957	9.1	Andreanof Islands
5.	4 November 1952	9.0	Kamchatka
6=.	31 January 1906	8.8	Colombia-Ecuador
6=.	27 February 2010	8.8	Chile
8=.	11 November 1922	8.7	Chile
8=.	25 March 2005	8.7	Sumatra
10=.	15 August 1950	8.6	Assam-Tibet
10=.	13 October 1963	8.6	Kuril Islands

The 2005 Sumatra earthquake, while not strictly an aftershock of the 26 December 2004 event, was certainly triggered by it. It was a very strange experience to be telling journalists that it was "only" 8.7 and small in comparison to the Boxing Day earthquake. It is somewhat remarkable that all the magnitude nines in the twentieth century, accounting for most of the seismic energy release of the whole century, occurred in the space of just 12 years, from 1952 to 1964. The years 2004–2010 appear to be making a similar spike in the energy release record.

The distribution is shown in Figure 2, which shows a very different pattern from Figure 1. The 2004 Sumatra earthquake is the only one to appear on both lists. This should emphasize that the deadliness of an earthquake generally owes more to its location with respect to population centers, than to its size.

Acknowledgments

The author would like to thank Professor N.N. Ambraseys and Dr. D.J. Kerridge for helpful reviews of the text.

Bibliography

Agnew, D. C., 2002. History of seismology. In Lee, W. H. K., Kanamori, H., Jennings, P. C., and Kisslinger, C. (eds.), *International Handbook of Earthquake and Engineering Seismology*. San Diego: Academic, pp. 3–12.

Ambraseys, N. N., 1962. A note on the chronology of Willis' list of earthquakes in Palestine and Syria. *Bulletin. Seismological Society of America*, **52**, 77–80.
Ambraseys, N. N., 2009. *Earthquakes in the Mediterranean and Middle East*. Cambridge: Cambridge University Press.
Ambraseys, N. N., and Adams, R. D., 1989. Long-term seismicity of Northern Armenia confirmed by the historical record. *Eos*, **70**, 152–154.
Ambraseys, N. N., and Melville, C. P., 1982. *A History of Persian Earthquakes*. Cambridge: Cambridge University Press.
Ambraseys, N. N., and Melville, C. P., 1988. An analysis of the Eastern Mediterranean earthquake of 20 May 1202. In Lee, W. H. K., Meyers, H., and Shimazaki, K. (eds.), *Historical Seismograms and Earthquakes of the World*. San Diego: Academic, pp. 181–200.
Ammon, C. J., et al., 2005. Rupture process of the 2004 Sumatra–Andaman earthquake. *Science*, **308**, 1133–1139.
Baptista, M. A., and Miranda, J. M., 2005. Evaluation of the 1755 earthquake source using tsunami modelling. In *International Conference of the 250th Anniversary of the 1755 Lisbon Earthquake, Proceedings*, Lisbon, pp. 574–577.
Baptista, M. A., Miranda, P. M. A., Miranda, J. M., and Mendes-Victor, L., 1998. Constraints on the source of the 1755 Lisbon tsunami inferred from numerical modelling of historical data. *Journal of Geodynamics*, **25**(2), 159–174.
Baptista, M. A., Miranda, J. M., Chierici, F., and Zitellini, N., 2003. New study of the 1755 earthquake source based on multi-channel seismic survey data and tsunami modelling. *Natural Hazards and Earth System Sciences*, **3**, 333–340.
Bilham, R., 1994. The 1737 Calcutta earthquake and cyclone evaluated. *Bulletin. Seismological Society of America*, **84**(5), 1650–1657.
Bizouard, C., 2005. Influence of the earthquakes on the polar motion with emphasis on the Sumatra event. In *Journées Systèmes de Référence Spatio-Temporels, Proceedings.* Warsaw, pp. 229–232.
Chen, Y., 1988. *The Great Tangshan Earthquake of 1976: An Anatomy of Disaster*. New York: Pergamon.
Davison, C., 1927. *The Founders of Seismology.* Cambridge: Cambridge University Press.
Engdahl, E. R., and Villasenor, A., 2002. Global seismicity: 1900–1999. In Lee, W. H. K., Kanamori, H., Jennings, P. C., and Kisslinger, C. (eds.), *International Handbook of Earthquake and Engineering Seismology.* San Diego: Academic, pp. 665–690.
Fradkin, P. L., 2006. *The Great Earthquake and Firestorms of 1906*. Berkeley CA: University of California Press.
Fu, Z.-X., Liu, G.-P., Shao, H.-C., and Ding, X., 2005. An overview on shallow strong earthquake activity and earthquake live losses of Chinese mainland in the centenary from 1901 to 2001. *Acta Seismologica Sinica*, **18**(4), 392–401.
Guidoboni, E., and Comastri, A., 2005. *Catalogue of Earthquakes and Tsunamis in the Mediterranean Area from the 11th to the 15th Century.* Bologna: INGV.
Gutscher, M.-A., 2004. What caused the Great Lisbon earthquake? *Science*, **305**, 1247–1248.
Gutscher, M.-A., et al., 2002. Evidence for active subduction beneath Gibraltar. *Geology*, **30**, 1071–1074.
Hoernes, R., 1878. Erdbeben studien. *Jahrbuch der Kaiserlich-Königlichen Geologischen Reichsanstalt*, **28**, 387–448.
Ji, C., 2004. Preliminary result of the 04/12/26 (Mw 9.0), off W coast of northern Sumatra earthquake (http://www.gps.caltech.edu/~jichen/Earthquake/2004/aceh/aceh.html).
Johnston, A. C., 1996. Seismic moment assessment of earthquakes in stable continental regions. III. New Madrid 1811–1812, Charleston 1886, and Lisbon 1755. *Geophysical Journal International*, **126**, 314–344.
Kendrick, T. D., 1956. *The Lisbon Earthquake*. London: Methuen.
Lawson, A. C. (ed.), 1908–1910. *The California Earthquake of April 18, 1906: Report of the State Earthquake Investigation Commission*. Washington, DC: Carnegie Institution of Washington, Vol. 2.
Levret, A., 1991. The Effects of the November 1, 1755 Lisbon earthquake in Morocco. *Tectonophysics*, **193**, 83–94.
Lyell, C., 1856. Sur les effets du tremblement de terre du 23 Janvier, 1855, à la Nouvell Zélande. *Bulletin de la Societe Geologique de France*, **13**, 661–667.
Mallet, R., 1851. First report on the facts of earthquake phaenomena. *Report of the British Association of the Advancement of Science*, **20**, 1–89.
Marza, V., 2004. On the death toll of the 1999 Izmit (Turkey) major earthquake. In *European Seismological Commission, Proceedings.* Potsdam.
Megawati, K., and Pan, T. C., 2009. Regional seismic hazard posed by the Mentawai Segment of the Sumatran Megathrust. *Bulletin. Seismological Society of America*, **99**, 566–584.
Nalbant, S. S., Steacy, S., Sieh, K., Natawidjaja, D. H., and McCloskey, J., 2005. Earthquake risk on the Sunda trench. *Nature*, **435**, 756–757.
Newcomb, K. R., and McCann, W. R., 1987. Seismic history and seismotectonics of the Sunda Arc. *Journal of Geophysical Research* **92**(B1), 421–439.
Reid, H. F., 1910. The mechanics of the earthquake. In Lawson, A. C. (ed.), *The California Earthquake of April 18, 1906: Report of the State Earthquake Investigation Commission*. Washington, DC: Carnegie Institution of Washington.
Segall, P., and Liskowski, M., 1990. Comparison of surface displacements in the 1906 San Francisco and 1989 Loma Prieta earthquakes. *Science*, **250**, 1241–1244.
Sibol, M. S., Bollinger, G. A., and Birch, J. B., 1987. Estimations of magnitudes in central and eastern North America using intensity and felt area. *Bulletin. Seismological Society of America*, **77**, 1635–1654.
Stover, C. W., and Coffman, J. L., 1993. *Seismicity of the United States, 1568–1989 (Revised)*. Washington: United States Government Printing Office.
Suárez, G., and Albini, P., 2009. Evidence for great tsunamigenic earthquakes (M 8.6) along the Mexican subduction zone. *Bulletin. Seismological Society of America*, **99**, 892–896.
Tiedemann, H., 1992. *Earthquakes and Volcanic Eruptions: A Handbook on Risk Assessment*. Zurich: Swiss Re. 951 pp.
Utsu, T., 2002. A list of deadly earthquakes in the World: 1500–2000. In Lee, W. H. K., Kanamori, H., Jennings, P. C., and Kisslinger, C. (eds.), *International Handbook of Earthquake and Engineering Seismology.* San Diego: Academic, pp. 691–718.
Vigny, C., et al., 2005. Insight into the 2004 Sumatra–Andaman earthquake from GPS measurements in southeast Asia. *Nature*, **436**, 201–206.
Vilanova, S. P., Nunes, C. F., and Fonseca, J. F. B. D., 2003. Lisbon 1755: a case of triggered onshore rupture? *Bulletin. Seismological Society of America*, **93**(5), 2056–2068.
Wald, D. J., Kanamori, H., Helmberger, D., and Heaton, T. H., 1993. Source study of the 1906 San Francisco earthquake. *Bulletin. Seismological Society of America*, **83**, 981–1019.
Walsham, A., 1999. *Providence in Early Modern England*. Oxford: Oxford University Press.
Willis, B., 1928. Earthquakes in the holy land. *Bulletin. Seismological Society of America*, **18**, 73–103.

Cross-references

Earthquake, Magnitude
Earthquakes, Energy
Earthquakes, Intensity
Earthquakes, Shake Map
Tsunami

H

HEAT FLOW MEASUREMENTS, CONTINENTAL

John H. Sass[1], Graeme Beardsmore[2,3]
[1]Great Lakes Geothermal Services, Port Huron, MI, USA
[2]Monash University, Melbourne, Australia
[3]Hot Dry Rocks Pty Ltd, Victoria, Australia

Definition

Continental heat flow is the flux density of heat being conducted to the surface of continents. It is expressed as mW m^{-2}. Heat flow is calculated as the product of temperature gradient (mK m^{-1} or °C per km – a scale appropriate to the dimensions of geological formations) and thermal conductivity (W m^{-1} K^{-1}).

Introduction

Terrestrial heat flow, Q, is estimated from the product of temperature gradient (Γ) and thermal conductivity (K), according to:

$$Q = K.\Gamma \tag{1}$$

In practice, Γ is derived from point measurements of temperature at two or more discrete depths. K can be determined in situ, but for routine determinations of heat flow, the most practical approach is to measure it on the surface, either in the field or in a laboratory.

Beardsmore and Cull (2001) reviewed the various options for applying equation 1. These usually involve combining the harmonic mean thermal conductivity and temperature gradient over discrete intervals of the temperature profiles. Where there is significant thermal conductivity stratification, the "Bullard plot" is a useful method for calculating average heat flow (see Figure 1). In this method, temperature is plotted against integrated thermal resistance, R, where $R = z/K$, z being the thickness of the interval with mean conductivity K. If heat flow is constant with depth, the relation is linear and the slope of the line is the heat flow.

Uncertainty is inherent in heat flow determinations:

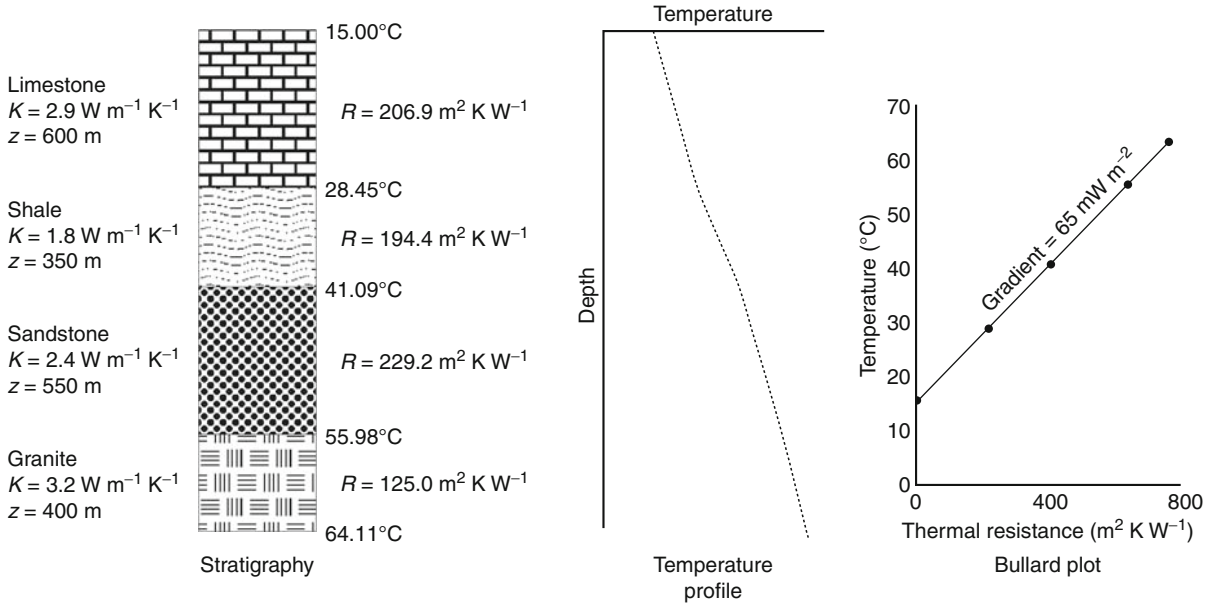
- As thermal gradient estimates become more precise with increasing depth, thermal conductivity estimates become less precise.
- The average thermal gradient and the average thermal conductivity must be determined over the same depth interval for a precise determination of heat flow.
- Because it is rarely practical to measure thermal conductivity over an entire depth interval, assumptions must be made as to how representative the measured samples are of the entire depth interval.
- The thermal disturbance due to seasonal surface temperature cycles may extend to a depth of several tens of meters, so meaningful results can be obtained only at greater depths.
- The diffusion of step changes in surface temperature (e.g., from the retreat of glaciation) may affect heat flow measurements down to hundreds of meters or more.
- Continental heat flow measurements, especially when compared to measurements of heat flow beneath the oceans, are complicated by significant lateral and temporal variations in surface temperature.
- Surface heat flow on the continents is more likely than oceanic heat flow to be distorted by near-surface variations in geological structures and topography.

Derived heat flow values may vary depending on the depth and sampling strategy of the survey, so careful thought should precede the design of any survey aimed at quantifying regional terrestrial heat flow.

Temperature

The earliest measurements of temperature beneath the earth's surface were obtained from mines and tunnels. As exploratory drilling for petroleum and minerals

Harsh K. Gupta (ed.), *Encyclopedia of Solid Earth Geophysics*, DOI 10.1007/978-90-481-8702-7,



Heat Flow Measurements, Continental, Figure 1 A hypothetical rock sequence with thermal properties and unit thicknesses. The thermal resistance, *R*, of each unit is equal to its thickness, *z*, divided by its thermal conductivity, K. A temperature–depth profile through the sequence is nonlinear, but a Bullard plot of temperature versus integrated thermal resistance is linear with slope equal to the conductive heat flow. Heat flow in this example is 65 mW m^{-2}.

became widespread, temperature profiles obtained in boreholes became the most common means of determining temperature gradients within the earth. Almost exclusively, electronic thermometers deployed on a wire line are currently used to measure temperatures within the earth.

Temperature and depth measurements for heat flow determinations need to be precise and accurate. Best practice calls for a precision of ~0.001°C at ~1 m intervals over several hundred meters. Common transducers include the following types:

- Thermistors: A thermistor is a bead of sintered metallic oxide, commonly germanium. It has a large negative temperature coefficient of resistance (typically −4% per K), and is precise and accurate to 1 mK using simple, unsophisticated readout devices like off-balance Wheatstone bridges and 4½ digit multimeters (DMMs) (McGee and Thomas, 1988). Thermistor probes for use in "normal" geothermal environments usually have a resistance of between 3 and 100 KΩ at 25°C and provide useable readouts between about 0°C and 80°C, above which their temperature sensitivity drops off because of low resistance. Thermistors must be individually calibrated.
- Platinum Resistance Elements (RTDs): Platinum RTDs have a small, positive temperature coefficient of resistance and depending upon their construction and configuration, can operate successfully up to 500°C (see Tew and Strauss, 2001, for principles and standards). Millidegree accuracy can be obtained with RTDs if 6½ digit DMMs are employed as readouts. RTD probes for geothermal applications typically have a resistance of 1 or 2 KΩ at 20°C. One advantage over thermistors is the interchangeability of RTD probes within a batch without the need for individual calibration.
- Diodes: Effective temperature sensors can be constructed from ordinary semiconductor diodes. Such sensors have a linear voltage or current response with temperature, are of relatively small size, operate over a limited temperature range (typically −40°C to +120°C), are of low cost, and are quite accurate if individually calibrated. Diode temperature sensors tend to be less robust (electrically and physically) than other sensor types, which tends to rule them out of most logging applications. For stationary applications, however, commercial semiconductor sensors are available that provide an output signal as a voltage, a current, a resistance, or as digital data.
- Fiber Optics: Distributed fiber-optic measurement of temperature using the phenomenon of backscattering is possible using commercially available instruments with temperature resolution down to 0.01°C (Selker et al., 2006). Instantaneous records of temperature distribution along entire lengths of fiber-optic cable at intervals of ~1 m up to 30,000 m can be obtained at time intervals of fractions of a minute. Temperature resolution increases relative to the length of time over which each measurement is recorded. This technology holds great promise for repeated temperature profiles in a well recovering from drilling-associated thermal

transients, or for monitoring the thermal effects of pumping or injection in "Hot Dry Rock" applications. Förster et al. (1997) presented a comparison between fiber-optic and conventional temperature logs.

Temperature Tools include the following types:

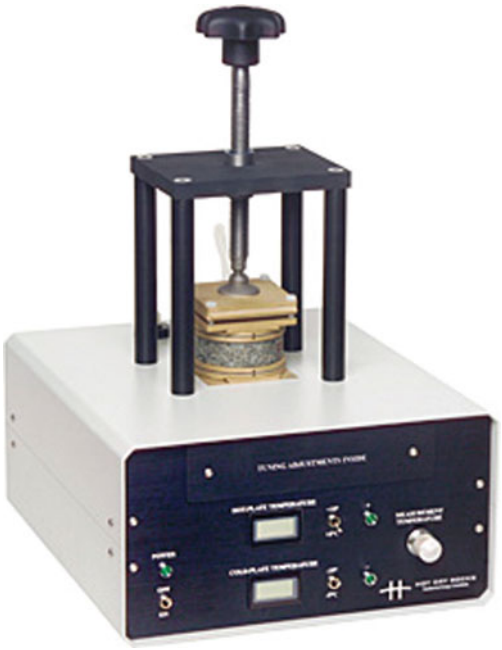
- Four-conductor connection to the surface. This mode provides for accurate measurement of the resistance of the temperature transducer. It requires that the leakage resistance between conductors be effectively infinite, a specification that can be met by manufacturers of geophysical logging cables. Significant leakage between conductors is usually manifested as noise on the detector. The relatively high resistance of thermistors with respect to typical logging cable makes them the most suitable transducers for these tools.
- Down-hole oscillator with single-conductor transmission. Where suitable multi-conductor cables are not available, the resistance element can be incorporated into an oscillator circuit with temperature–frequency dependence, independent of leakage between conductors (Doig et al., 1961).
- "Memory tools." With or without onboard pressure gauges, memory tools are deployed on a wire line. They record down hole and store data in onboard memory. The data are downloaded and read on the surface (Suto et al., 2008).
- Fiber-optic cables: Henninges et al. (2005) described the deployment of fiber-optic cables for both synoptic temperature logs and long-term monitoring of temperature.

Thermal conductivity

There are two basic approaches for thermal conductivity measurement:

1. Steady State (divided bar)
2. Transient (point-, line- or plane- source)

The concept of the divided bar was originally introduced by Prof. C.H. Lees in the nineteenth century. The device is essentially a vertical "stack" within which the temperature drop across a standard substance is compared with that across a disc of unknown thermal conductivity. Early versions were introduced in the 1930s by Prof. E.C. Bullard at Cambridge and by Prof. Francis Birch at Harvard. These adaptations featured a constant heat source at the top of the "stack" and an isothermal sink at the base. Adoption of a constant temperature difference resulted in more rapid measurements (Beck, 1957). Substitution of Peltier devices for the constant temperature water baths usually used provides hitherto unattainable portability, economy and speed of measurement (Antriasian, 2010; Figure 2). In the late 1960s, the divided bar was adapted to the measurement of the thermal conductivity of rock fragments, by far the most common source of material (Sass et al., 1971).



Heat Flow Measurements, Continental,
Figure 2 A commercial divided bar utilizing Peltier devices as heat source/sink. (Photo courtesy of Hot Dry Rocks Pty Ltd, Australia.)

Transient heat source methods can be applied in a permanent laboratory, in mobile field laboratories, or in certain circumstance within boreholes. In these techniques, the rock is heated, and the temperature-time curve of the heater element or a nearby point is analyzed to obtain the thermal conductivity. In general, higher conductivity rocks conduct heat away from the heater at a greater rate and suppress the increase in temperature.

Line- or cylindrical- source methods present difficulties in characterizing uniaxial (e.g., vertical) conductivity because they conduct heat radially, rather than linearly, away from the source. In homogenous sequences, conductivity is the same in all directions, but the values may be significantly different in heterogeneous sequences. Point source methods overcome this to some extent but require higher operating power and may initiate convection (thus negating the assumption of pure conduction).

J.C. Jaeger pioneered the concept of line- and cylindrical heat sources for the measurement of thermal conductivity. Early attempts at implementation of his models called for ingenuity in the use of analog recording devices of fairly low sensitivity. Line-source methods in common use include variations on the "needle probe" and half-space techniques; e.g., Showa Denko (Ito et al., 1977) and GeoForschungZentrum (GFZ) line-source units, see e.g., Pribnow and Sass (1995), which also has a description of the United States Geological Survey (USGS) divided bar apparatus. Both units are available commercially. K. Horai (1971) adapted the line-source

technique to measurements on fragments. The USGS versions of line-source techniques were described in detail by Sass et al. (1984).

Optical scanning

In the 1980s, Yuri Popov developed a rapid, nondestructive method for measuring the thermal conductivity of large numbers of rocks using an optical scanning technique. Flat and black-painted rock faces are aligned with a known standard material along a track. A device containing a laser heat source and an infrared radiometer temperature sensor moves at a constant speed across the surface of the samples. The radiometer measures the rise in temperature of the samples due to the applied laser pulse, and the temperature rise is then related back to the thermal conductivity of the rocks. This method was initially met with skepticism in many laboratories, but a three-way comparison among Popov's method, the USGS divided bar, and the GFZ line-source half-space apparatus resulted in remarkable agreement over a large range of thermal conductivity and a variety of crystalline rock types (Popov et al., 1999).

In situ thermal conductivity measurement in boreholes

The in situ measurement of thermal conductivity in vertical boreholes generally presents great experimental and logistical difficulties. In anisotropic rocks, it also measures conductivity in the wrong direction so that independent measurements of anisotropy on core are required to calculate vertical conductivity. Conductivity was measured in situ over a ~2 m interval of the KTB pilot core hole (Burkhardt et al., 1990) as a demonstration of the feasibility of the method.

During the boom in geothermal exploration in the mid- to late 1970s, the importance of determining heat flow, as well as temperature gradients became apparent. This was coupled with concerns over the security of exploration wells in competitive lease areas. To address these issues, the Geothermal Studies Project of the U.S. Geological Survey developed a method for determining temperature gradient, thermal conductivity, and thus, heat flow in situ over discrete intervals during the drilling process (Sass et al., 1981)

Details of the technology can be found in Sass et al. (1979) with the caveat that the electronics and computer hardware described in that reference have since been superseded.

More recently, Freifeld et al. (2008) have demonstrated thermal conductivity profiling of entire boreholes by coupling a down-hole line-source heating loop with a distributed fiber-optic temperature system.

Sampling

Thermal conductivity can rarely be practically measured over the full interval of interest. Certain sampling protocols must be devised and followed to estimate the conductivity over significant depth intervals. This may involve:

- Characterizing the conductivity of individual geological formations where these are largely homogenous
- Selecting samples from evenly spaced intervals as a "random" sampling strategy
- Measuring mixtures of chips from sequential depth intervals

Conductivity is temperature dependent and values measured in the laboratory should be corrected for in situ conditions especially for very deep wells, and in geothermal areas with high temperature gradients.

When no rock samples are available for measurement of thermal conductivity, estimates can be obtained from empirical relations between conductivity and well-log parameters or by choosing a "flux plate", a layer of more than a few meters thickness composed of a rock type whose conductivity has been characterized by measurements elsewhere. (See Blackwell and Steele, 1989 for a review of these methods.)

Summary

Heat flowing from the surface of the continents cannot be measured directly. It is calculated as the product of the temperature gradient (most often measured in boreholes) and the thermal conductivity of the formation within which the temperature gradient was measured. Thermal conductivity can be measured in situ, but by far the majority of measurements are made on rock samples from the borehole, or from nearby outcrops of the relevant formations.

Temperatures are usually measured by electronic transducers deployed on multi-conductor electrical cables or incorporated into "memory tools," which are lowered into the borehole on steel or alloy wires. Fiber-optic cables can also be inserted into boreholes to provide synoptic temperature profiles or a time series of profiles.

A variety of techniques, based on both steady-state and transient heat flow, can be used to characterize the thermal conductivity of rock specimens. Optical scanning methods provide rapid estimates of thermal conductivity and are valuable reconnaissance tools complementary to the conventional techniques.

Bibliography

Antriasian, A. M., 2010. The Portable Electronic Divided Bar (PEDB): a tool for measuring thermal conductivity of rock samples. In *Proceedings, World Geothermal Congress 2010*, Bali, Indonesia, 25–29 April 2010. 7 pp. See also http://www.hotdryrocks.com/content/view/130/73/

Beardsmore, G. R., and Cull, J. P., 2001. *Crustal Heat Flow: A Guide to Measurement and Modeling*. Cambridge: Cambridge University Press, 324 pp.

Beck, A. E., 1957. A steady-state method for the rapid measurement of the thermal conductivity of rocks. *Journal of Scientific Instruments*, **34**, 186–189.

Blackwell, D. D., and Steele, J. L., 1989. Thermal conductivity of sedimentary rocks: measurements and significance. In Naeser, N. D., and McCulloh, T. H. (eds.), *Thermal History of Sedimentary Basins*. New York: Springer, pp. 13–36.

Burkhardt, H., Honarmend, H., and Pribnow, D., 1990. *First Results of Thermal Conductivity Measurements with a Borehole Tool for Great Depths, KTB Report 90–6a*. New York: Springer, pp. 245–258.
Doig, R., Saull, V. A., and Butler, R. A., 1961. A new borehole thermometer. *Journal of Geophysical Research*, **66**, 4263–4264.
Förster, A., Schrötter, J., Merriam, D. F., and Blackwell, D. D., 1997. Application of optical-fiber temperature logging – an example in a sedimentary environment. *Geophysics*, **62**, 1107–1113.
Freifeld, B. M., Finsterle, S., Onstott, T. C., Toole, P., and Pratt, L. M., 2008. Ground surface temperature reconstructions: using *in situ* estimates for thermal conductivity acquired with a fiber-optic distributed thermal perturbation sensor. *Geophysical Research Letters*, **35**, L14309, doi:10.1029/2008GL034762.
Henninges, J., Zimmermann, G., Büttner, G., Schrötter, J., Erbas, K., and Huenges, E., 2005. Wireline distributed temperature measurements and permanent installations behind casing. In *Proceedings World Geothermal Congress*, Antalya, Turkey, 24–29 April 2005, 5 pp.
Horai, K. I., 1971. Thermal conductivity of rock-forming minerals. *Journal of Geophysical Research*, **76**, 1278–1308.
Ito, Y., Saito, T., and Nagumeo, M., 1977. Shotherm QTM measurement of Rock Specimens, Shotherm Sales Information #111: Abstracted and translated with permission from the original paper in Japanese:Shotherm QTM measurement of Rock Specimens, *Chinetsu*, 14, 21.
McGee, and Thomas, D., 1988. *Principles and Methods of Temperature Measurement*. New York: Wiley-Interscience, 608 pp.
Popov, Y. A., Pribnow, D. F. C., Sass, J. H., Williams, C. F., and Burkhardt, H., 1999. Characterization of rock thermal conductivity by high-resolution optical scanning. *Geothermics*, **28**, 253–276.
Pribnow, D. F. C., and Sass, J. H., 1995. Determination of thermal conductivity for deep boreholes. *Journal of Geophysical Research B*, **100**, 9981–9994.
Sass, J. H., Lachenbruch, A. H., and Munroe, R. J., 1971. Thermal conductivity of rocks from measurements on fragments and its application to heat-flow determinations. *Journal of Geophysical Research*, **76**, 3391–3401.
Sass, J. H., Kennelly, J. P., Jr., Wendt, W. E., Moses, T. H., Jr., and Ziagos, J. P. 1979. In Situ determination of heat flow in unconsolidated sediments, *U.S. Geological Survey Open-File Report,* 79–593, 73pp. Available in digital form as http://pubs.er.usgs.gov/usgspubs/ofr/ofr79593
Sass, J. H., Kennelly, J. P., Wendt, W. E., Moses, T. H., Jr., and Ziagos, J. P., 1981. *In situ* determination of heat flow in unconsolidated sediments. *Geophysics*, **46**, 76–83.
Sass, J. H., Kennelly, J. P., Jr., Smith, E. P., and Wendt, W. E., 1984. Laboratory line-source methods for the measurement of thermal conductivity of rocks near room temperature. *U.S. Geological Survey Open-File Report*, 84–91, 20pp. Available in digital form as http://pubs.er.usgs.gov/usgspubs/ofr/ofr8491
Selker, J. S., Thévenaz, L., Huwald, H., Mallet, A., Luxemburg, W., van de Giesen, N., Stejskal, M., Zeman, J., Westhoff, M., and Parlange, M. B., 2006. Distributed fiber-optic temperature sensing for hydrologic systems. *Water Resources Research*, **42**, W12202, doi:doi:10.1029/2006WR005326.
Suto, Y., Sakuma, S., Takahashi, H., Hatakeyama, N., and Henfling, J., 2008. Temperature memory gauge survey and estimation of formation temperature of the USDP-4 conduit hole at Unzen Volcano, Japan. *Journal of Volcanology and Geothermal Research*, **175**(1–2), 20–27.
Tew, W. L., and Strouse, G. F., 2001. Standard Reference Material 1750: Standard Platinum Resistance Thermometers, 13.8033 K to 429.7485 K, NIST Special Publication 260–139, U.S.Government Printing Office, Washington, D.C., 37 pp. Specific information on RTD devices is best obtained by entering "resistance temperature devices" into an Internet Search Engine. An excellent contemporary source is http://www.omega.com/temperature/Z/TheRTD.html).

Cross-references

Heat Flow, Continental
Heat Flow, Seafloor: Methods and Observations
Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat
Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity

HEAT FLOW, CONTINENTAL

Paul Morgan
Colorado Geological Survey, Denver, CO, USA

Synonyms

Heat flow; Heat flow density; Heat flux, land; Terrestrial heat flow

Definition

Heat flow. The outward flow of thermal energy by conduction from within the Earth through the solid surface of the Earth.
Continental heat flow. Heat flow from the continental crust or lithosphere, which are those portions of the plates that are not directly formed by seafloor spreading at mid-ocean ridges and are not generally subducted.

Introduction

Heat flow is a fundamental property of the Earth. The Earth is a heat engine that drives plate tectonics and intraplate processes such as hot-spot volcanism unrelated to plate margins. Heat flow, the outward flow of thermal energy from within the Earth through the solid surface of the Earth, is a measure of the thermal energy that drives this heat engine. In addition, temperature in the Earth, which is closely related to heat flow, is a primary parameter controlling physical properties of materials within the Earth. In general, areas of high heat flow are underlain by higher temperatures, at least at shallow depth, than areas of low heat flow. We seek therefore to know the distribution of heat flow from the surface of the Earth and to understand the factors that control its variations.

The primary factor controlling the distribution of heat flow is the type of lithosphere, continental or oceanic. Continental lithosphere is generally relatively old (>500 Ma) and the dominant mechanism of heat transfer through this lithosphere is conduction. A significant portion of heat flow from this lithosphere is generated by radiogenic heat production in the continental crust. In contrast, oceanic lithosphere is young (<200 Ma), heat is convected into this lithosphere as it is formed at mid-ocean ridges and it has very little internal radiogenic heat production (see *Heat Flow, Seafloor: Methods and Observations*). A variety of

geologic processes modify continental lithosphere and most of these processes also modify the thermal regime. In addition, the upper continental crust commonly hosts moving groundwater that redistributes heat in the upper crust. Thus, the long and complex history of continental lithosphere, significant and lateral variability in radiogenic heat generation, and redistribution of heat by groundwater combine to make continental heat flow heterogeneous.

Fourier's law

Under conditions of steady state, heat flow, q, is a function of thermal conductivity, K, and temperature, T, given by Fourier's law:

$$q = -K\nabla T, \tag{1}$$

where ∇ is the gradient operator. This equation assumes that thermal conductivity is isotropic, that is, it does not change with direction. In general, spatial changes in thermal conductivity and anisotropy must be considered. The negative sign in the equation recognizes that heat flows from higher temperatures to lower temperatures and that flow of heat is in the opposite direction to the temperature gradient. On average, Earth heat flow is primarily vertical and the vertical component of heat flow, q_z, is commonly quoted:

$$q_Z = -K\frac{\partial T}{\partial z}, \tag{2}$$

where z is depth. In practice, the subscript and negative sign are commonly omitted:

$$q = K\,\partial T/\partial z. \tag{2a}$$

The working units of q, K, and $\partial T/\partial z$ are $\mathrm{mW\,m^{-2}}$, $\mathrm{W\,m^{-1}\,^{\circ}C^{-1}}$ or $\mathrm{W\,m^{-1}\,K^{-1}}$, and $\mathrm{^{\circ}C\,km^{-1}}$, respectively.

Under transient conditions the heat conduction equation in one dimension becomes:

$$\frac{\partial T}{\partial t} = \kappa\frac{\partial^2 T}{\partial z^2}, \tag{3}$$

where t is time and κ is thermal diffusivity, given by $\kappa = K/\rho c$, where ρ is density and c is specific heat. The general working units for t and κ are s and $\mathrm{mm^2\,s^{-1}}$, although Ma and $\mathrm{km^2\,Ma^{-1}}$ are also sometimes used.

If heat generation, A, is also included, the heat flow equation becomes:

$$\frac{\partial^z T}{\partial z^z} - \frac{1}{\kappa}\frac{\partial T}{\partial t} + \frac{A}{K} = 0. \tag{4}$$

The working units for radiogenic heat generation, A, are $\mu\mathrm{W\,m^{-3}}$.

Geothermal gradient

The rate of increase in temperature with depth, $\partial T/\partial z$, is the *geothermal gradient*. This term is usually reserved for zones over which heat is transferred by conduction and the rate of increase in temperature with depth is linear (see *Heat Flow Measurements, Continental*), but if detailed temperature data are not available or gradients are estimated from oil-well bottom-hole temperatures, an approximate average geothermal gradient may be given. The profile of temperature plotted as a function of depth is the *geotherm*.

Geothermal gradients range from very low to near zero in fore-arc regions, the zones between subduction zones and their associated volcanic belts where heat is convected downward by the subducting lithosphere. Similarly, low gradients are found in zones of groundwater recharge. Gradients in excess of $20{,}000^{\circ}\mathrm{C\,km^{-1}}$ have been measured at one location in the upper few meters of the sediments beneath Yellowstone Lake inside the caldera of Yellowstone National Park, Wyoming, USA (Morgan, et al., 1977), but such high gradients are buffered by the boiling point curve of water as soon as temperatures reach the boiling temperature at depth. Thus, there is a very wide range of geothermal gradients in the upper continental crust. Temperatures are ultimately controlled by the boiling point curve of water in the uppermost crust and then the solidus curves of the crust and mantle lithosphere, and temperatures at depth generally do not range as widely as shallow geothermal gradients might suggest. Thermal conductivity varies by a factor of about 5 in the main rock types that comprise the continental crust (see *Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat*): there is some variation in heat flow associated with changes in thermal conductivity, but the most significant variations in heat flow are accompanied by changes in geothermal gradient.

Sources of heat

There are two sources of heat contributing to continental heat flow, heat that enters continental lithosphere from below, and heat generated within the continental lithosphere (see *Energy Budget of the Earth*). Although the Earth beneath the lithosphere has thermal energy that must ultimately be lost to the surface either through the lithosphere or in the process of the formation of new oceanic lithosphere at mid-oceanic ridges, there is no requirement where or when that heat must be lost, that is, there is no constant basal heat flux that flows into continental lithosphere. The thermal boundary condition at the base of the lithosphere is probably more accurately described as a constant temperature boundary, and this constant temperature may be maintained either by small-scale convection and/or lateral flow in the asthenosphere. Heat flow into the base of the lithosphere is then determined by the thermal conductivity structure of the lithosphere, the thickness of the lithosphere, heat production within the lithosphere, and its surface temperature. The resulting thermal profile is disrupted during major deformation of the lithosphere or changes in its boundary conditions.

By far the largest source of heat within the lithosphere is radiogenic heat production from the decay series of

the unstable isotopes ^{232}Th, ^{235,238}U, and ^{40}K. These isotopes all have half-lives of the same order as the age of the Earth and thus remain sufficiently abundant to be significant and continue to decay at a rate that is significant in terms of heat production (see *Radiogenic Heat Production of Rocks*). Other radiogenic isotopes are important for age dating, such as ^{87}Rb, ^{147}Sm, and ^{187}Re (see *Absolute Age Determinations: Radiometric*), but the heat produced by the decay of these isotopes and other isotopes used for age dating is insignificant. Uranium, thorium, and potassium are incompatible elements and tend to be concentrated in silicic magmas. Hence, they are depleted in the mantle during crustal formation and generally concentrated in the crust in silicic rocks. Heat flow data show that their concentrations in large silicic plutons are relatively predictable from surface measurements (Roy et al., 1968), but their distributions are less predictable in metamorphic and sedimentary terrains (see *Lithosphere, Continental: Thermal Structure*). Measurements indicate that from as little as less than 20% to as much as 80% of heat flow may be derived from crustal radiogenic heat generation. Typical values of heat production for crustal rocks range from $<0.1\ \mu W\,m^{-3}$ for basic igneous rocks to $>6\ \mu W\,m^{-3}$ for radiogenic granites.

Other crustal heat sources include frictional heat from faults and heat from chemical reactions. Attempts to measure heat associated with active faults have generally failed to detect heat or produced inconclusive results. The primary chemical reaction that has been observed as a heat source is the oxidation of pyrite, but any exothermic metamorphic or weathering reaction is a potential heat source. However, friction and chemical heat sources are likely to be only local in extent and minor in terms of the total continental thermal budget.

Mechanisms of heat transport

In stable continental crust, the primary mechanism of heat transfer is conduction as described by Fourier's law (Equations 1 and 2). In the mantle, the effects of radiative heat transfer may be significant as olivine has a significant component of radiative heat transfer above a temperature of about 500°C. This radiative component may be included in Fourier's law by substituting a temperature-dependent value of thermal conductivity that includes a correction for radiative heat transfer.

Even in stable regions, heat may be redistributed in the upper few kilometers by groundwater flow. In stable regions the geothermal gradient is seldom high enough to drive free convection (see below), and groundwater flow is driven by differences in elevation of the piezometric surface (water table). Heat transferred by the forced vertical flow of groundwater, q_{gw}, may be approximated by:

$$q_{gw} = \bar{v} c_w |dT|, \tag{5}$$

where $\bar{v}$ is the equivalent vertical velocity if the porosity were 100%, c_w is the specific heat of water, and $|dT|$ is the absolute temperature drop over the vertical extent of the water flow. Heat transfer will be in the direction of water flow and heat flow is decreased in areas of recharge and increased in areas of discharge. This type of convection is termed forced convection or advection.

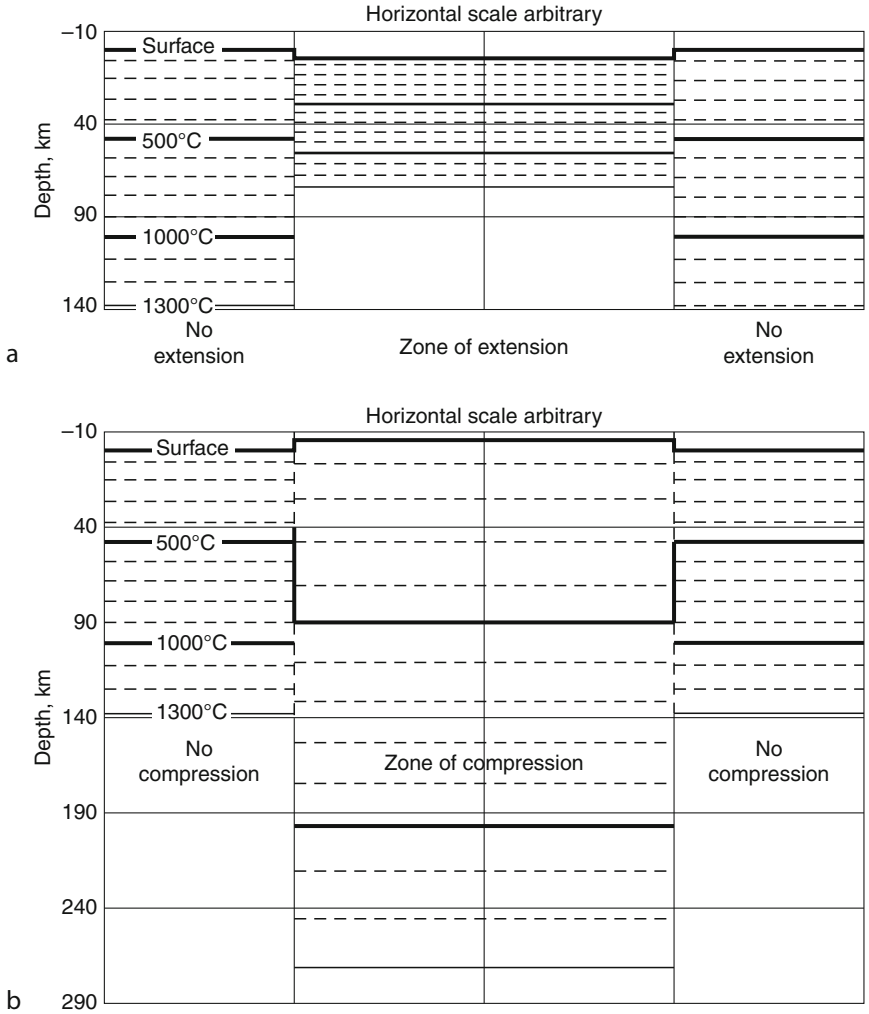
If the geothermal gradient is sufficiently high, water flow may be driven by thermal buoyancy forces. The magnitude of the gradient required for the onset for thermal convection depends on the geometry of the heat source and the medium in which convection occurs (e.g., Turcotte and Schubert, 2002, p. 393 et seq.). For a homogeneous porous layer of permeability $k\,m^2$ and thickness b km heated uniformly from below, the minimum thermal gradient, $(\partial T/\partial z)_{min}$, required for convection to occur is given by:

$$\left(\frac{\partial T}{\partial z}\right)_{min} = \frac{4.2x10^{-10}}{kb^z}. \tag{6}$$

This minimum gradient may not sustain convection and a higher gradient may be required for continuous convection. Convection driven by thermal buoyancy forces is termed free convection.

Heat will also be convected when there are net vertical movements within the lithosphere associated with deformation or changes in either the top or bottom boundaries of the lithosphere. In general, extensional deformation causes thinning of the lithosphere resulting in a net upward movement of the material in the lithosphere relative to its surface and a compression of the isotherms, or horizons of equal temperature within the lithosphere (Figure 1a). In contrast, compressional deformation causes thickening of the lithosphere resulting in a net downward movement of the material in the lithosphere relative to its surface and an increase of the spacing of the isotherms within the lithosphere (-Figure 1b). For simplicity, deformation in Figure 1 is shown to be 100% by pure shear and uniformly distributed throughout the lithosphere, a highly unlikely distribution of geological strain. As shear strength changes from brittle to ductile laterally and with depth throughout the lithosphere, the strain distribution is always much more complex: zones of simple shear, imbricated layers, and other complexities typically characterize lithospheric deformation. However, at a lithospheric scale, the net material movements and compression and extension of the geotherms illustrated in Figure 1 are valid.

Stable continental lithosphere has a relatively low heat flow corresponding to a low geothermal gradient and strong lithosphere. These conditions are not favorable for deformation. Observations indicate that deformation commonly occurs in lithosphere that is still recovering from a previous thermal event or has experienced a preceding thermal event. This thermal event may be manifested in the form of volcanism in which heat is advected in the lithosphere by ascending magmas. Heat transfer by magmas is conceptually similar to heat transfer by groundwater described by Equation 5, if the specific heat of magma is substituted for the specific heat of water. Additional heat is released from a magma during



Heat Flow, Continental, Figure 1 Diagrammatic examples of the effects of instantaneous pure shear deformation on lithospheric isotherms. Both diagrams have the same thermal conditions in undeformed lithosphere (no extension and no compression) calculated for a 35 km thick crust with a surface heat flow of 40 mW m^{-2}, a surface heat production of 2 μW m^{-3}, and a crustal thermal conductivity of 2.5 W m^{-1} K^{-1}. The crustal heat production is assumed to decay exponentially with depth with a depth parameter of 10 km, and the mantle is assumed to have a heat production of 0.084 μW m^{-3} and a temperature-dependent thermal conductivity appropriate for olivine. The base of the lithosphere is assumed to be the 1,300 °C isotherm. (**a**) shows the effect of thinning the lithosphere by a factor of 2 by extending the lithosphere by a factor of 2. (**b**) shows the effect of doubling the thickness of the lithosphere by compressing the lithosphere by a factor of 2.

solidification, and this heat is represented by the latent heat of solidification. Many analytical solutions for conductive modeling of magma intrusions may be found in Carslaw and Jaeger (1959) and other references, and these processes may also be modeled numerically. Magma generation may be initiated from below the lithosphere by the ascent of hot material to the base of the lithosphere (mantle plumes?) generating magmas by pressure-release melting and raising the temperature at the lower boundary of the lithosphere. Additionally magmas may be generated within the lithosphere by pressure-release melting and the melting of lower-melting temperature components within the lithosphere.

Once deformation is initiated additional magma generation will be initiated. The net upward movement of material during extension may trigger pressure-release melting, especially if the lithosphere has been preheated by pre-deformation magmatism. During compression, there is a net syn-deformational lowering of the geotherm. However, thickening of the crustal layer of radiogenic heat-producing elements will generally cause the geothermal gradient to return to a higher value than its initial value. This higher geotherm may result in melting of the thickened crust and crustal derived magmatism, the results of which are commonly observed in belts of compressional deformation. Thus, magmatic

convection of heat is an important heat transfer mechanism that commonly accompanies heat convection by tectonic deformation.

Finally, there is heat convection associated with changes at the Earth's surface. In discussing deformation and magmatism, the effects of these processes on surface elevation were ignored. Cooling oceanic lithosphere as it ages away from oceanic ridges results in a relatively simple change in depth of the ocean floor (see *Heat Flow, Seafloor: Methods and Observations*). A similar, but more complex relation exists for the thermal and density structure of continental lithosphere (Lachenbruch and Morgan, 1990; *Isostasy, Thermal*). Changes in surface elevation are generally accompanied by either erosion or sedimentation with the net result that the material in the lithosphere has a net movement either up or down relative to the surface, respectively. If the rates of erosion or sedimentation are known, a correction may be applied for these effects. If erosion creates three-dimensional topography, a terrain correction may also be applied. Not strictly a heat transfer mechanism, but changes in surface temperature associated with past climate change or changes in land use may also change surface heat flow and a correction may be applied for these perturbations.

The duration of a thermal disturbance associated with thermal convection in the lithosphere depends primarily on the stable thickness of the lithosphere and its thermal diffusivity. There is no standard for defining this duration as temperatures relax asymptotically to their stable values, and the form of this return depends on the structure of the thermal disturbance, which may be highly variable and very complex in continental deformation with magmatism. However, a useful guideline for the duration of the thermal disturbance, τ, is:

$$\tau = \frac{L^z}{4\kappa} \tag{7}$$

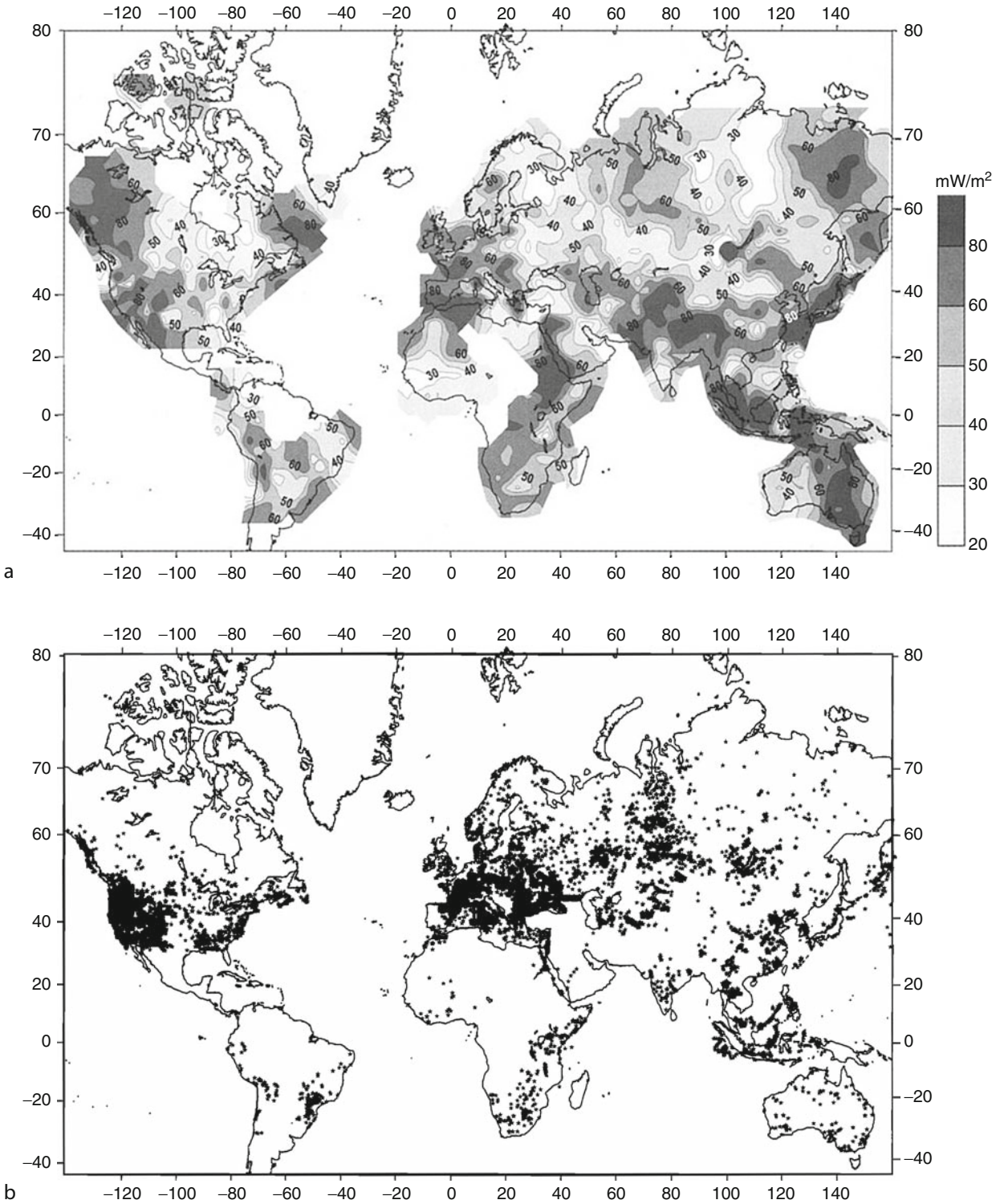
where L is the stable thickness of the lithosphere and κ is its thermal diffusivity. For example, using an average thermal diffusivity of 32 $km^2\ Ma^{-1}$ and a thickness of 100 km for stable oceanic lithosphere, the calculated duration of the thermal disturbance is about 78 Ma. Oceanic crust is observed to flatten at about 70–80 Ma. For continental lithosphere, if the stable thickness of this lithosphere is 100 km, the duration of the thermal disturbance would be about 78 Ma; if the stable thickness were 150 km, the duration would be ~175 Ma; if the thickness were 200 km, the duration would be ~310 Ma; and if the stable thickness were 250 km, the duration would be ~500 Ma. These durations are from the start of stabilization: if a deformation and magmatism event were followed by significant erosion or sedimentation, the duration of the thermal stabilization would not begin until the end of the erosion or sedimentation. However, as the effects of sedimentation and erosion are relatively minor, their measurable effects would probably be dissipated before the end of the formal disturbance duration.

Global distribution of continental heat flow

The global distribution of continental heat flow is shown in Figure 2a. The density of data is far from uniform (Figure 2b), but there are sufficient data that some generalizations about the heat flow distribution may be made. In general, high heat flow corresponds with zones of active seismicity, which in turn correlate with plate margins. Where the zones of seismicity are diffuse, such as the western US and the Himalaya, the plate margin deformation is distributed and the zones of high heat flow are extensive. Some portions of active plate margins are associated with low heat flow, such as the western margin of South America and the Pacific Northwest of the USA, west of the volcanic Cascade Range. These areas overlie subducting oceanic lithosphere where heat is being convected down into the mantle. Associated with these low heat flow zones is a rapid lateral transition to high heat flow in the active volcanic zones of the subduction zones. In addition, plate margins are generally associated with elevated topography, which received more precipitation than low elevations and in which heat flow is generally more variable because of redistribution of heat by forced and free groundwater convection.

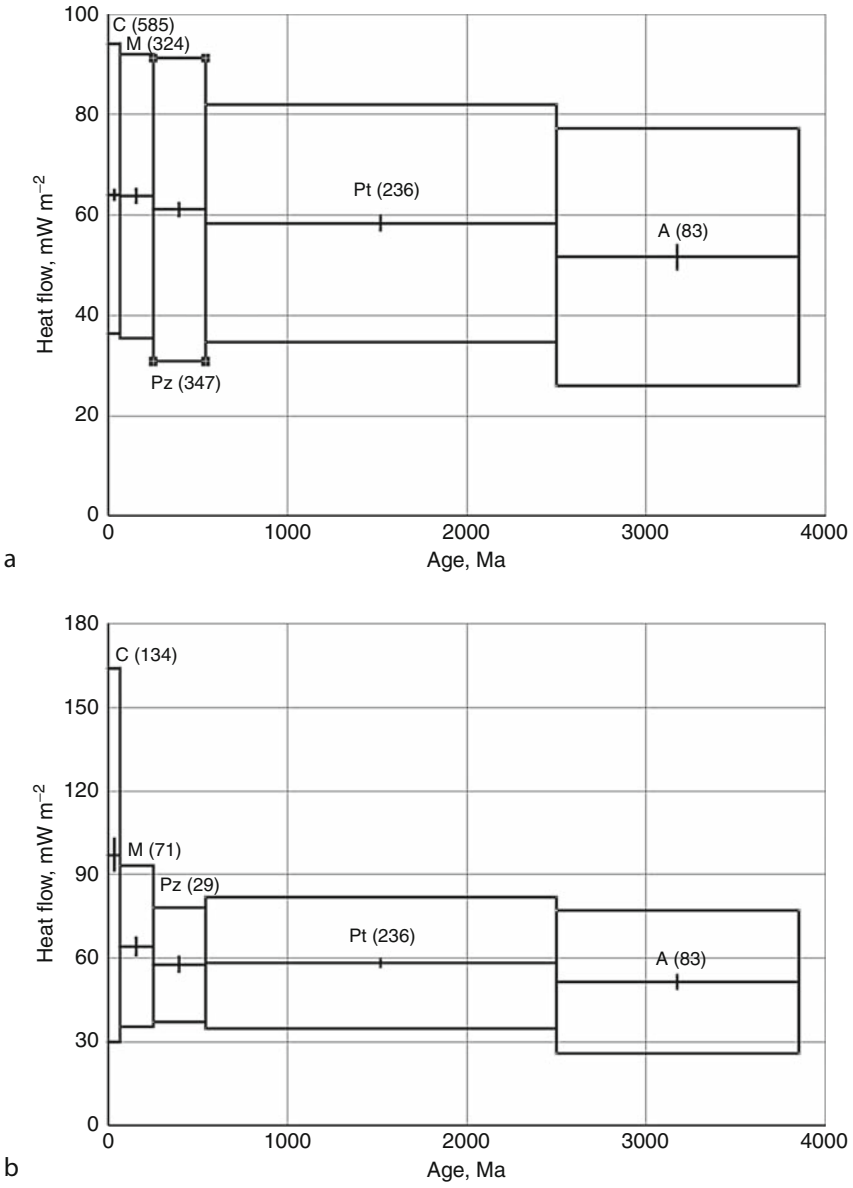
The association of high and low heat flow with active plate margins is demonstrated in a plot of mean heat flow as a function of age of heat flow site (Figure 3). Data are shown for sedimentary and metamorphic sites in Figure 3a and igneous sites in Figure 3b. The data for sedimentary and metamorphic sites (Figure 3a) show a small decrease from Cenozoic and Mesozoic to Paleozoic sites and from Paleozoic to Proterozoic sites, and a larger decrease from Proterozoic to Archean sites. This result suggests that the plate margin disturbances are relatively minor in the Phanerozoic sedimentary and metamorphic site data. In the igneous site data (Figure 3b), however, there is a significant decrease in both the mean and standard deviation from both Cenozoic to Mesozoic and from Mesozoic to Paleozoic sites. There is almost no difference in the mean from Paleozoic to Proterozoic sites. The data are the same for Proterozoic and Archean sites as Figure 3a. The igneous site data clearly show thermal perturbations with ages extending into Mesozoic age sites relative to older sites. This result is consistent with a continental lithospheric thickness of the order of 200 km. The observation that the igneous site data show a perturbation at young ages but there is only a small signal in the sedimentary and metamorphic site data suggests that the sedimentary and metamorphic site data are dominated by sedimentary sites and that these sites are either distant from plate boundaries or high heat flow is suppressed at these sites by downward advection of heat by advection during sedimentation.

The difference between the mean heat flow at Proterozoic and Archean sites has be commented upon from previous compilations of global heat flow data (Morgan, 1985) in addition to this data set (Artemieva and Mooney, 2001). To explain this difference as a transient effect would require a lithosphere of the order of 500–600 km



Heat Flow, Continental, Figure 2 (**a**) Distribution of continental heat flow, based on the data of Pollack et al. (1993) and subsequent updates, presented with a 10° × 10° interpolation (Artemieva and Mooney, 2001, Figure 3). (**b**). Locations of continental heat flow measurements used in constructing Figure 2(**a**) (From Artemieva and Mooney, 2001, Figure 4).

in thickness, which is probably unrealistic. However, a mean heat flow difference of 7 $mW\,m^{-2}$ may be explained in terms of a difference in the contributions of the mean heat production of Proterozoic and Archean crusts to their heat flows. Assuming an average crustal thickness of 35 km, an average difference of 0.2 $\mu W\,m^{-3}$ would explain the difference between the mean measured heat flow at Archean and Proterozoic sites. This difference is not universally observed in all areas (Jaupart and Mareschal, 1999; *Lithosphere, Continental: Thermal Structure*), but it is consistent with bulk observations of the differences in chemistry between the upper crust of Proterozoic and Archean terrains (McLennan et al., 2006). Archean upper crust is statistically depleted in the heat-producing elements relative to younger crust. In addition, xenolith data indicate that Archean lithosphere is statistically distinct from younger lithosphere so the upper crustal thermal observation may indicate that Archean lithosphere is statistically thermally distinct (Griffin et al., 2009).



Heat Flow, Continental, Figure 3 Continental heat flow as a function of age of data site. Widths of boxes and bars indicate age range. Heights of boxes indicate ± one standard deviation about the mean; short vertical line indicates ± one standard error about the mean. Ages of sites are grouped as C – Cenozoic, M – Mesozoic, Pz – Paleozoic, Pt – Proterozoic, and A – Archean. Numbers in parentheses are the number of 1° by 1° blocks represented in each age group. (**a**) shows sedimentary and metamorphic sites for Phanerozoic data. (**b**) shows igneous sites for Phanerozoic data. Data re-plotted from Pollack et al. (1993).

In summary, the global distribution of continental heat flow is complex. Neglecting measurement errors (see *Heat Flow Measurements, Continental*), variations associated with surface temperature changes, uncorrected topographic effects, and anomalies induced by changes in surface cover, heat flow may be described as steady state or transient. At any site, near-surface heat flow may be redistributed by groundwater flow, but this is most likely at sites with high elevation and rugged topography that are likely to have experienced relatively young (Cenozoic or Mesozoic) mountain building and/or igneous activity. Transient thermal perturbations in the lithosphere associated with tectonic deformation and/or igneous activity will dissipate with time and are likely to be restricted to heat flow sites in lithosphere with Cenozoic or Mesozoic tectonic or thermal ages. To distinguish tectonic and igneous ages from sedimentary ages, the term *tectonothermal age* is commonly used to describe the age of heat flow sites. Data from sites with tectonothermal ages Paleozoic and older are generally expected to be in steady state unless erosion or sedimentation are significant.

Igneous activity always heats continental lithosphere but deformation may either raise or lower the geothermal gradient (Figure 1). There is therefore no predictable anomaly for continental lithosphere at or near plate margins. The wide range in observed heat flow measured at young sites is in part associated with redistribution of heat by groundwater flow, but also associated with the real range in anomalies, both positive and negative, generated by deformation.

Whether the heat flow and geotherm be transient or steady state, significant variation among sites may occur associated with crustal radiogenic heat generation. Heat generation in the upper crust is very variable being generally low in mafic rocks and high in silicic rocks. Its contribution may vary from $<20\%$ to $>70\%$ of heat flow, and much of the range in heat flow represented in the standard deviations about the means of heat flow in Figure 3 are probably associated with chemical heterogeneity in the continental crust resulting in lateral variations in the contributions of radiogenic heat production to heat flow.

Paleo-heat flow

The discussion above has been restricted to heat flow measured at present although there is some lag in heat flow measured at the surface as it is conducted from depth: the order of that time lag may be estimated from Equation 7. This heat flow and associated temperatures are appropriate for most use for geophysical parameters, such as seismic wave velocities, electrical conductivities, and density calculations, but the time lag may be significant in some area where temperatures at depth in the lithosphere are changing. However, for some geological observations, including most techniques of radiometric age dating, the former temperature of a rock or mineral are preserved in a chemical or radiometric system when the rock cools as it is brought to the surface. This process may allow the temperature and sometimes the geothermal gradient of a site to be reconstructed at the time the rock was uplifted and this general topic is sometimes referred to as paleo-heat flow.

The rates of most chemical reactions increase at higher temperatures. For example, the maturation of organic material in sediments roughly follows the Arrhenius' equation, which describes the velocity of the reaction, r, as:

$$r = \lambda \exp(-E/RT_K) \tag{8}$$

where λ is the frequency factor, E is the activation energy, R is the gas constant, and T_K is temperature in Kelvin (Stegena, 1988). The combination of the frequency factor and temperature gives a time-temperature index that, together with composition of the organic matter, provides a useful predictor of hydrocarbon maturation. Temperatures in sedimentary basins are usually estimated by the increase in coal rank of carbonaceous matter in the sediments as measured by its vitrinite reflectance (e.g., Suggate, 1998), and time may be estimated by the sedimentation history of the basin. Vitrinite reflectance measures the maximum temperatures attained by the sediments, which may exceed modern temperatures.

Radiometric methods for age dating (see *Absolute Age Determinations: Radiometric*) do not determine rock ages but the time at which the radiometric system on which the age determination is based cools through its closure temperature, the temperature at which the age is set. Thus, radiometric ages may be used for the additional purpose of determining the thermal history of a rock, especially if some estimate of the depth at which the age was set may be determined (see below). Radiometric systems do not close at a single temperature but follow the Arrhenius' equation; at high temperatures cooling is generally relatively rapid. Examples of approximate mineral-closure temperatures for the uranium-lead method are monazite $>1{,}000°C$, zircon $>1{,}000°C$, titanite 650–600°C, apatite 500–450°C, and rutile 450–400°C (Flowers, 2005). Examples of approximate mineral-closure temperatures for the potassium-argon system are hornblende $530 \pm 40°C$, muscovite $\sim 350°C$, and biotite $280 \pm 40°C$ (McDougall and Harrison, 1999). Different dates from these minerals may be used to calculate rates of erosion or exhumation of terrains. However, young ages may also indicate reheating for minerals with relatively low-temperature closure temperatures.

The minerals used for some radiogenic systems have geologically very low closure temperatures. For example, the fission track technique is based on radiation damage (tracks) in different minerals caused by the spontaneous fission of ^{238}U (Garver, 2008). Closure of this system is by thermal annealing of the fission tracks, which occurs at $>120°C$ for apatite, $\sim 200°C$ for zircon, and $\sim 300°C$ for sphene. At these low temperatures, a geothermal gradient is usually assumed and the fission track ages are commonly used to indicate recent uplift/erosion rates (uplift/erosion rate = closure temperature/[fission track age × assumed geothermal gradient]).

Finally, paleo-thermal conditions are preserved by metamorphism, and in particular in mineral systems called geothermometers. Geothermometers are typically pairs of minerals in which concentration of one element in one of the minerals has a strong temperature dependence but is essentially pressure independent; the other mineral, which must be in intimate contact with the first mineral grain, acts as a reservoir for this element and allows the first mineral to be in thermal equilibrium with the element. Geothermometers tend to equilibrate rapidly at high temperatures but re-equilibrate slowly as the temperature decreases, so if cooling and exposure at the surface is relatively rapid, the geothermometer records peak thermal conditions. Geobarometers also exist in which the concentration of an element in the first mineral is strongly pressure (depth) dependent but insensitive to temperature. If mineral pairs representing a geothermometer and a geobarometer occur in the same outcrop, then temperature and depth may be estimated allowing calculation of the paleo-geothermal gradient and heat flow at the time of peak thermal conditions.

Studies of paleo-thermal gradients from metamorphic terrains suggest that, on average, geothermal gradients were higher than modern average geothermal gradients. However, although global heat flow was almost certainly higher in the past than at present (see *Energy Budget of the Earth*), these observations do not demonstrate that conclusion for the following reasons: (1) only peak or near-peak thermal conditions are preserved resulting in an overestimation of average continental heat flow; (2) peak preserved geothermal gradients are no higher than geothermal gradients measured today – modern geothermal gradients are buffered in the uppermost crust by the water boiling point curve and at greater depths by the crust solidus, and the same conditions would have applied in the earliest Earth preserved in the continental crust; and (3) if there were zones of low temperature/high pressure metamorphism (blueschist) in the Archean/Early Proterozoic, they are unlikely to be preserved because a geotherm required to maintain these metamorphic conditions requires active subduction. With the cessation of subduction, the geotherm would relax to a warmer geotherm, and blueschist metamorphism is likely to be overprinted. Similarly, low temperatures would not be recorded by geothermometers as the lithospheric temperatures increase.

Summary

Continental heat flow is a complex, but fundamental parameter that, together with oceanic heat flow, constrains the energy budget of the Earth heat engine. It is particularly complex because continents are complex and heterogeneous. It has components from beneath the lithosphere and radiogenic heat production within the lithosphere. Heat is advected into and within the lithosphere by magmatic activity, and heat is redistributed by lithospheric deformation associated with plate tectonics and at relatively shallow depths by groundwater flow. Changes in surface conditions also modify heat flow in the shallow crust. The time constant of thermal relaxation of continental lithosphere is of the order of 300 Ma for stable lithosphere 200 km in thickness. Typically heat flow data from igneous sites of Cenozoic and Early Mesozoic ages reflect transient thermal conditions; data from sites of older ages do not. However, continental heat flow remains very variable even at Precambrian sites because of lateral variability in crustal radiogenic heat production.

Bibliography

Artemieva, I. M., and Mooney, W. D., 2001. Thermal thickness and evolution of Precambrian lithosphere. *Journal of Geophysical Research*, **106**, 16387–16414.

Carslaw, H. S., and Jaeger, J. C., 1959. *Heat Conduction in Solids*. Oxford: Oxford University Press, 510. pp.

Flowers, R. M., 2005. Tempo of burial and exhumation within the deep roots of a magmatic arc, Fiordland, New Zealand. *Geology*, **33**, 17–20.

Garver, J. I., 2008. Fission-track dating. In Gornitz, V. (ed.), *Encyclopedia of Paleoclimatology and Ancient Environments*. Dordrecht: Springer, pp. 247–249.

Griffin, W. L., O'Reilly, S. Y., Afonso, J. C., and Begg, G. C., 2009. The composition and evolution of lithospheric mantle: a re-evaluation and its tectonic implications. *Journal of Petrology*, **50**, 1185–1204.

Jaupart, C., and Mareschal, J.-C., 1999. The thermal structure and thickness of continental roots. *Lithos*, **48**, 93–114.

Lachenbruch, A. H., and Morgan, P., 1990. Continental extension, magmatism and elevation; formal relations and rules of thumb. *Tectonophysics*, **174**, 39–62.

McDougall, I., and Harrison, T. M., 1999. *Geochronology and Thermochronology by the $^{40}Ar/^{39}Ar$ Method*. New York: Oxford University Press.

McLennan, S. M., Taylor, S. R., and Hemming, S. R., 2006. Composition, differentiation, and evolution of continental crust: constraints from sedimentary rocks and heat flow. In Brown, M., and Rusher, T. (eds.), *Evolution and Differentiation of Continental Crust*. New York: Cambridge University Press, pp. 92–134.

Morgan, P., 1985. Crustal radiogenic heat production and the selective survival of ancient continental crust. *Journal of Geophysical Research*, **90**, C561–C570. Supplement.

Morgan, P., Blackwell, D. D., Spafford, R. E., and Smith, R. B., 1977. Heat flow measurements in Yellowstone Lake and the thermal structure of the Yellowstone Caldera. *Journal of Geophysical Research*, **82**, 3719–3732.

Pollack, H. N., Hurter, S. J., and Johnson, J. R., 1993. Heat flow from the Earth's interior: analysis of the global data set. *Reviews of Geophysics*, **31**, 267–280.

Roy, R. F., Blackwell, D. D., and Birch, F., 1968. Heat generation of plutonic rocks and continental heat flow provinces. *Earth and Planetary Science Letters*, **5**, 1–12.

Stegena, L., 1988. Paleogeothermics. In Haenel, R., Rybach, L., and Stegena, L. (eds.), *Handbook of Terrestrial Heat-Flow Density Determination*. Dordrecht: Kluwer, pp. 391–419.

Suggate, R. P., 1998. Relations between depth of burial, vitrinite reflectance and geothermal gradient. *Journal of Petroleum Geology*, **21**, 5–32.

Turcotte, D. L., and Schubert, G., 2002. *Geodynamics*. Cambridge: Cambridge University Press.

Cross-references

Absolute Age Determinations: Radiometric
Energy Budget of the Earth

Heat Flow, Continental
Isostasy
Lithosphere, Continental: Thermal Structure
Lithosphere, Oceanic: Thermal Structure
Radiogenic Heat Production of Rocks
Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat

HEAT FLOW, SEAFLOOR: METHODS AND OBSERVATIONS

Earl E. Davis[1], Andrew T. Fisher[2]
[1]Pacific Geoscience Centre, Geological Survey of Canada, Sidney, BC, Canada
[2]Earth and Planetary Sciences Department, University of California at Santa Cruz, Santa Cruz, CA, USA

Definition

Heat flow. The rate of thermal energy transfer in a medium driven either conductively along a thermal gradient or advectively via mass transport. The standard unit is watts, W. The term is also used to describe a subdiscipline of geophysics, as in the title of this entry.

Conductive heat flux. The heat flow per unit area diffusing by conduction along a thermal gradient, determined as the product of the thermal gradient and thermal conductivity. The standard unit is W m^{-2}. The term *heat flow density* has been used correctly as a synonym; the term *heat flow*, traditionally but inexactly used as a synonym for heat flux, more strictly applies to the integrated heat flux over a specified area or region (watts).

Advective heat flux. The rate of heat transfer per unit area carried by a moving medium, proportional to the velocity and the heat capacity of the medium. The standard unit is W m^{-2}.

Thermal conductivity. The quantity that defines the ability of a medium to transfer heat by steady-state diffusion. The standard unit is W m^{-1} K^{-1}.

Hydrothermal circulation. Large-scale pore-fluid convection driven by thermal buoyancy, in which fluid flux and advective heat flux are strongly influenced by the permeability structure of the host formation.

History of observations

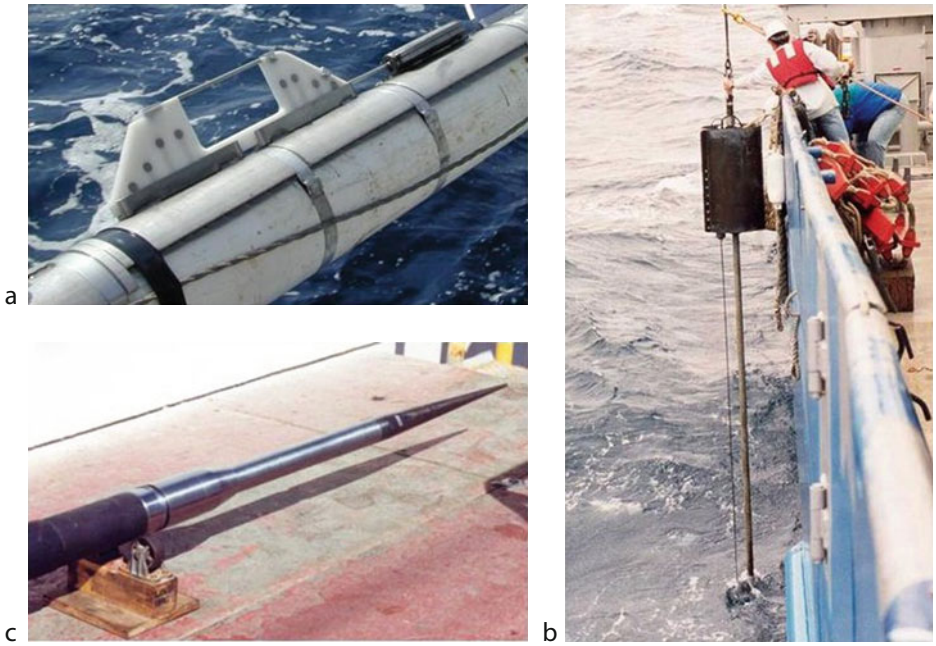
Pioneering measurements of temperature below the seafloor (e.g., Petterson, 1949; Revelle and Maxwell, 1952; Bullard, 1954) were made to compare the thermal state of the ocean crust with that of continents and thus to improve the knowledge of the present-day heat loss from the Earth. The early data demonstrated that gravity-driven probes and corers that penetrated a few meters into seafloor sediments could provide meaningful geothermal gradients, and they provided a foundation for the marine heat flow discipline. Methods for measuring the seafloor thermal gradient and thermal conductivity (the product of which is conductive heat flux) improved in subsequent years, the number and geographic distribution of determinations increased, and patterns of seafloor heat flux were gradually revealed. Initially, the average seafloor heat flux appeared to be similar to that through continents, despite the contribution from crustal radiogenic heat production, which is significant in the continental crust but not in the oceanic crust. Heat-flux values over midocean ridges were found to be significantly higher on average than in the flanking basins, but locally, values were often inexplicably scattered (Von Herzen and Uyeda, 1963; Lee and Uyeda, 1965). Higher heat flux at midocean ridges was consistent with emerging ideas about seafloor spreading, although the values measured were lower than expected from early theoretical models for the formation of ocean lithosphere. For nearly 2 decades, regionally low and scattered seafloor heat-flux values remained unexplained.

By the 1970s, studies began to be done with improved navigation, and with more closely spaced measurements made in the context of local sediment and igneous crustal structure. Results provided a sound basis for the hypothesis that hydrothermal circulation in the igneous crust and advective loss through unsedimented igneous outcrops caused both the scatter and the lower-than-expected values in young areas (Lister, 1972). Further improvements to instrumentation and observational strategies, in particular the development of probes that could be used with great efficiency for multiple measurements during a single instrument lowering, and the practice of making measurements in the context of geologic structure, led to the use of heat flux in the study of the process of hydrothermal circulation itself (Williams et al., 1974). With this new knowledge, it was possible to decipher the variability in measurements in a way that could lead to a better quantification of deep-seated heat flux, the goal of the original marine heat flow studies, and to understand the hydrologic processes behind the perturbations. Thus began a diverse range of applications of marine heat flow over a broad range of scales. A summary of the suite of tools currently in use for these studies is provided in the next section, along with a brief description of how heat-flux determinations are made. This is followed by a few examples of data from specific studies that illustrate how data are used, and a summary of some of the major conclusions that have been made through such studies.

Methods

Shallow measurements in marine sediments

Heat flux through the seafloor is often determined using temperatures measured with a series of sensors mounted on the outside of gravity-driven corers (Figure 1a), and thermal conductivities measured on the recovered sediment cores. Depths of penetration in excess of 10 m can be achieved in soft sediment, providing a valuable check on potential perturbations from bottom-water temperature variations, although accuracy is often limited by physical disturbances caused by the coring process, by changes in the physical properties of the recovered material, by



Heat Flow, Seafloor: Methods and Observations, Figure 1 Sensors and probes for measuring temperatures and thermal conductivity in marine sediments, including (**a**) outrigger temperature sensors mounted to the outside of a sediment corer (core barrel is 12 cm diameter), (**b**) a devoted heat-flux probe for measuring sediment temperatures and thermal conductivities (total length is 4.5 m), and (**c**) a high-strength probe that extends below a drill bit for bottom-hole temperature measurements (length is 1.2 m, tip diameter 1 cm).

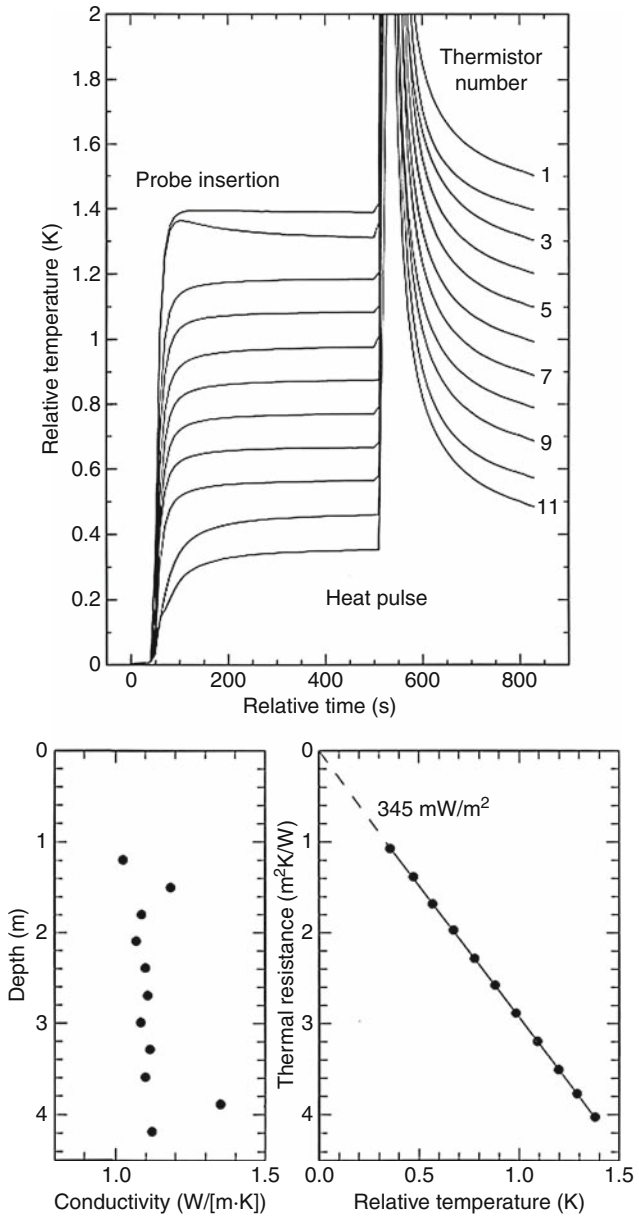
incomplete recovery, and by the imperfect depth registration between the cores and the intervals between the temperature sensors. Probes devoted exclusively to heat-flux measurements are typically limited to lengths of a few meters, but they have several distinct advantages. They allow thermal conductivity to be measured under in situ conditions and at depths that are co-registered with temperature measurements, and they allow transects of many measurements to be made efficiently during single instrument lowerings. A typical multipenetration heat-flux probe (Figure 1b) employs a heavy strength member that resists bending during repeated penetration and withdrawal from the sediment, and a small-diameter, rapidly responding tube containing thermistor sensors and a linear heater element. In situ temperatures are estimated by extrapolating transient decays following probe penetration, and conductivities are determined from the rate of change of temperature following steady or impulsive activation of the heater. A typical data record is shown in Figure 2, as are the resulting determinations of temperature and thermal conductivity. Heat flux is determined as the linear regression fit of temperature versus cumulative thermal resistance, R:

$$R = \sum [\Delta z / \lambda(z)]$$

where λ is thermal conductivity measured at a series of depths, z, and Δz is the depth interval assumed to be represented by each measurement. This is equivalent to calculating heat flux as the product of thermal gradient and the harmonic mean of thermal conductivity between temperature measurements, as was done in early marine heat flow studies. Errors associated with possible bottom-water temperature variations are evaluated by examining systematic deviations from linearity as a function of the number of thermistors included in the fit, working progressively up toward the shallowest measurement point. Complete descriptions of instruments and discussions of data reduction methods can be found in Lister (1979); Hyndman et al. (1979); Davis (1988), Wright and Louden (1989), and Villinger and Davis (1987).

Deep borehole measurements

Where observations are needed in hard formations or at depths greater than can be penetrated with a gravity-driven device, drilling is required. For research objectives, this has been done primarily through the Deep Sea Drilling Project, the Ocean Drilling Program, and the Integrated Ocean Drilling Program. In relatively unconsolidated sediments (typically the uppermost 50–100 m below the seafloor), hydraulically driven piston corers are deployed from the bottom of the drill string, and temperatures are measured at the tip of the core barrel (Horai and Von Herzen, 1985). At greater depths below the seafloor, high-strength probes can be pushed in with the weight of the drill string sufficiently far below the bottom of the hole (c. 1 m) to gain an unperturbed measurement (Uyeda and



Heat Flow, Seafloor: Methods and Observations, Figure 2 Typical data (upper panel) collected with a marine heat-flux probe like that shown in Figure 1b. Thermal conductivities are determined from the rates of decay following the metered pulse of heat, and natural sediment temperatures are determined by extrapolating the transients following probe insertion. In this example, high conductivities associated with two turbididic sand layers are present.

Horai, 1980; Davis et al., 1997a) (Figure 1c). Deeper than a few hundred meters in sediment, or at any level in crystalline rock, bottom-hole measurements are not feasible; instead, long-term borehole measurements are required to discriminate the natural formation thermal state from the commonly large and long-lived perturbations from drilling and subsequent fluid flow into or out of the hole. The most reliable method for determining the natural thermal state of crustal rocks has been to seal holes and install thermistor strings for long-term monitoring (Davis et al., 1992) (Figure 3).

Example studies

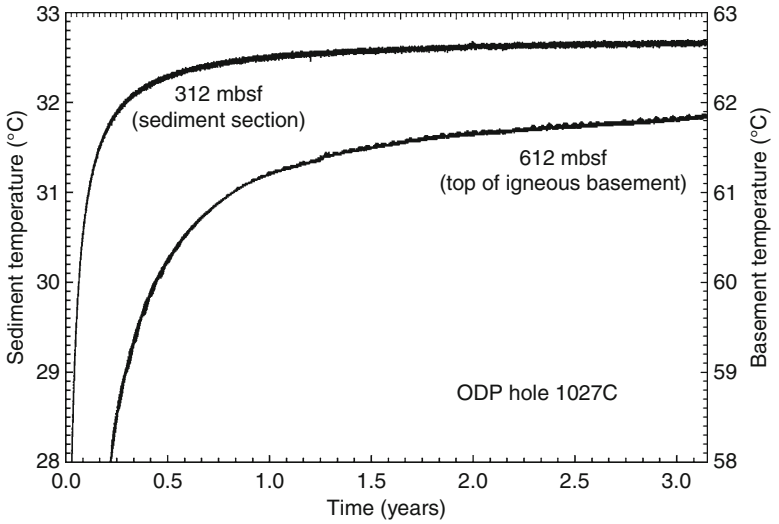
Bottom-water temperature perturbations

Seafloor heat-flux measurements rely on the assumption that both long- and short-term bottom-water temperature variations are small; the volume of ocean bottom water is very large and the temperature of the source in polar regions is regulated by the formation of sea ice. Many early measurements were made with temperatures determined at only three depths over a span of a few meters, however. Good checks on the validity of this assumption were not possible until temperature observations began to be made in deep-sea boreholes and long records of bottom-water temperature were acquired (Hyndman et al., 1984; Davis et al., 2003). An ideal suite of observations that would allow errors associated with bottom-water temperature variations to be quantified throughout the world's oceans – one that is broadly distributed both geographically and with ocean depth – does not yet exist, but the available data show that gradients measured a few meters below the seafloor generally do permit accurate determinations of heat flux in large areas of the oceans where depths are greater than ~2,000 m. One example where errors are demonstrated to be small is illustrated in Figure 4, where closely colocated seafloor probe and borehole observations are compared. Significant bottom-water temperature variations are ruled out by the linearity of the plots of temperature versus cumulative thermal resistance, and by the agreement between the shallow probe and deep borehole determinations. This illustration also shows the importance of precise colocation when doing such a comparison, given the local spatial variability of heat flux as defined by neighboring probe measurements.

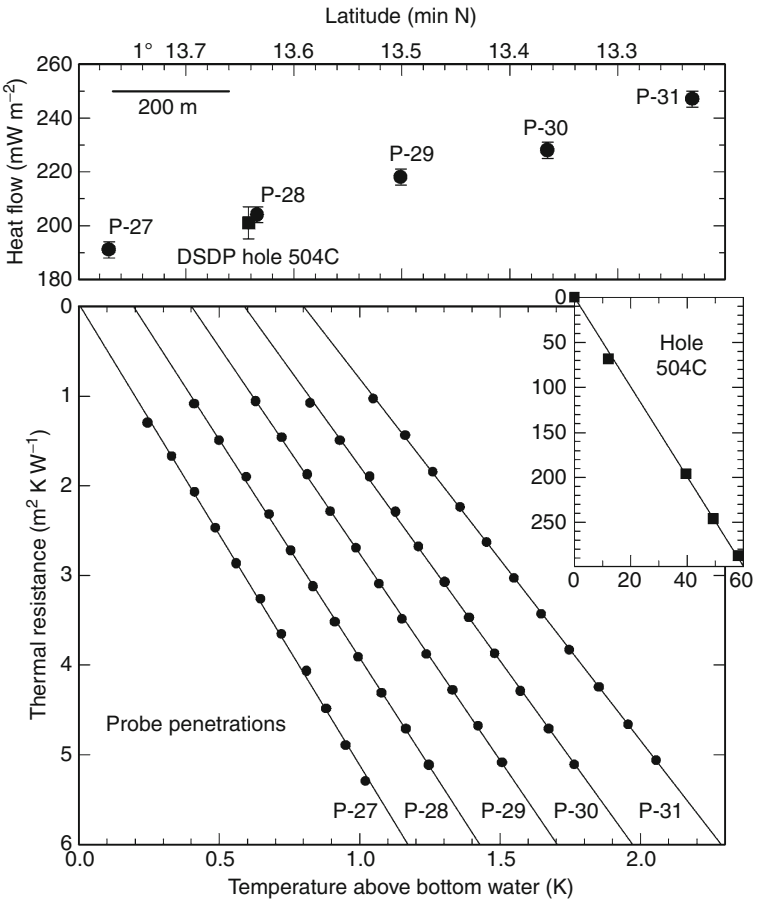
A more direct approach uses observations of bottom-water fluctuations. An example from a 5,000-m-deep site in the western Atlantic (Figure 5a) shows that in this area, oceanographic perturbations might have a modest influence on measured gradients. Estimated gradient perturbations at depths of less than 2–3 m below the seafloor range up to 10 mK m^{-1} (Figure 5b), that is, up to 20% of the geothermal gradient if the heat flux were 50 mW m^{-2}. A second example from a 2,600-m-deep site in the eastern Pacific shows smaller variability (Figure 5c), although perturbations estimated at a depth of 2 m could still result in a heat-flux determination error of up to 10%. Observations like these are clearly useful for guiding measurement strategy (e.g., depth of penetration) wherever heat flux is low and precise determinations are required.

Heat-flux signals from hydrothermal circulation

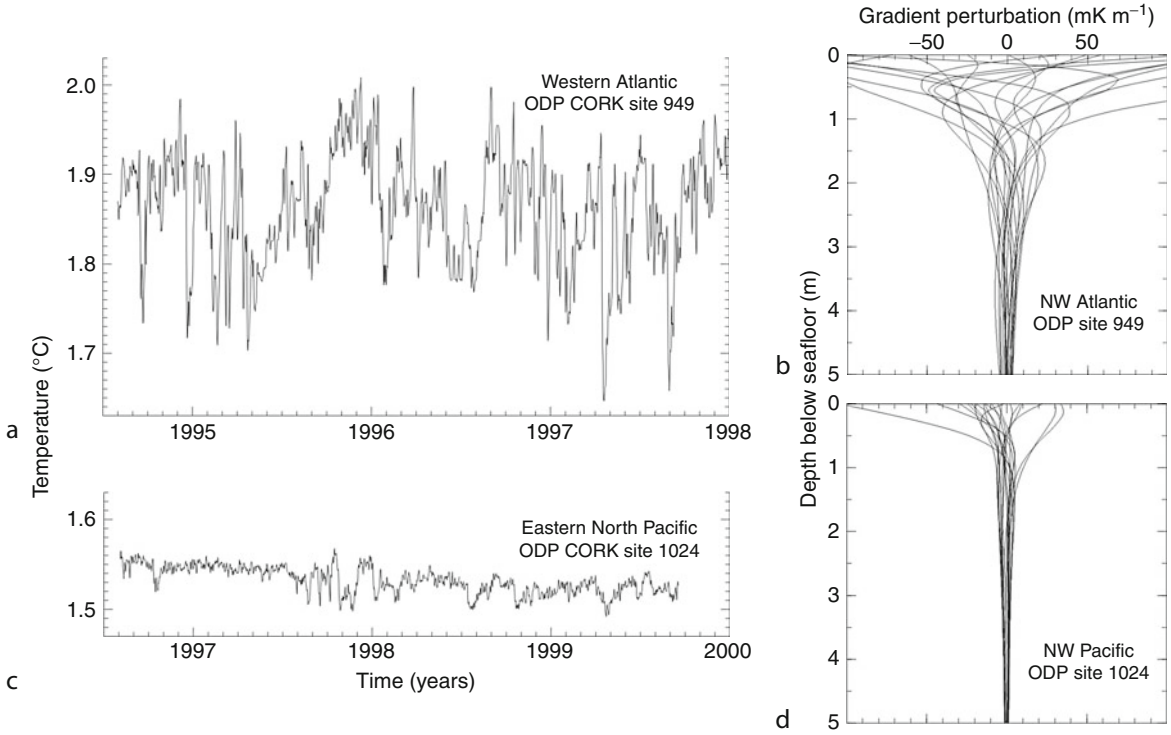
Hydrothermal circulation is a major source of error when determinations of deep-seated heat flux are sought, but it



Heat Flow, Seafloor: Methods and Observations, Figure 3 Two records from a 10-thermistor cable suspended in a borehole that penetrates a c. 600-m-thick sediment layer and into the underlying uppermost igneous oceanic crust. The longer recovery time at the deeper level reflects the large volume of water that invaded the uncased and permeable igneous section during the 5 days between the time of drilling and when the hole was sealed and instrumented.



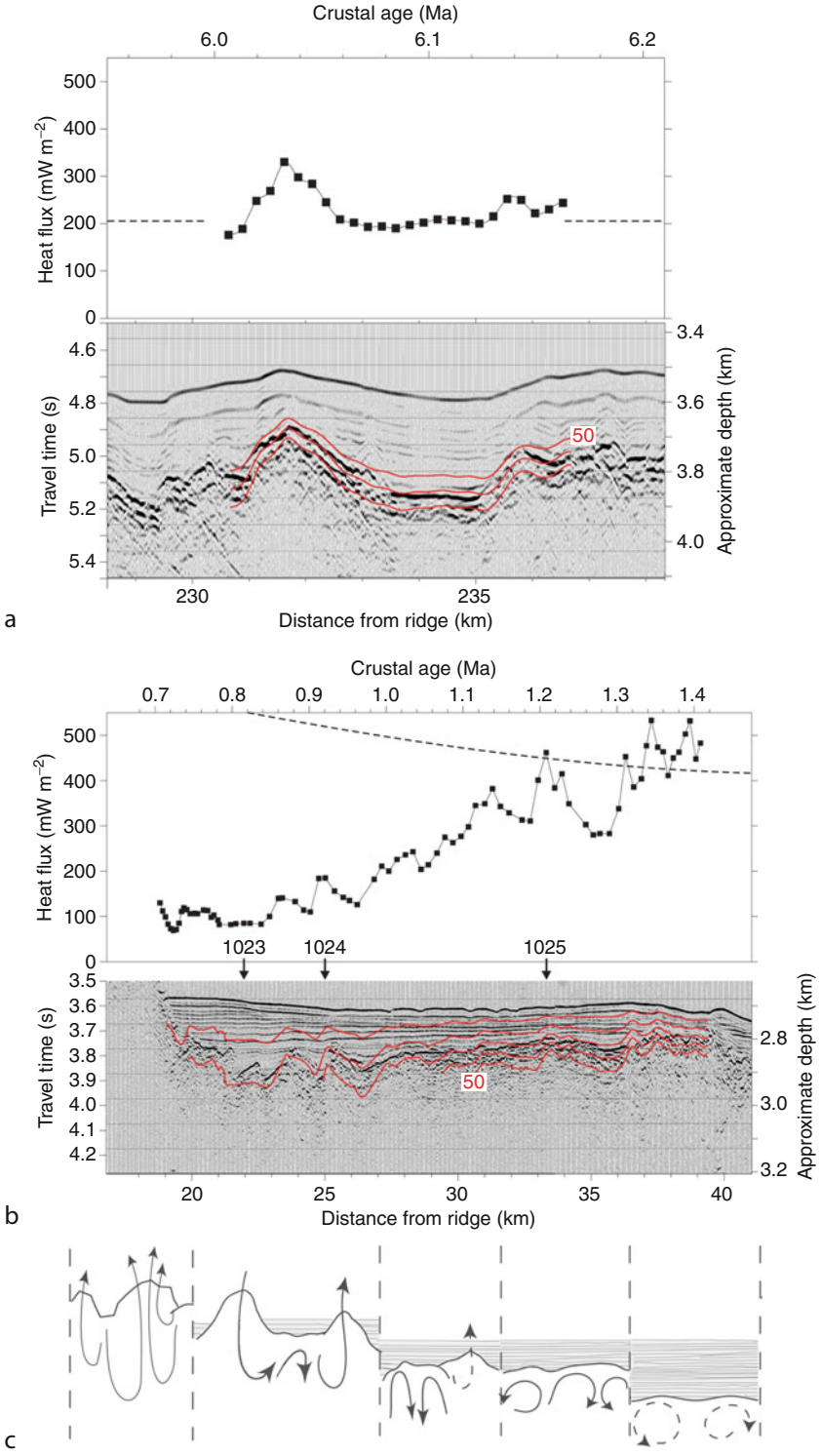
Heat Flow, Seafloor: Methods and Observations, Figure 4 Comparison of colocated seafloor probe and borehole heat-flux observations. Variations in probe measurements along the transect adjacent to the borehole site illustrate how carefully such comparisons must be done.



Heat Flow, Seafloor: Methods and Observations, Figure 5 Bottom-water temperature variations observed in the Atlantic (**a**) and Pacific Oceans (**c**), and estimated perturbations of the geothermal gradient as a function of depth (**b** and **d**) stemming from the variations, calculated at an evenly distributed suite of times during the temperature time series.

is also an important geological process. Hence it has been the focus of a large number of studies. Instructive examples of the influence of hydrothermal circulation are shown in Figure 6, where closely spaced measurements were made along transects striking perpendicular to basement structure. The first (Figure 6a) is in an area where sediment cover is continuous over a broad region spanning several tens of kilometers. Locally, heat flux varies inversely with sediment thickness. Such variations are common in areas of young seafloor, but can occur across relatively old seafloor as well (Embley et al., 1983; Von Herzen, 2004; Fisher and Von Herzen, 2005). They are the consequence of thermally efficient, local convection in permeable igneous rocks beneath low-permeability sediment cover. If the permeability of the igneous "basement" formation is high, vigorous convective flow maintains nearly constant temperatures at the sediment/basement interface despite variations in the thermal resistance of the overlying sediment layer. In this instance, the average seafloor heat flux is close to that expected from the underlying lithosphere, suggesting that from a thermal perspective, the circulation in the upper igneous crust is sealed in by the extensive sediment cover. The second transect (Figure 6b) crosses a sediment-covered area immediately adjacent to an area of outcropping igneous crust (where measurements are impossible). Local variations like those in Figure 6a are present, but even more apparent is a systematic variation of a larger scale, with heat flux increasing with distance from the area of basement outcrop, opposite to the expected trend of decreasing heat flux with increasing seafloor age. Temperatures estimated at the top of the igneous section increase systematically as well, suggesting that heat is transported laterally by fluid circulation and mixing in the sediment-sealed igneous crust. Heat exchange between the well-ventilated and sediment-sealed areas, indicated by the heat-flux deficit in this example, suggests a lateral heat-transfer scale of 20 km. Examples elsewhere suggest that the effects of advective heat loss may be felt laterally as far as 50–100 km (e.g., Fisher et al., 2003).

Ever since the early work of Lister (1972), the mere presence of local variability has been used as a diagnostic indicator of hydrothermal circulation in both young and old areas (e.g., Figure 6c), but with widely spaced observations, neither of the signals exemplified in Figure 6a and b could be resolved coherently; values were simply scattered and averages were often low. When systematic, detailed transects of observations began to be completed in context of colocated seismic data, the vigor of the convection could be inferred quantitatively from the nonconductive thermal regime (Fisher and Becker, 1995; Davis et al., 1997b), and the amount of heat lost



Heat Flow, Seafloor: Methods and Observations, Figure 6 Transects of heat flux on the flanks of the Costa Rica Rift (**a**) where an extensive sediment cover is present, and Juan de Fuca Ridge (**b**) striking away from an area of extensive basement at the left end of the figure. Both show the effects of hydrothermal circulation on conductive seafloor heat flux and on the crustal thermal regime. Heat flux estimated on the basis of the local lithosphere age (see text) is shown as the dashed lines. Temperatures estimated below the seafloor are shown at intervals of 10 °C. The cartoons in (**c**) show the influence of sediment burial on hydrothermal circulation and advective heat loss under a variety of burial states.

from the crust by fluid advection could be estimated with growing confidence (e.g., Anderson and Skilbeck, 1980; Stein and Stein, 1994; Harris and Chapman, 2004).

Two major lessons are learned from detailed observations like these for drawing conclusions about deep-seated heat flux. First, to ensure that observations do not suffer from the bias caused by convective ventilation, it must be demonstrated that there are no exposures of permeable rock at faults or volcanic edifices within distances of several tens of kilometers. Second, large numbers of closely spaced observations must be made, ideally colocated with seismic reflection data, so that the local variability can be understood and meaningfully averaged, and the locally relevant lateral transport scale can be assessed (e.g., Sclater et al., 1976; Davis et al., 1999).

Lessons learned about the way that the seawater interacts with the oceanic crust are far-reaching and continuously expanding. Estimates for the temperatures of circulation, the chemistry of the fluids, the volumetric rates of exchange between the crust and the ocean, and the consequent effects on crustal alteration and ocean chemistry have become reasonably well understood (e.g., Mottl and Wheat, 1994; Elderfield and Schultz, 1996). Studies of the actual distribution of crustal permeability, the percentage of rock affected by hydrothermal alteration, and the potential for chemosynthetic microbial populations are the focus of current investigations.

Dependence of heat flux on age and the global average

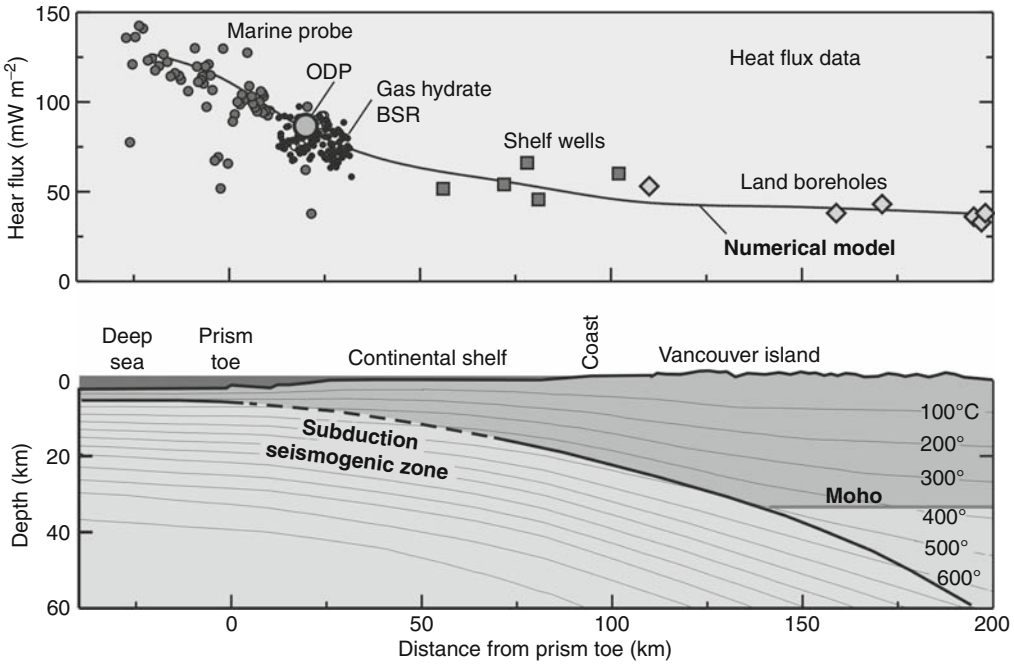
With the potentially large influence of hydrothermal circulation in mind, it is clear that a simple compilation of heat-flux data will provide a deceiving view of global heat loss. Except in old ocean basins, values are likely to be scattered and low relative to the heat loss expected from the underlying lithosphere. But by taking only those measurements that are sufficiently far from known permeable crustal outcrops and sufficiently numerous to provide a reliable local average, a subset of data can be gathered that provides a reliable determination of deep-seated heat flux. When considered in the context of lithospheric age, the results have been found to be consistent with both the characteristics of age-dependent seafloor subsidence and with simple lithospheric cooling theory (see *Lithosphere, Oceanic: Thermal Structure*). In young areas, heat flux is found to decline linearly with the inverse square root of age, following the simple relationship $Q = \mathrm{C}\, t^{-1/2}$ (where Q is heat flux in $\mathrm{mW\,m^{-2}}$, t is age in Ma, and C is a constant estimated between 475 and 510; Lister, 1977; Harris and Chapman, 2004). High-quality observations in older regions ($>$100 Ma) are generally uniform, in the range of 45–50 $\mathrm{mW\,m^{-2}}$ (Lister et al., 1990), suggesting that the thermal structure of the lithosphere may become stabilized in a state regulated either by the convectively supplied heat flux from the underlying asthenosphere, or by the combination of a compositionally established lithospheric thickness and the relatively uniform temperature of the vigorously convecting asthenosphere.

With the relationship between heat flux and age thus defined, the problems of the unknown bias and large scatter in young areas and the sparse distribution of measurements in large portions of the oceans can be overcome. A reliable estimate for the total heat flow through the floor of the ocean can be had by using the area/age relationship for the oceans defined by seafloor magnetic anomalies (e.g., Parsons, 1982; Wei and Sandwell, 2006), along with a robust heat-flux/age relationship. Several such estimates have been made (e.g., Williams and Von Herzen, 1974; Sclater et al., 1980; see summary in Jaupart et al., 2007), and all fall in a relatively narrow range centered around 32 TW (with contributions from marginal seas and hotspot swells included). This yields an average seafloor flux of roughly 107 $\mathrm{mW\,m^{-2}}$, a number that has little physical significance, but is considerably greater than that estimated in the early days of marine heat flow, and greater than the average though continents (c. 67 $\mathrm{mW\,m^{-2}}$), particularly when the latter is adjusted for the contribution of continental crustal radiogenic heat (c. 33 $\mathrm{mW\,m^{-2}}$).

With the total heat flow thus constrained, the heat lost advectively by ventilated circulation can be estimated from the difference between this and the age-binned average of unfiltered observations. Such estimates of this "heat deficit" fall in the neighborhood of 10 TW (Stein and Stein, 1994; Harris and Chapman, 2004). Most of this deficit occurs in seafloor less than 8–10 Ma in age, and it becomes insignificant on average by an age of 65 Ma. The actual age at which advective loss becomes insignificant is locally variable, depending primarily on the continuity of accumulating sediments that bury the igneous crust (Anderson and Skilbeck, 1980; Harris and Chapman, 2004), and the associated increase in spacing between basement outcrops that are essential for hydrothermal recharge and discharge on older ridge flanks (Fisher and Wheat, 2010).

The signature of subduction

Marine heat flux is used extensively to constrain deep thermal structure in studies of continental margins and marginal basins. A transect crossing the forearc prism of the Cascadia subduction zone illustrates one such application (Figure 7). This transect begins with standard gravity-driven probe measurements over the incoming plate and outermost accretionary prism, where bottom-water temperature variability is small. Where the seafloor is shallower than 1,500–2,000 m, other measurement techniques are used, including borehole measurements and estimates made using the depth to a bottom-simulating seismic reflector (BSR), which marks the limit of methane-hydrate stability. This reflector defines a unique set of pressure-temperature conditions, and with constraints on seismic velocity and thermal conductivity of the section above the BSR, the thermal gradient and heat flux can be estimated. Alternatively, a small set of seafloor



Heat Flow, Seafloor: Methods and Observations, Figure 7 Structural and heat-flux transect across the Cascadia subduction forearc, with temperatures estimated from a numerical model for underthrusting and sediment thickening, constrained by the heat-flux data (following compilation of Hyndman and Wang, 1993).

heat-flux measurements can be used as a "calibration." In either case, the travel-time depth to BSRs can serve as a widespread proxy for thermal data (e.g., Yamano et al., 1982). This technique is valuable where bottom-water temperature variability is too large to permit accurate heat-flux determinations with shallow probes, where sediments are too hard to allow probe penetration, or where there are few conventional measurements. Observations like these allow the thermal structure to be inferred deep within subduction zones, providing a critical constraint on the rheology of the rocks and the potential for seismogenic slip along the subduction thrust interface. In the example shown (Figure 7), the seafloor heat flux is variable locally, but regional values and trends are consistent with the expected thermal state of the thickly sedimented subducting plate.

In another study of the subducting Cocos Plate seaward of the Middle America Trench, variations in the thermal state of the plate are strongly influenced by regional differences in hydrothermal heat loss, and these correlate with differences in seismic processes occurring at depth (Newman et al., 2002; Fisher et al., 2003). One part of the plate is extensively cooled by hydrothermal circulation before the Cocos Plate is subducted, and earthquakes observed within the subduction wedge in this area are relatively deep (>20 km). Earthquakes tend to be shallower (<20 km) along an adjacent segment of the subduction zone, where there is no evidence for regional advective heat extraction. One explanation for the different earthquake depths is that cooling of part of the Cocos Plate slows dewatering and the transition of smectite to illite in subducting sediments (Spinelli and Saffer, 2004). Illite-rich sediments are more likely to undergo brittle deformation at depth, so the delay in heating associated with hydrothermal circulation in the crust prior to subduction causes a landward shift of the locked region where earthquakes are most likely to occur.

Summary

Seafloor heat flux can be measured with high accuracy in most deep-ocean settings with gravity-driven probes that penetrate a few meters into seafloor sediment. Improvements in heat-flux measurement technology, improvements in navigation, and integration with swath bathymetric, seismic, and other data that provide geological context for heat-flux measurements have greatly advanced our understanding of many global, regional, and local heat flow processes. Comparison of seafloor and borehole data has demonstrated that measurements made with short probes are accurate, provided that bottom-water temperature variations are relatively small. Compilations of global heat-flux data show that heat flux tends to vary systematically with seafloor age, following a $t^{-1/2}$ relation, at least until seafloor age exceeds 100 Ma, after which heat flux tends to become relatively constant. Determining the deep-seated lithospheric heat flux requires quantification of the potentially large influence of hydrothermal circulation in the permeable igneous rocks of the upper oceanic crust. This is best accomplished

through closely spaced transects of heat-flux measurements colocated with seismic reflection profiles that constrain the hydrologic structure, and regional maps that allow identification of basement outcrops. This approach has been applied in numerous settings, providing valuable constraints on the flow of water within the oceanic crust, the exchange of water, heat, and solutes between the crust and the oceans, the formation of hydrothermal mineral deposits, the accumulation of gas hydrates, and the development and maintenance of a subseafloor microbial biosphere. Individual heat-flux measurements and transects of measurements can be extended across broad regions using the depth to bottom-simulating seismic reflectors. These and other applications were never imagined by those who developed the original techniques for acquisition of seafloor heat-flux data 6 decades ago, but they illustrate how acquiring this kind of data has remained valuable for multidisciplinary studies of thermal, hydrogeologic, tectonic, and microbiological conditions and processes within the lithosphere.

Bibliography

Anderson, R. N., and Skilbeck, J. N., 1980. Oceanic heat flow. In Emiliani, C. (ed.), *The Sea*. New York: Wiley Interscience, Vol. 7, pp. 489–523.

Bullard, E. C., 1954. The flow of heat through the floor of the Atlantic Ocean. *Proceedings of the Royal Society A*, **222**, 408–429.

Davis, E. E., Wang, K., He, J., Chapman, D. S., Villinger, H., and Rosenberger, A., 1997a. An unequivocal case for high Nusselt number hydrothermal convection in sediment-buried igneous oceanic crust. *Earth and Planetary Science Letters*, **146**, 137–150.

Davis, E. E., 1988. Oceanic heat-flow density. In Haenel, R., Rybach, L., and Stegena, L. (eds.), *Handbook of Terrestrial Heat-Flow Density Determination*. Dordrecht: Kluwer, pp. 223–260.

Davis, E. E., Becker, K., Pettigrew, T., and Carson, B., 1992. CORK: a hydrologic seal and downhole observatory for deep-sea boreholes. *Proceedings ODP, Init. Repts.* 139, pp 45–53.

Davis, E. E., Villinger, H., Macdonald, R. D., Meldrum, R. D. J., and Grigel, J., 1997b. A robust rapid-response probe for measuring bottom-hole temperatures in deep-ocean boreholes. *Marine Geophysical Researches*, **19**, 267–281.

Davis, E. E., Chapman, D. S., Wang, K., Villinger, H., Fisher, A. T., Robinson, S. W., Grigel, J., Pribnow, D., Stein, J. S., and Becker, K., 1999. Regional heat flow variations on the sedimented Juan de Fuca Ridge eastern flank: constraints on lithospheric cooling and lateral hydrothermal heat transport. *Journal of Geophysical Research*, **104**, 17675–17688.

Davis, E. E., Wang, K., Becker, K., Thomson, R. E., and Yashayaev, I., 2003. Deep-ocean temperature variations and implications for errors in seafloor heat flow determinations. *Journal of Geophysical Research*, **108**, 2034, doi:10.1029/2001JB001695.

Elderfield, H., and Schultz, A., 1996. Mid-ocean ridge hydrothermal fluxes and the chemical composition of the ocean. *Annual Review of Earth and Planetnery Sciences*, **24**, 191–224.

Embley, R. W., Hobart, M. A., Anderson, R. N., and Abbott, D., 1983. Anomalous heat flow in the northwest Atlantic: a case for continued hydrothermal circulation in 80-M.y. crust. *Journal of Geophysical Research*, **88**, 1,067–1,074.

Fisher, A. T., and Becker, K., 1995. The correlation between heat flow and basement relief: observational and numerical examples and implications for upper crustal permeability. *Journal of Geophysical Research*, **100**, 12,641–12,657.

Fisher, A. T., and Von Herzen, R. P., 2005. Models of hydrothermal circulation within 106 Ma seafloor: Constraints on the vigor of fluid circulation and crustal properties below the Madeira Abyssal Plain, *Geochemistry, Geophysics, Geosystems*, **6**, 17 pp doi:10.1029/2005GC001013.

Fisher, A. T., and Wheat, C. G., 2010. Seamounts as conduits for massive fluid, heat and solute fluxes on ridge flanks. *Oceanography*, **23**(1), 74–87.

Fisher, A. T., Stein, C. A., Harris, R. N., Wang, K., Silver, E. A., Pfender, M., Hutnak, M. C., Cherkaoui, A., Bodzin, R., and Villinger, H., 2003. Abrupt thermal transition reveals hydrothermal boundary and role of seamounts within the Cocos plate. *Geophysical Research Letters*, **30**(11), 1550, doi:10.1029/2002GL016766.

Harris, R. N., and Chapman, D. S., 2004. Deep-seated oceanic heat flux, heat deficits, and hydrothermal circulation. In Davis, E. E., and Elderfield, H. (eds.), *Hydrogeology of the Oceanic Lithosphere*. New York: Cambridge University Press, pp. 311–336.

Horai, K., and Von Herzen, R. P., 1985. Measurement of heat flow on Leg 86 of the Deep Sea Drilling Project, in *Init. Repts., DSDP* 86, pp. 759–777.

Hyndman, R. D., 1984. Review of Deep Sea Drilling Project geothermal measurements through Leg 71. *Init. Repts. DSDP* 78B, pp. 813–823.

Hyndman, R. D., Davis, E. E., and Wright, J. A., 1979. The measurement of marine geothermal heat flow by a multipenetration probe with digital acoustic telemetry and in situ thermal conductivity. *Marine Geophysical Researches*, **4**, 181–205.

Hyndman, R. D., and Wang, K., 1993. Thermal constraints on the zone of major thrust earthquake failure: the Cascadia subduction zone. *Journal of Geophysical Research*, **98**, 2039–2060.

Jaupart, C., Labrosse, S., and Mareschal, J.-C., 2007. Temperatures, heat, and energy in the mantle of the Earth. In Bercovici, D. (ed.), *Treatise on Geophysics.* Oxford: Elsevier Vol. 7, pp 253–303.

Lee, W. H. K., and Uyeda, S., 1965. Review of heat flow data in Lee, W. H. K. (ed.), *Terrestrial Heat Flow, Geophysical Monograph 8*, Washington: American Geophysical Union, pp. 87–190.

Lister, C. R. B., 1972. On the thermal balance of a mid-ocean ridge. *Geophysical Journal International*, **26**, 515–535.

Lister, C. R. B., 1977. Estimators for heat flow and deep rock properties based on boundary layer theory, Tectonophysics, **41**, 157–171.

Lister, C. R. B., 1979. The pulse-probe method of conductivity measurement. *Geophysical Journal of Royal Astronomical Society*, **57**, 451–461.

Lister, C. R. B., Sclater, J. G., Davis, E. E., Villinger, H., and Nagihara, S., 1990. Heat flow maintained in ocean basins of great age: investigations in the north-equatorial west Pacific. *Geophysical Journal International*, **102**, 603–630.

Mottl, M. J., and Wheat, C. G., 1994. Hydrothermal circulation through mid-ocean ridge flanks: fluxes of heat and magnesium. *Geochimica et Cosmochimica Acta*, **58**, 2,225–2,237.

Newman, A. V., Schwartz, S. Y., Gonzalez, V., DeShon, H. R., Protti, J. M., and Dorman, L. M., 2002. Along-strike variability in the seismogenic zone below Nicoya Peninsula, Costa Rica. *Geophysical Research Letters*, **29**, 1–4, doi:10.1029/2002GL015409.

Parsons, B., 1982. Causes and consequences of the relation between area and age of the ocean floor. *Journal of Geophysical Research*, **87**, 289–303.

Petterson, H., 1949. Exploring the bed of the ocean. *Nature*, **4168**, 468–470.

Revelle, R. R., and Maxwell, A. E., 1952. Heat flow through the floor of the eastern North Pacific Ocean. *Nature*, **170**, 199–202.

Sclater, J. G., Crowe, J., and Anderson, R. N., 1976. On the reliability of ocean heat flow averages. *Journal of Geophysical Research*, **81**, 2,997–3,006.
Sclater, J. G., Jaupart, C., and Galson, D., 1980. The heat flow through oceanic and continental crust and the heat loss of the earth. *Reviews of Geophysics*, **18**, 269–311.
Spinelli, G. A., and Saffer, D., 2004. Along-strike variations in underthrust sediment dewatering on the Nicoya margin, Costa Rica related to the updip limit of seismicity. *Geophysical Research Letters*, **31**, 1–5, doi:10.1029/2003GL018863.
Stein, C. A., and Stein, S., 1994. Constraints on hydrothermal heat flux through the oceanic lithosphere from global heat flow. *Journal of Geophysical Research*, **99**, 3,081–3,095.
Uyeda, S., and Horai, K. 1980. Heat flow measurements on Deep Sea Drilling Project Leg 60, in *Init. Repts., DSDP* 60, pp 789–800, Washington, U. S. Govt. Printing Office.
Von Herzen, R. P., and Uyeda, S., 1963. Heat flow through the eastern Pacific Ocean floor. *Journal of Geophysical Research*, **68**, 4,219–4,250.
Villinger, H., and Davis, E. E., 1987. A new reduction algorithm for marine heat flow measurements. *Journal of Geophysical Research*, **92**, 12,846–12,856.
Von Herzen, R. P., 2004. Geothermal evidence for continuing hydrothermal circulation in older (> 60 M.y.) ocean crust. In Davis, E. E., and Elderfield, H. (eds.), *Hydrogeology of the Oceanic Lithosphere*. Cambridge: Cambridge University Press, pp. 414–447.
Wei, M., and Sandwell, D., 2006. Estimates of heat flow from Cenozoic seafloor using global depth and age data. *Tectonophysics*, **417**, 325–335.
Williams, D. L., and Von Herzen, R. P., 1974. Heat loss from the earth: new estimate. *Geology*, **2**, 327–330.
Williams, D. L., Von Herzen, R. P., Sclater, J. G., and Anderson, R. N., 1974. The Galapagos spreading center: Lithospheric cooling and hydrothermal circulation. *Geophysical Journal of Royal Astronomical Society*, **38**, 587–608.
Wright, J. A., and Louden, K. E., 1989. *Handbook of Seafloor Heat Flow.* Boca Raton: CRC Press. 498 pp.
Yamano, M., Uyeda, S., Aoki, Y., and Shipley, T. H., 1982. Estimates of heat flow derived from gas hydrates. *Geology*, **10**, 339–343.

Cross-references

Heat Flow Measurements, Continental
Heat Flow, Continental
Lithosphere, Continental: Thermal Structure

IMPACT CRATERS ON EARTH

Richard A. F. Grieve[1], Gordon R. Osinski[2]
[1]Earth Sciences Sector, Natural Resources Canada, Ottawa, ON, Canada
[2]Department of Earth Sciences/Physics and Astronomy, University of Western Ontario, London, ON, Canada

Definition

Impact crater. An approximately circular or polygonal depression formed by the hypervelocity impact of an interplanetary body (asteroid or comet) on a planetary surface.

Shock metamorphism. Irreversible changes to rocks and minerals, resulting from the passage of a shock wave produced by a hypervelocity impact.

The terrestrial record

Introduction

The results of planetary exploration have demonstrated that impact cratering is a ubiquitous process that has occurred throughout solar system history. On bodies that have retained portions of their early crust, such as the Moon, older surfaces are dominated by impact craters, attesting to the importance of impact cratering in early surface and crustal evolution. The Earth is the most geologically active of the terrestrial planets and is constantly renewing its surface. As impact craters are surface features, the net result of this high level of endogenous geological activity is that most of the original population of Earth's impact craters have been destroyed. What is left is a small preservation sample. The terrestrial record, however, is invaluable in the understanding of impact processes for it is the only source of ground-truth data on the three-dimensional structural and lithological character of natural impact craters.

Since the time of the Apollo missions in the 1960s, the ground-truth data supplied by terrestrial impact craters has served as analogues in understanding cratering processes on the other terrestrial planets; however, impact was generally not regarded as important to Earth's evolution. This changed in 1980, with evidence for a major impact as the cause of the mass extinction event at the Cretaceous-Paleogene (K-Pg) boundary ~65.5 million years ago. The actual crater, the ~180 km diameter Chicxulub crater, was identified in 1991, underneath ~1 km of sediments in the Yucatan peninsula, Mexico. The terrestrial impact record currently stands at ~180 craters or crater fields, with the largest being the ~300 km diameter Vredefort structure, South Africa. Evidence of distal ejecta from over two dozen impact events, some of which can be linked to their source craters, is also recorded in the stratigraphic record. They include four strewn fields of impact glasses known as tektites, which can occur over considerable areas, for example, the Australasian strewn field, which has currently no known source crater, is in excess of 50×10^6 km^2.

Distribution

The basic properties of known terrestrial impact craters are listed at http://www.unb.ca/passc/ImpactDatabase. The spatial distribution of known terrestrial impact structures is not random, with concentrations on the geologically stable cratons of the Australian, European, and North American continents. They are areas with relative tectonic stability and low rates of erosion, that is, they are the best available surfaces for the preservation of terrestrial impact craters. These are also the regions where there have been active programs to study impact craters. Terrestrial processes, such as differential erosion, can result in terrestrial impact craters evolving from an originally negative to a positive landform, that is, by definition, no longer a "crater." Thus, it is more appropriate to use the term

Harsh K. Gupta (ed.), *Encyclopedia of Solid Earth Geophysics*, DOI 10.1007/978-90-481-8702-7,

terrestrial impact "structure," which is not limited to a particular landform. Approximately one-third of the known terrestrial impact structures are buried by post-impact sediments. Detected initially as geophysical anomalies, their identity was confirmed by drilling. There are a number that are now underwater and several that were formed underwater, on relatively shallow continental shelves. At present, no impact structures are known on the true ocean floors, reflecting their relatively young age and the poorer geologic knowledge of this realm.

At large (>20 km) diameters, the cumulative size-frequency distribution of terrestrial impact structures is similar to that on the other terrestrial planets. This likely represents some form of steady-state condition between the formation of impact structures and their removal by terrestrial geologic processes. At smaller diameters, however, there is a deficit in the number of known terrestrial impact structures. This is due to the crushing of small impacting bodies during atmospheric passage, the relative ease with which smaller structures can be removed by erosion or buried, and the intrinsic difficulties in recognizing smaller structures. There is also a bias in the ages of known terrestrial impact structures, with most being <200 Ma old, which also reflects the problems of preservation in the terrestrial environment.

Morphology

Unless buried almost immediately, the forms of terrestrial impact structures are generally modified from their original. There is, however, the same overall progression in form, with increasing diameter, as on the other terrestrial planets. Smaller structures have the form of a bowl-shaped depression, with an upraised rim, and a depth of ~1/6 the diameter, and are termed simple structures (Figure 1). The exposed rim, walls, and floor define what is called the apparent crater. Beneath the apparent floor is an allochthonous breccia lens, which is roughly parabolic in cross section and floored by parautochthonous fractured target rocks. This breccia lens–parautochthonous target rock boundary defines what is called the true crater, with a depth of ~1/3 the rim diameter (Figure 1). For simple structures, the rim diameter is ~10 times that of the impacting body.

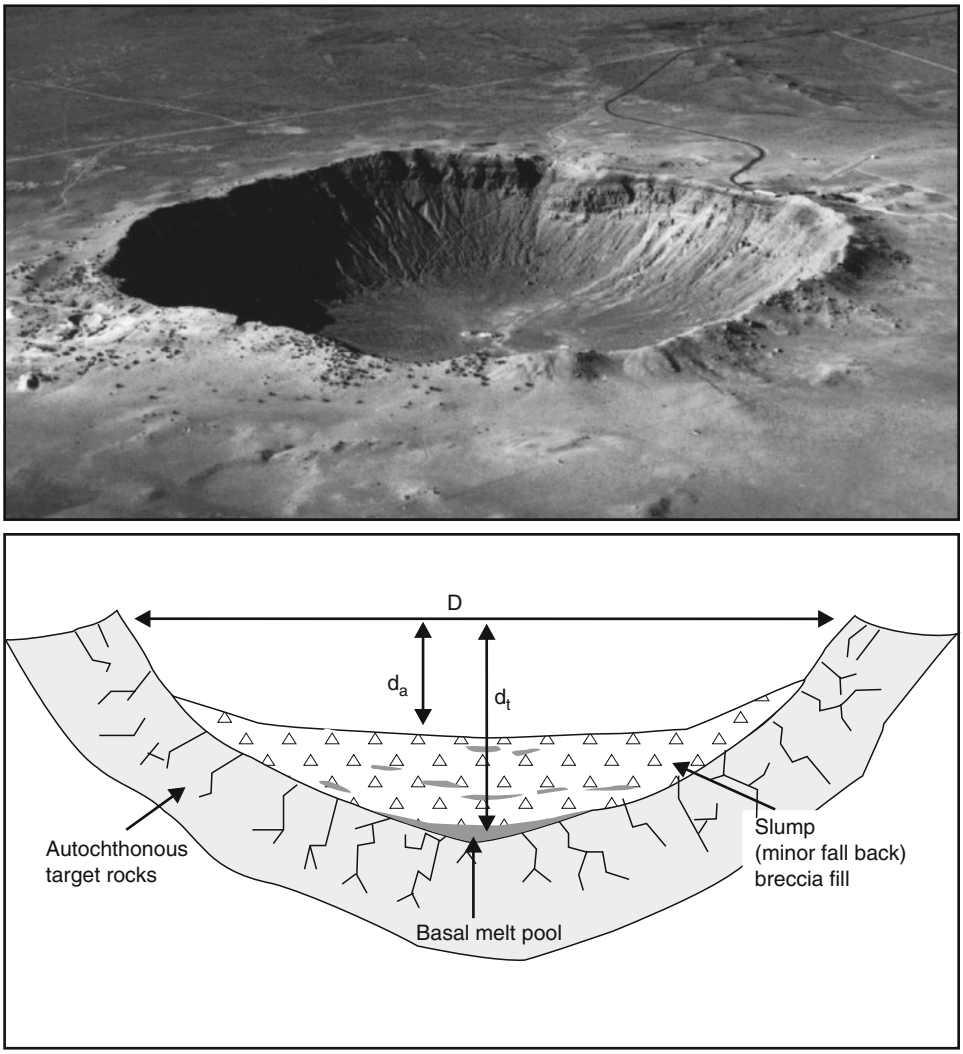
As the diameter increases, there is a change in form, with increasing wall and rim collapse, as simple structures evolve into complex impact structures, with rims consisting of structurally faulted terraces (Figure 2). Interior to the rim is a down-faulted annulus or trough, which is partly filled by allochthonous breccias and/or rocks melted by the impact. At the center is a structurally complex central peak of parautochthonous target rocks, which is the topographic manifestation of a more extensive volume of uplifted target rocks that lie beneath the floor in the center (Figure 2). The transition diameter from simple to complex structures appears to be an inverse function of planetary gravity and target rock strength, with the transition occurring at a diameter of ~4 km in crystalline and ~2 km in sedimentary target rocks, in the terrestrial environment. A number of relatively young (i.e., only slightly eroded) terrestrial complex structures – Haughton, Canada; Ries, Germany; Zhamanshin, Kazakhstan – lack an emergent central peak. These structures are in mixed targets of sediments overlying crystalline basement, with the lack of a peak most likely due to target strength effects. For complex structures, the rim diameter is ~20 times that of the impacting body.

With still increasing diameter, a fragmentary inner ring appears, marking the transition from a central peak structure to peak-ring basin. The larger impact structures in the terrestrial environment, however, tend to be older, eroded structures that have been modified, with their current elements being structural features rather than topographic features. Chicxulub is the best-preserved large terrestrial impact structure, due to burial. As such, however, the definition of its morphological elements depends on the interpretation of geophysical data. It has an interior topographic peak-ring, a terraced rim area, and exterior ring faults and, thus, appears to correspond to the definition of a multi-ring basin, as on the Moon. Details of the physics of cratering mechanics and impact, in general, can be found in Melosh (1989).

Geology of impact structures

Only small, young simple impact structures, where the atmosphere has partially slowed the impacting body, have physical evidence of the impacting body in the form of spalled fragments of iron or stony-iron meteorites, which are relatively quickly destroyed by weathering. (Equivalent-sized, weaker stony meteorites are generally crushed during atmospheric passage.) Larger impacting bodies (>100 m in diameter) impact the Earth with largely undiminished velocity. The average impact velocity of an asteroidal body is 20–25 km s^{-1} and cometary impacts have even higher velocities. Upon impact, most of the impacting body's considerable kinetic energy is partitioned into the target by means of a hemi-spherical propagating shock wave, which sets the target rocks in motion (kinetic energy) and compresses them (internal energy). A cratering flow-field is established in the target, leading to the excavation of target materials. During compression, a considerable amount of pressure-volume work is done. Not all this mechanical work is recovered on decompression to ambient pressure and the excess is manifested as waste heat. This leads to melting and even vaporization of some of the target rocks and the destruction of the impacting body.

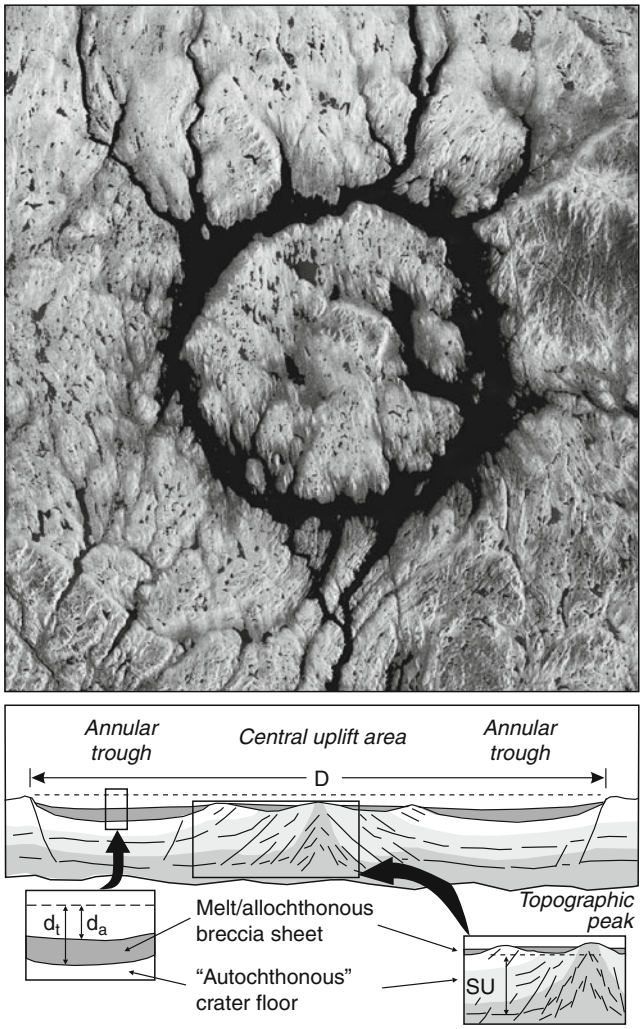
The transient compression, decompression, and heating of the target rocks lead to shock metamorphic effects, which record pressures, temperatures, and strain rates well beyond those produced in terrestrial regional or contact metamorphism. The exact physical conditions recorded depend of the peak pressure on impact, as determined by the impacting body's impact velocity and density, and distance from the point of impact. Given the



Impact Craters on Earth, Figure 1 *Top*: Oblique aerial photograph of the classic example of a terrestrial simple impact structure: The 1.2 km diameter Barringer or Meteor Crater, Arizona, USA. *Bottom*: Schematic cross section of a terrestrial simple impact structure, showing various lithological, structural, and morphologic attributes. Note allochthonous breccia lens, with impact-melted lithologies, partially filling the structure. *D* rim diameter, d_a apparent depth, d_t true depth.

highly transient nature of shock metamorphic processes, disequilibrium and metastable equilibrium are common. The nomenclature of shocked rocks or impactites (various melt rocks, breccias, and shocked target rocks) and the pressures required to generate particular shock metamorphic effects are summarized in Stöffler and Grieve (2007). Examples of impactites and shock metamorphic effects are well illustrated in French (1998).

The highest shock metamorphic effect preserved at impact structures is impact-melted material, which results from shock pressures >60 GPa, and takes the form of glass clasts in breccias, melt dikes in the floor of the structure, and lenses and coherent sheets of impact melt rock within the crater interior (Figure 3). When crystallized, impact melt rocks have an igneous-looking matrix and are often charged with mineral and lithic debris. Compositionally, these melted lithologies can be modeled as a mixture of target lithologies, in geologically reasonable proportions. Isotopic systems, such Sr, Nd, and Pb, mirror those of the initial ratios of the target rocks, while others (e.g., Ar) can be reset and record the time of impact. In some cases, siderophile-element enrichments above target rock levels occur, reflecting an admixture of material from the impacting body. In some cases, their relative abundances have identified the type of impacting body, for example, C-1 chondrite. In a few cases, Cr and Os isotopes have also been used to identify the impacting body. At lower shock pressures (~60–30 GPa), selective mineral melting occurs. In addition, the shock wave can destroy the internal crystallographic order of feldspars and quartz and convert them to solid-state glasses, which still have the original crystal shapes. These are



Impact Craters on Earth, Figure 2 *Top*: RADARSAT image of the Manicouagan, Canada, complex impact structure. The annular lake is 65 km in diameter and is, in part, man-made, serving as reservoir for hydroelectric power generation. The original diameter of this partially eroded structure was ~100 km and the central peak (Mont de Babel) rises to 600 m above the level of the lake. *Bottom*: Schematic cross section of a terrestrial complex impact structure, indicating various lithological, structural, and lithological attributes. Example shown has a central peak and an annular ring, which may have been the original form of Manicouagan. *D* rim diameter, d_a apparent depth, d_t true depth, *SU* structural uplift, which is ~0.1 D.

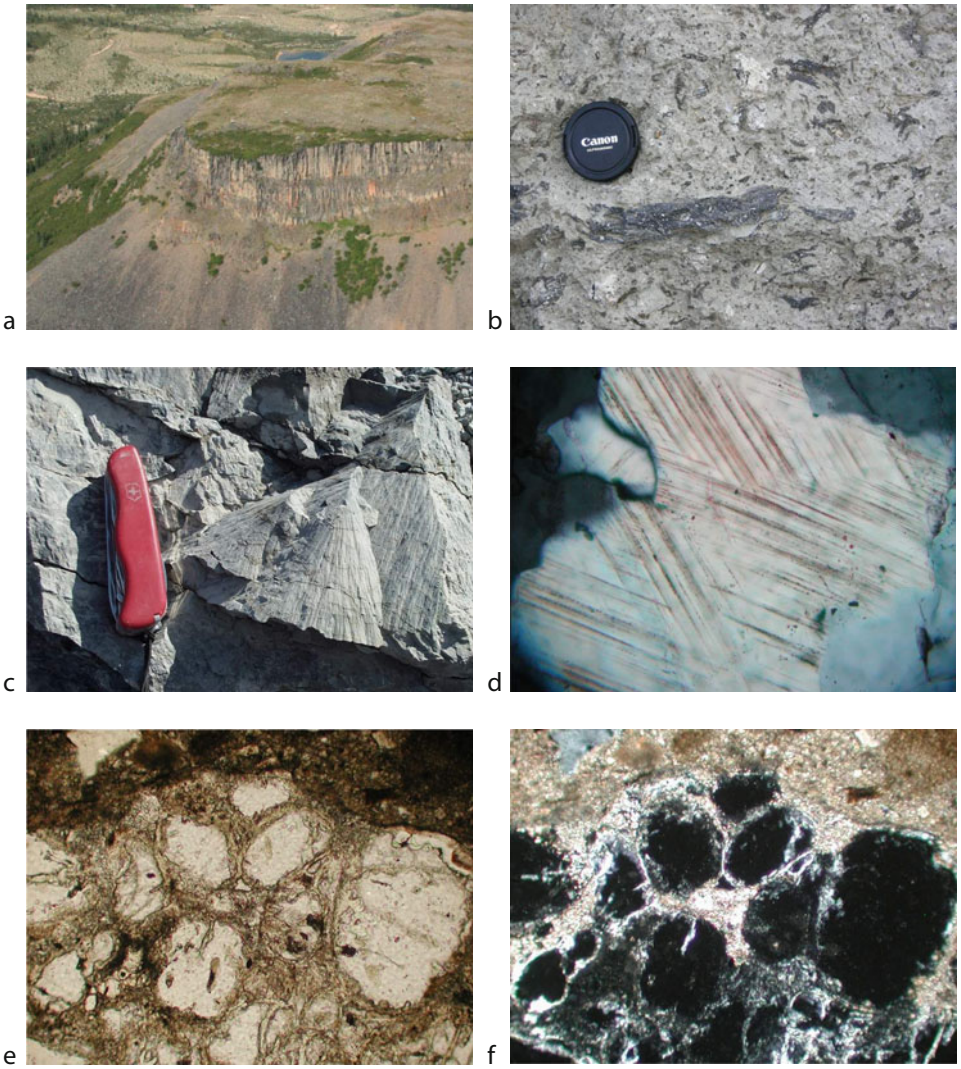
"diaplectic" glasses (Figure 3), with the required pressures being 30–45 GPa for feldspar and 35–50 GPa for quartz (diaplectic plagioclase glass is also known as maskelynite). The extremely rapid compression and then decompression also produces metastable polymorphs, for example, coesite and stishovite from quartz and diamond and lonsdaleite from graphite.

The most-documented shock metamorphic feature is the occurrence of so-called planar deformation features, particularly in quartz (Figure 3), although they do occur in other minerals, for example, feldspar, zircon. Their utility is a function of the common occurrence and stability of quartz in terrestrial rocks and the relative ease with which they can be observed and measured by optical microscopy. When fresh, planar deformation features are parallel planes of glass, with specific crystallographic orientations as a function of shock pressures of ~10–35 GPa. The only megascopic shock features are shatter cones, which are distinctive, striated, and horse-tailed conical fractures ranging in size from millimeters to tens of meters (Figure 3). They are best developed in fine-grained, structurally isotropic rocks, such as carbonates and quartzites, but more poorly developed variants can occur in coarser-grained rocks. They are most common in rocks shocked to relatively low levels, ~2–6 GPa, but can occur in rocks recording pressures as high as ~25 GPa.

Geophysics of impact structures

For buried impact structures, a geophysical anomaly is the first indicator of their existence, although confirmation of their genesis requires drilling and the documentation of shock metamorphic effects. The most common geophysical anomaly is a localized low in the regional gravity field (Figure 4), due to lowering of rock density from brecciation and fracturing. In general, modeling of the gravity anomalies is consistent with the morphometric parameters of terrestrial impact structures, with the anomaly extending to the rim of the structure. The anomaly increases to a maximum of ~200–300 g.u. at rim diameters of 20–30 km (Pilkington and Grieve, 1992), with the lack of further increases in magnitude at larger diameters due to lithostatic pressure closing any deep-seated impact-related fracturing. The gravity signature can be complicated by erosion and variations in target rock densities. Larger complex impact structures tend to have a central, relative gravity high, which can extend out to approximately half the diameter of the structure. This central, relative gravity high is from the uplift of originally deeper (more dense) lithologies in the structural uplift.

Magnetic anomalies are more varied. Most commonly, they are a magnetic low, with the disruption of any regional trends in the magnetic field. This is due to an overall lowering of magnetic susceptibility and the randomizing of pre-impact lithologic trends in the target rocks. Larger structures commonly have high-amplitude, short wave-length anomalies in the center, due to changes to remanent magnetization in the target rocks from the effects of heat, shock, and/or post-impact hydrothermal alteration. Seismic velocities are reduced at impact structures, due to fracturing, and reflection seismic images are extremely useful in characterizing buried structures in sedimentary targets. Subsurface brecciation, fracturing, and incoherency of seismic reflectors increase toward the center of complex structures due to structural uplift, resulting in a central volume that is isotropic in terms of a reflection seismic signal. Electrical methods have been used less but,

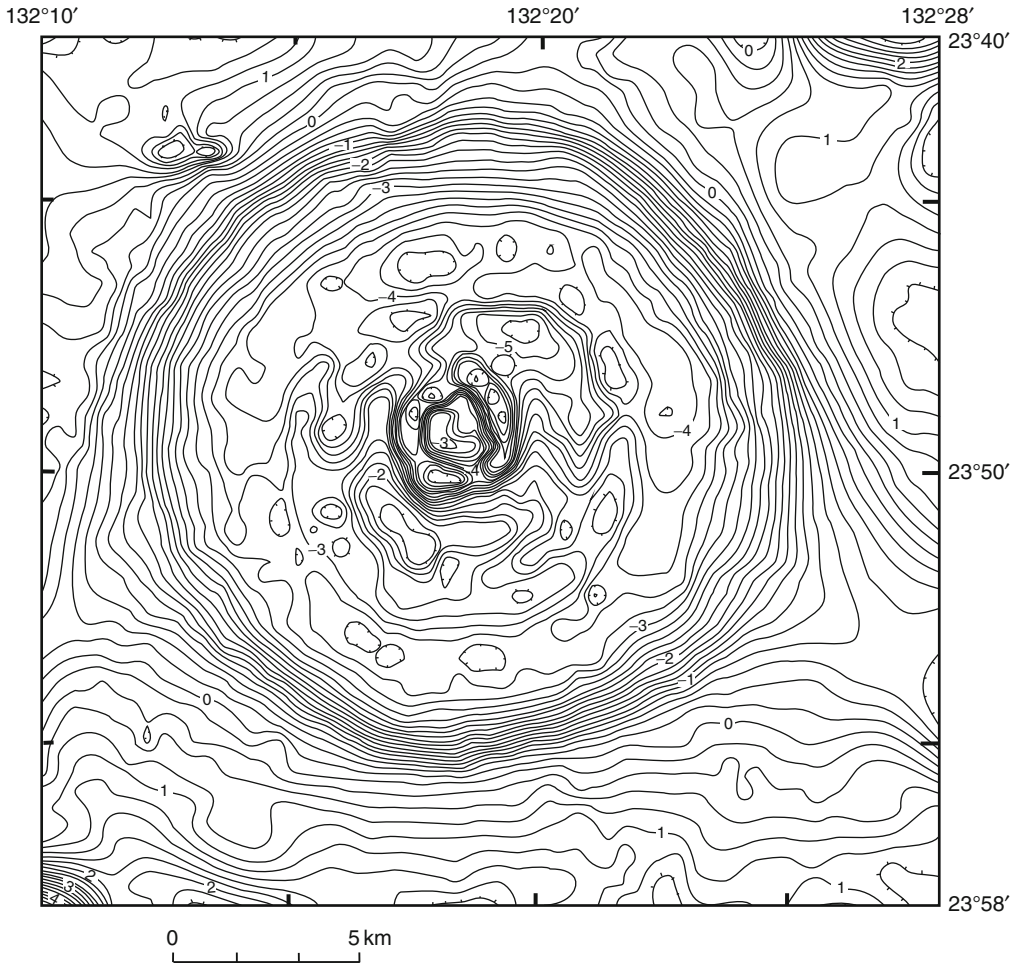


Impact Craters on Earth, Figure 3 Montage of impactites. (**a**) The impact melt sheet of the 28-km diameter Mistastin impact structure, Canada. Note the columnar jointing in the melt sheet. (**b**) Impact melt-bearing breccias (glassy melt rock occurs as black clasts) at the Mistastin structure. Six-centimeter diameter camera lens for scale. (**c**) Shatter cones from the Haughton impact structure, Canada. Eleven-centimeter long penknife for scale. (**d**) Photomicrograph (*plane light*) of Planar Deformation Features (PDFs) in quartz. Image is 0.35 mm across. (**e**) and (**f**) Photomicrograph of diaplectic quartz glass seen in plane (**e**) and cross-polarized light (**f**), respectively. Images are 5 mm across.

in general, there is a reduction in resistivity, which is related to fracturing. Resistivity contrasts can also delineate lithological contacts, for example, between allochthonous breccia deposits and the fractured target rock and/or post-impact sediments.

Impacts and Earth history

The formation of the terrestrial planets by the accretion of small bodies through collisions and subsequent growth to proto-planets includes potential collisions between these proto-planets. One such collision between the proto-Earth and a Mars-sized body is currently the best working hypothesis for the origin of the Moon. Such an impact would result in an Earth-orbiting accretionary disk of material from the impacting body and proto-Earth, which was the source of volatile- and siderophile-depleted material for building the Moon. The consequences for the proto-Earth include massive remelting and the loss of any original atmosphere. Without a Moon and lunar tides, there would be little or no littoral zone, which is the most important area in the terrestrial ecosystem, to Earth's oceans. With no littoral zone, the development and evolutionary path of Earth's biosphere would have been very different. Following planetary formation, the subsequent high rate of bombardment is attested to by the heavily



Impact Craters on Earth, Figure 4 Residual Bouguer gravity anomaly at the 22-km diameter Gosses Bluff impact structure, Australia, which is in the form of a symmetric gravity low, with a central, short-wave anomaly due to structural uplift. The target was a flat-lying sequence of sedimentary rocks and the simplicity in the form of the anomaly results from the lack of major asymmetric lithological/density contrasts. Contour interval is 1 g.u.

cratered lunar highland crust and the formation of the very large (up to 2,500 km in diameter) multi-ring impact basins. The Earth, however, has a larger gravitational cross section and will capture more incoming interplanetary bodies than the Moon. Thus, it has been subject to a greater number of impacts than the Moon throughout geological time, including the basin-forming impacts in the period of heavy bombardment. There have been speculation as to the potential effects of such impacts (e.g., Grieve et al., 2006), but no rocks are preserved from this period of the early Earth. Such impacts would retard any development of life on Earth but once they had subsided smaller impact events may have been beneficial for microbial life, with the generation of fertile hydrothermal systems, impact crater lakes, and new habitats within shocked rocks and minerals.

The temporal association of a large impact event (Chicxulub), world-wide ejecta, and a mass extinction at the K-Pg boundary is well established. As other large impact structures have not been linked to mass extinctions in the terrestrial biosphere, it may be that there was a unique characteristic to the Chicxulub event. Sulfate-bearing rocks occur at Chicxulub, raising the possibility of massive amounts of sulfur aerosols in the atmosphere after the impact. They would reduce light levels below those required for photosynthesis for the best part of a year, with the collapse of the food chain. Chicxulub-scale impacts occur every 100 ± 50 Ma. Smaller impacts occur on shorter timescales and loading of the atmosphere with dust from craters in the 20-km size-range would be global in scale. This would not produce a mass extinction but could severely disrupt or even destroy the relatively fragile infrastructure that supports human civilization.

Past-impacts have also been beneficial for the human condition. Approximately 25% of known terrestrial

impact structures have some form of natural resources, ranging from unexploited local deposits to world-class mining camps. For example, the ~1.5 billion tonnes of ores at the Sudbury impact structure, Canada, have so far produced over $ 300 billion worth of metals. Approximately 50% of impact structures in hydrocarbon-bearing sedimentary basins are producers. The world-class Campeche Bank oilfield in the Gulf of Mexico produces most of its hydrocarbon products (~ 12 billion barrels of oil and >3 trillion cubic feet of gas, since discovery) from Chicxulub-related breccias and has reserves exceeding the entire known reserves of the USA.

Concluding remarks

Although the first known recorded observations of impact craters was made by Galileo Galilei in 1609, it was not until the 1960s that the importance of impact cratering as a geological process began to be recognized. This came with the recognition of shock metamorphism and its effects on rocks and minerals and with impetus provided by the planned Apollo landings on the Moon. It is now widely acknowledged that impact cratering is one of the most important surface geological processes in the solar system. It has also become apparent that impact events have also played an important role throughout Earth's history, shaping the geological landscape, producing economic benefits, and affecting the evolution of life. The study of impact craters and related phenomena is relatively young when compared to other fields of scientific study and, as such, there is still considerable future potential for new and exciting discoveries.

Bibliography

French, B. M., 1998. Traces of catastrophe: a handbook of shock-metamorphic effects in terrestrial meteorite impact structures. LPI Contribution No. 954. Houston: Lunar and Planetary Institute.

Grieve, R. A. F., Cintala, M. J., and Therriault, A. M., 2006. Large-scale impact and evolution of the earth's crust: the early years. In Reimold, W. U., and Gibson, R. L. (eds.), *Processes on the Early Earth*. Boulder: Geological Society of America Special Paper 405, pp. 23–31.

Melosh, H. J., 1989. *Impact Cratering: A Geologic Process*. Oxford: Oxford University Press.

Pilkington, M., and Grieve, R. A. F., 1992. The geophysical signature of terrestrial impact craters. *Reviews of Geophysics*, **30**, 161–181.

Stöffler, D., and Grieve, R. A. F., 2007. A systematic nomenclature for metamorphic rocks: Impactites. In Fettes, D., and Desmons, J. (eds.), *Metamorphic Rocks: A Classification and Glossary of Terms*. Cambridge: Cambridge University Press, pp. 82–92.

Cross-references

Deep Scientific Drilling
Gravity Anomalies, Interpretation
Magnetic Anomalies, Interpretation
Seismic Imaging, Overview
Seismic Properties of Rocks

INSTRUMENTATION, ELECTRICAL RESISTIVITY*

Meng H. Loke[1], Jonathan E. Chambers[2], Oliver Kuras[2]
[1]Geotomo Software Sdn. Bhd., Gelugor, Penang, Malaysia
[2]Geophysical Tomography Team, British Geological Survey, Keyworth, Nottingham, UK

Definition

Electrical survey. Mapping subsurface resistivity by injecting an electrical current into the ground.

Resistivity meter. An instrument used to carry out resistivity surveys that usually has a current transmitter and voltage-measuring circuitry.

Electrode. A conductor planted into the ground through which current is passed, or which is used to measure the voltage caused by the current.

Apparent resistivity. The apparent resistivity is the resistivity of an equivalent homogeneous earth model that will give the same potential value as the true earth model for the same current and electrodes arrangement.

Multi-core cable. A cable with a number of independent wires.

Introduction

The resistivity survey method is more than 100 years old and is one of the most commonly used geophysical exploration methods (Reynolds, 1997). It has been used to image targets from the millimeter scale to structures with dimensions of kilometers (Linderholm et al., 2008; Storz et al., 2000). It is widely used in environmental and engineering (Dahlin, 2001; Chambers et al., 2006) and mineral exploration (White et al., 2001; Legault et al., 2008) surveys. There have been many recent advances in instrumentation and data interpretation resulting in more efficient surveys and accurate earth models. In its most basic form, the resistivity meter injects a current into the ground through two metal stakes (electrodes), and measures the resulting voltage difference on the ground surface between two other points (Figure 1). The current (I) and voltage (V) values are normally combined into a single quantity, the apparent resistivity, which is given by the following relationship:

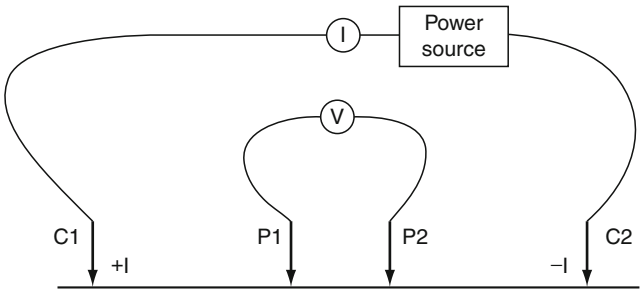
$$\rho_a = kV/I \quad (1)$$

The apparent resistivity is the resistivity of an equivalent homogeneous earth model that will give the same potential value as the true earth model for the same current and electrodes arrangement. The geometric factor k depends on the arrangement of the electrodes (see *Electrical Resistivity Surveys and Data Interpretation*). The flow of the current through the ground depends on the resistivity of the material (rocks) through which it passes and is reflected in the measured voltage values. Subsurface

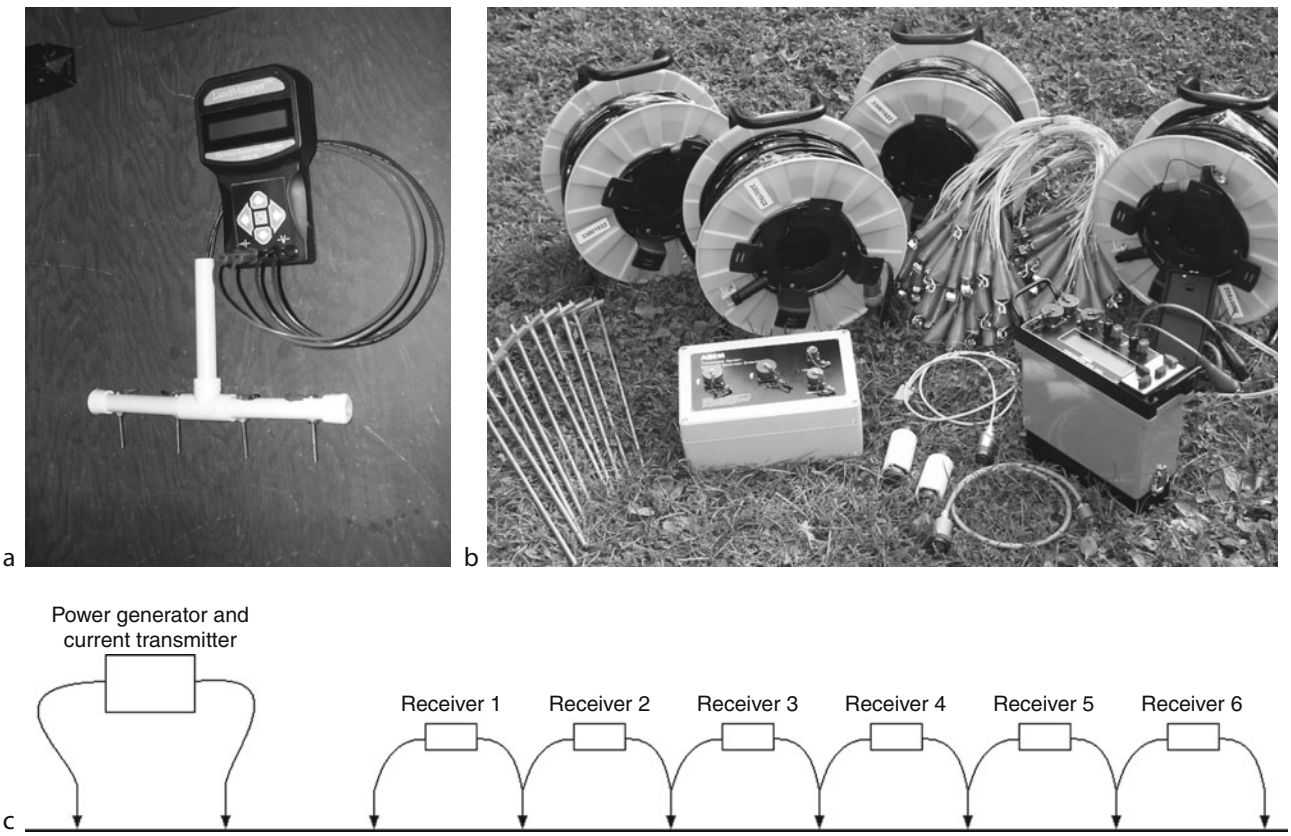
resistivity distribution can be mapped by making measurements with the current and voltage electrodes at different positions.

Basic resistivity instruments

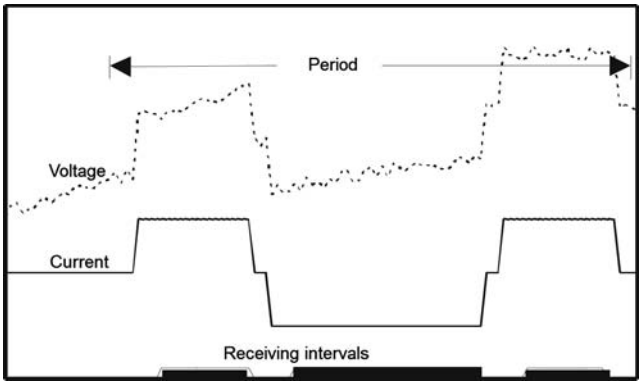
The basic parts of a resistivity measurement system include a source of electrical current, a voltage measuring system, and the cables to connect these components to the electrodes. These components have undergone major modifications over the years to improve the efficiency of the survey procedure, the quality of the data, and to increase the depth of investigation. The power source for the current is typically a battery for shallow surveys. Some small portable resistivity meter systems (Figure 2a) for very shallow surveys of up to a few meters depth weigh less than a kilogram (Pozdnyakov et al., 2009). A typical system (with associated cables and electrodes) for environmental and engineering surveys that uses an internal battery is shown in Figure 2b. Such systems usually weigh between 10 and 50 kg. The current source and voltage measuring circuitry are integrated into a single unit (the resistivity meter). Such battery-based systems can provide currents of up to about 2 A and have been used for survey depths of up to about 200 m. For deeper surveys where currents of up to 10 A are used, a petrol/diesel engine–powered electric generator is usually used. Such systems can weigh several hundred kilograms, and are commonly used for induced polarization (IP) surveys in mineral exploration (White et al., 2001). The current transmitter is separate from the receivers that measure the potential signal (Figure 2c).



Instrumentation, Electrical Resistivity, Figure 1 Basic setup for an electrical resistivity survey. The current is injected into the ground through the C1 and C2 electrodes, and the resulting voltage difference is measured by the P1 and P2 electrodes.



Instrumentation, Electrical Resistivity, Figure 2 Different classes of resistivity instruments. (**a**) A lightweight portable system (courtesy of Landviser LLC). The distance between the electrodes in the above picture is 5 cm. (**b**) A typical system used for shallow environmental and engineering surveys (courtesy of Torleif Dahlin and Abem Instruments AB.), and (**c**) a schematic diagram of a system for deep surveys.



Instrumentation, Electrical Resistivity, Figure 3 Example of input current and resulting voltage signal. (Courtesy of Torleif Dahlin and Abem Instruments AB.)

In the interpretation of the survey data, it is assumed that a direct current (DC) is used. However, in practice, a low-frequency square-wave alternating current (AC) is normally used. The frequency is sufficiently low (typically less than 100 Hz) such that EM effects are negligible. Figure 3 shows an example of the shape of the input current used and the resulting voltage signal. The amplitude of the voltage signal is typically in the millivolt range, but smaller voltages can be resolved if the signal-to-noise ratio is sufficiently high or signal-stacking techniques are employed. A 24-bit ADC (analog-to-digital converter) is widely used in modern systems to digitize the voltage signal. In the absence of an input current, the ground has a natural self-potential (SP) voltage between different points on the surface. This can be caused by variations in the chemical properties of the soil or currents induced by changes of the geomagnetic field with time. The SP voltage can vary significantly with time. The voltage measuring system in the resistivity meter has a SP buck-out mechanism to remove the SP voltage signal. It usually also has notch filters (e.g., 50, 60, 16.7 Hz) to remove noise due to electric power lines and other technical infrastructure such as electrified railway systems.

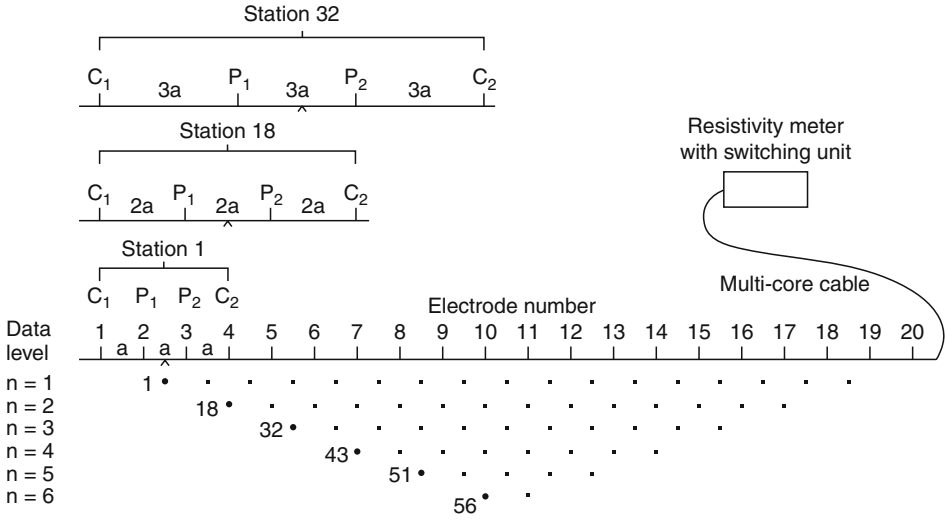
IP surveys are widely used in mineral exploration, and recently there has been increasing interest in the environmental and engineering sector (Aristodemou and Thomas-Betts, 2000; Dahlin et al., 2002). While both time and frequency domain (Telford et al., 1990; Reynolds, 1997) measurement techniques have been used in mineral exploration, the time domain method appears to be the norm for the less powerful equipment used in environmental and engineering surveys.

Electrode types

Metal stakes (Lu and Macnae, 1998) are commonly used for the current electrodes as well as the potential electrodes in resistivity surveys. For some ground conditions where it is difficult to insert a stake electrode, flat-base (or plate) electrodes have been used (Athanasiou et al., 2007; Tsokas et al., 2008); galvanic contact with the ground is achieved using an electrically conductive gel or mud at the base of the electrode. In IP surveys, nonpolarizable electrodes (Reynolds, 1997) are widely used as the potential electrodes to reduce SP noise. This makes it difficult to carry out IP surveys using multielectrode systems for 2-D (two-dimensional) imaging surveys. There has, however, been significant progress in using normal steel stakes as the potential electrodes (Dahlin et al., 2002; LaBrecque and Daily, 2008).

Multielectrode and multichannel systems

The development of multielectrode systems over the past 20 years has sparked a revolution in resistivity surveying. The advent of 2-D and 3-D (three-dimensional) resistivity tomography has opened up whole new application areas to electrical methods. Before early 1990s, the electrical resistivity method was mainly used in resistivity sounding, profiling, and mapping surveys (Telford et al., 1990). Quantitative interpretation was mainly confined to 1-D (one-dimensional) structure of the subsurface consisting of horizontal layers (Koefoed, 1979). The multielectrode systems made it practical to carry out 2-D imaging surveys that give a more accurate picture of the subsurface (Dahlin, 2001) in a routine manner. Unlike the conventional 4 electrodes system, a multielectrode system has about 25 or more electrodes connected to the resistivity meter via a multi-core cable (Figure 4). A switching circuitry controlled by an internal microcomputer within the resistivity meter automatically selects the appropriate 4 electrodes for each measurement. The system is usually programmable and almost any electrode array configuration can be used. Several variants of the multielectrode cable system exist, where the distinguishing factor tends to be the physical location of the switching circuitry. Some resistivity meters contain all the switching capability within the main unit, so that two or more long multi-core cables can be attached directly to the meter. The cables then merely serve as a physical extension between the switch and the electrodes, which have no other function than to act as a galvanic contact between the meter and the ground (also known as "dumb" electrodes). These systems have a practical limit on the number of electrode takeouts (commonly 32) in each cable. In the second type of system, an external switch box is connected to the main meter via a system bus. The switching capacity can then be extended by daisy-chaining switch boxes, each connecting to cable segments with multiple "dumb" electrodes. This design can vastly increase the possible number of electrodes available for simultaneous measurements, but it can also be less portable than the first arrangement, particularly in forested and rugged terrain. The third type of system uses "intelligent" electrodes, where the switching circuitry is decentralized and placed at individual electrodes. Even the actual voltage measurement and analog-to-digital conversion may be carried out locally in such systems, potentially reducing interference problems (Stummer and Maurer, 2001).



Instrumentation, Electrical Resistivity, Figure 4 Schematic diagram of a multielectrode system used for a 2-D electrical survey and an example sequence of measurements used to build up a pseudosection using the Wenner array.

A more recent development is multichannel capability for multielectrode systems that can greatly reduce the survey time. Only two electrodes can be used as the current electrodes at a single time, but the voltage can be measured between different pairs of potential electrodes. Commercial systems with 4–10 channels are available and some research systems have more than 100 channels (Stummer and Maurer, 2001).

Full-waveform systems

There has been considerable research on systems that record the full waveform of the potential signal in an effort to extract more information from the data and to improve the signal-to-noise ratio using signal processing techniques (Friedel and Jacobs, 1998; Storz et al., 2000; Matthews and Zonge, 2003; Rowston et al., 2003; Zhe et al., 2007). This is often the only practical approach for resistivity surveys in which large depths of investigation are required (such as, regional surveys or crustal studies) or where the nature of the target requires large dipole offsets (for example, investigations on volcanoes).

Towed systems for dynamic measurements

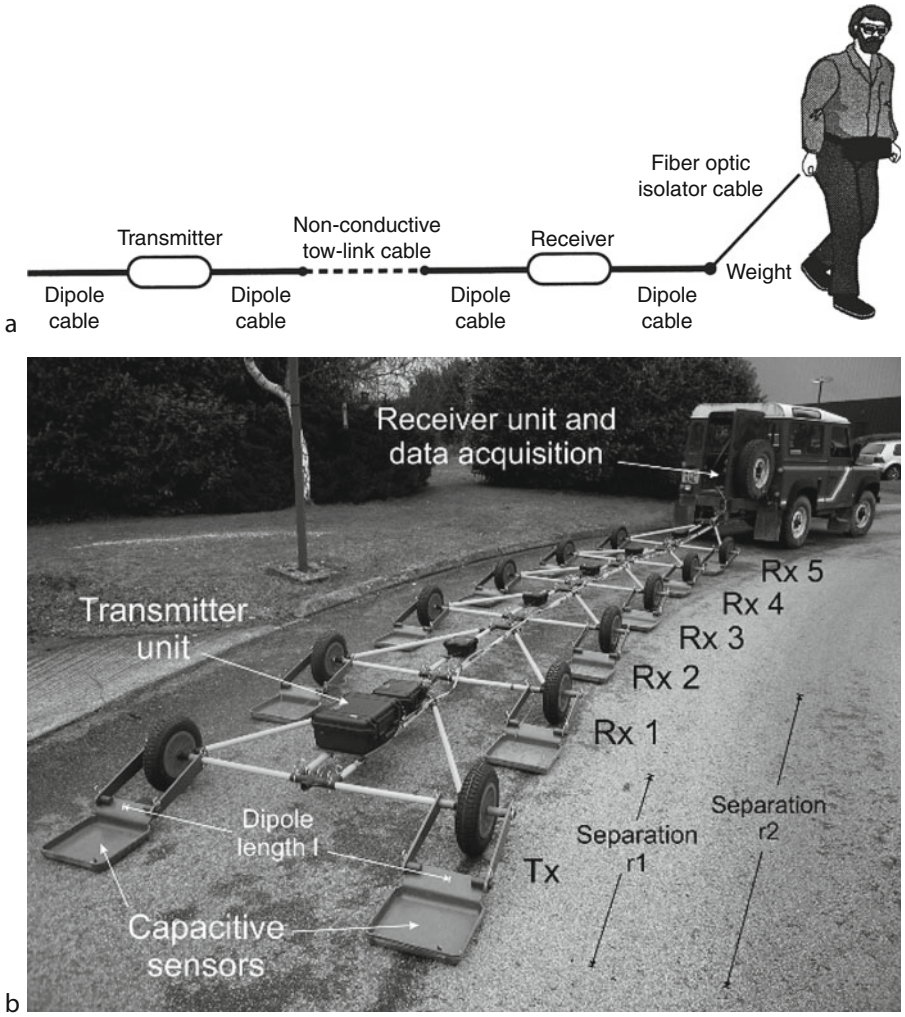
The multichannel multielectrode systems have been adapted for continuous profiling water-borne surveys using floating electrodes attached to a cable pulled by a boat (Loke and Lane, 2005; Mansoor and Slater, 2007; Goto et al., 2007). A PC coordinates the resistivity meter system data acquisition together with a GPS and water depth sounder. Continuous measuring systems for land surveys have also been developed. Some systems use cylindrical steel electrodes based on an in-line array geometry (Sørensen, 1996), while others use spiked wheels to achieve continuous galvanic contact with the soil (Panissod et al., 1998; Dabas, 2009).

Capacitively coupled systems

It is difficult or impossible to use conventional resistivity meter systems that inject a galvanic current into the ground in areas with very resistive surface materials or paved surfaces. Capacitively coupled systems can be used in such areas. These instruments use an oscillating, nongrounded electric dipole to generate current flow in the ground and a second similar dipole to measure the resulting potential distribution at the ground surface (Kuras et al., 2006). Two major configurations have been used in commercial and research instruments of this type. The first configuration (line antenna type) uses cylindrical transmitters and receivers, which are towed behind an operator (Figure 5a). It gives measurements that are comparable to the galvanically coupled in-line dipole–dipole array (Møller, 2001) and have maximum depths of investigation of 1–20 m, depending on the lengths of transmitter and receiver and the distance between them. The second type (electrostatic quadrupole) uses flat metallic conductors (Figure 5b) in an equatorial dipole–dipole configuration (Panissod et al., 1998; Kuras et al., 2007). This type has been used for survey depths of up to a few meters. Both configurations are dynamic measuring systems where the array configuration is fixed but the entire setup is moved during the survey for lateral coverage. 2-D and 3-D surveys with dense lateral coverage can be rapidly conducted with these systems (Kuras et al., 2007).

Automated resistivity monitoring systems

Automatic PC controlled monitoring systems have been developed for detecting transient phenomena such as water seepage from dams, landslides, and solute transport (Oldenborger et al., 2007; Sjödahl et al., 2008; Supper et al., 2008; Kuras et al., 2009). Monitoring systems are becoming increasingly sophisticated, and are now being



Instrumentation, Electrical Resistivity, Figure 5 Types of capacitively coupled systems. (**a**) Line antenna (courtesy of Geometrics Inc.) and (**b**) plate-wire capacitive sensors (electrostatic quadrupole) arranged in a towed array. (Courtesy of British Geological Survey, NERC.)

deployed remotely with permanently installed electrode arrays, telemetric control, and data transfer, supported by automated data management systems to handle the large data volumes generated through time-lapse data acquisition (Ogilvy et al., 2009).

Summary

There have been major advancements over the past few decades in the instrumentation for the electrical resistivity method. In addition to traditional 1-D surveys, the new multielectrode and multichannel systems have made it possible to efficiently carry out 2-D and 3-D imaging surveys that provide more accurate models of the subsurface geology. Computerized dynamic towed systems can map large areas on land and water rapidly. Capacitively coupled systems make it possible to survey areas with resistive surfaces where conventional galvanic systems cannot be used. Sophisticated PC controlled monitoring systems are now available to automatically detect and record transient phenomena.

Acknowledgments

We wish to thank Geometrics Inc, Landviser LLC, Abem Instrument AB and the British Geological Survey (NERC) for permission to use their illustrations in this paper.

Bibliography

Aristodemou, E., and Thomas-Betts, A., 2000. DC resistivity and induced polarisation investigations at a waste disposal site and its environments. *Journal of Applied Geophysics*, **44**, 275–302.

Athanasiou, E. N., Tsourlos, P. I., Vargemezis, G. N., Papazachos, C. B., and Tsokas, G. N., 2007. Non-destructive DC resistivity surveying using flat-base electrodes. *Near Surface Geophysics*, **5**, 263–272.

Chambers, J. C., Kuras, O., Meldrum, P. I., Ogilvy, R. D., and Hollands, J., 2006. Electrical resistivity tomography applied to geologic, hydrogeologic, and engineering investigations at a former waste-disposal site. *Geophysics*, **71**, B231–B239.
Dabas, M., 2009. Theory and practice of the new fast electrical imaging system ARP. In Campana, S., and Piro, S. (eds.), *Seeing the Unseen. Geophysics and Landscape Archaeology*. London: CRC Press, pp. 105–126.
Dahlin, T., 2001. The development of DC resistivity imaging techniques. *Computers and Geosciences*, **27**, 1019–1029.
Dahlin, T., Leroux, V., and Nissen, J., 2002. Measuring techniques in induced polarization imaging. *Journal of Applied Geophysics*, **50**, 279–298.
Friedel, S., and Jacobs, F., 1998. Large scale resistivity imaging at Merapi volcano. *Annales Geophysicae*, **16**(1), C180.
Goto, T., Kasaya, T., Machiyama, H., Takagi, R., Matsumoto, R., Okuda, Y., Satoh, M., Watanabe, T., Seama, N., Mikada, H., Sanada, Y., and Kinoshita, M., 2007. A marine deep-towed DC resistivity survey in a methane hydrate area, Japan Sea. *Exploration Geophysics*, **39**, 52–59.
Koefoed, O., 1979. *Geosounding Principles 1: Resistivity sounding measurements*. Amsterdam: Elsevier Science.
Kuras, O., Beamish, D., Meldrum, P. I., and Ogilvy, R. D., 2006. Fundamentals of the capacitive resistivity technique. *Geophysics*, **71**, G135–G152.
Kuras, O., Meldrum, P. I., Beamish, D., Ogilvy, R. D., and Lala, D., 2007. Capacitive resistivity imaging with towed arrays. *Journal of Environmental and Engineering Geophysics*, **12**, 267–279.
Kuras, O., Pritchard, J. D., Meldrum, P. I., Chambers, J. E., Wilkinson, P. B., Ogilvy, R. D., and Wealthall, G. P., 2009. Monitoring hydraulic processes with automated time-lapse electrical resistivity tomography (ALERT). *Comptes Rendus Geoscience*, **341**, 868–885.
LaBrecque, D., and Daily, W., 2008. Assessment of measurement errors for galvanic-resistivity electrodes of different composition. *Geophysics*, **73**, F55–F64.
Linderholm, P., Marescot, L., Loke, M. H., and Renaud, P., 2008. Cell culture imaging using microimpedance tomography. *IEEE Transactions on Biomedical Engineering*, **55**, 138–146.
Legault, J. M., Carriere, D., and Petrie, L., 2008. Synthetic model testing and distributed acquisition dc resistivity results over an unconformity uranium target from the Athabasca Basin, northern Saskatchewan. *The Leading Edge*, **27**(1), 46–51.
Loke, M. H., and Lane, J. W., Jr., 2005. Inversion of data from electrical resistivity imaging surveys in water-covered areas. *Exploration Geophysics*, **35**, 266–271.
Lu, K., and Macnae, J. C., 1998. The international campaign on intercomparison between electrodes for geoelectrical measurements. *Exploration Geophysics*, **29**, 484–488.
Matthews, P., and Zonge, K. L., 2003. 50 years – State of the art in IP and complex resistivity. *KEGS 50th Anniversary Symposium, Mining and Environmental Geophysics – Past, Present and Future*, Toronto.
Mansoor, N., and Slater, L., 2007. Aquatic electrical resistivity imaging of shallow-water wetlands. *Geophysics*, **72**, F211–F221.
Møller, I., 2001. OhmMapper field tests at sandy and clay till sites in Denmark. In *Proceedings of the EEGS-ES 7th Annual Meeting*. Environmental and Engineering Geophysical Society-European Section, pp. 100–101.
Ogilvy, R. D., Meldrum, P. I., Kuras, O., Wilkinson, P. B., Chambers, J. E., Sen, M., Pulido-Bosch, A., Gisbert, J., Jorreto, S., Frances, I., and Tsourlos, P., 2009. Automated monitoring of coastal aquifers with electrical resistivity tomography. *Near Surface Geophysics*, **7**, 367–375.
Oldenborger, G. A., Knoll, M. D., Routh, P. S., and LaBrecque, D. J., 2007. Time-lapse ERT monitoring of an injection/withdrawal experiment in a shallow unconfined aquifer. *Geophysics*, **72**, F177–F187.
Panissod, C., Michel, D., Hesse, A., Joivet, A., Tabbagh, J., and Tabbagh, A., 1998. Recent developments in shallow depth electrical and electrostatic prospecting using mobile arrays. *Geophysics*, **63**, 1542–1550.
Pozdnyakov, A. I., Rusakov, A. V., Shalaginova, S. M., and Pozdnyakova, A. D., 2009. Anisotropy of the properties of some anthropogenically transformed soils of podzolic type. *Eurasian Soil Science*, **42**, 1218–1228.
Reynolds, J. M., 1997. *An Introduction to Applied and Environmental Geophysics*. England: Wiley.
Rowston, P., Busuttil, S., and McNeill, G., 2003. Cole–cole inversion of telluric cancelled IP data. In *ASEG 16th Geophysical Conference and Exhibition Extended Abstracts*. Australian Society of Exploration Geophysicists, ab021.
Sjödahl, P., Dahlin, T., Johansson, S., and Loke, M. H., 2008. Resistivity monitoring for leakage and internal erosion detection at Hällby embankment dam. *Journal of Applied Geophysics*, **65**, 155–164.
Sørensen, K., 1996. Pulled Array Continuous Profiling. *First Break*, **14**, 85–90.
Storz, H., Storz, W., and Jacobs, F., 2000. Electrical resistivity tomography to investigate geological structures of the earth's upper crust. *Geophysical Prospecting*, **48**, 455–471.
Stummer, P., and Maurer, H. R., 2001. Real-time experimental design applied to high-resolution direct-current resistivity surveys. In *Expanded Abstracts, The International Symposium on Optical Science and Technology*, SPIE 4491, pp. 143–150.
Supper, R., Romer, A., Jochum, B., Bieber, G., and Jaritz, W., 2008. A complex geo-scientific strategy for landslide hazard mitigation – from airborne mapping to ground monitoring. *Advances in Geosciences*, **14**, 195–200.
Telford, W. M., Geldart, L. P., and Sheriff, R. E., 1990. *Applied Geophysics*, 2nd edn. Cambridge: Cambridge University Press.
Tsokas, G. N., Tsourlos, P. I., Vargemezis, G., and Novack, M., 2008. Non-destructive electrical resistivity tomography for indoor investigation: the case of Kapnikarea Church in Athens. *Archaeological Prospection*, **15**, 47–61.
White, R. M. S., Collins, S., Denne, R., Hee, R., and Brown, P., 2001. A new survey design for 3D IP modelling at Copper Hill. *Exploration Geophysics*, **32**, 152–155.
Zhe, J., Greenhalgh, S., and Marescot, L., 2007. Multichannel, full waveform and flexible electrode combination resistivity-imaging system. *Geophysics*, **72**, F57–F64.

Cross-references

Electrical Properties of Rocks
Electrical Resistivity Surveys and Data Interpretation

INSTRUMENTATION, EM

Steven Constable
Institute of Geophysics and Planetary Physics, Scripps Institution of Oceanography, La Jolla, CA, USA

Definition

CSEM Controlled source EM
EM Electromagnetic
GDS Geomagnetic depth sounding
MT Magnetotelluric
nT Nanotesla, unit of magnetic flux density

Introduction

Here we look at instrumentation used for land and marine electromagnetic (EM) induction studies of Earth, which seek to characterize the electrical conductivity of the subsurface. Conductivity is driven by porosity and fluid content in the crust, temperature and mineralogy in the mantle, and minor proportions of highly conductive graphite, oxides, or sulfides in both. The principal tool for deep Earth conductivity sounding is the *magnetotelluric* (MT) method, in which time series measurements of the surface horizontal magnetic and electric fields are recorded and converted into a frequency-dependent transfer function. Loosely speaking, variations in the external magnetic field act as the input and induced electric fields in the earth act as the output. The frequency dependence of the amplitude and phase of the transfer function may be interpreted to recover electrical conductivity as a function of depth; longer periods propagate into deep structure while higher frequencies are absorbed in shallow structure.

A related method is called *geomagnetic depth sounding* (GDS), in which a transfer function between the vertical and horizontal magnetic fields is used in a similar manner. GDS data are mainly sensitive to lateral conductivity structure, although if assumptions are made about the morphology of the external magnetic source field, long-period data can be interpreted to recover conductivity as a function of depth in a similar manner to MT data. Alternatively, gradients in the magnetic field can become a proxy for the electric field, and an MT response calculated in this way.

When near-surface crustal geology is of interest, man-made sources of EM energy (transmitters) can be used to probe conductivity, using either current passed through a grounded wire (an electric field source) or current passed through a loop (or loops) of wire (a magnetic field source). Electric or magnetic fields are measured as a function of transmitter frequency (frequency domain sounding) or, alternatively, transmitter turn-off time (time domain, or transient, sounding). Transmitter–receiver position, separation, and geometry (orientation, etc.) are also survey parameters. Measurements of this type fall into the category of controlled-source EM (CSEM) methods, although particular configurations often have distinct names or even trade names (e.g., Turam, LoTEM, MTEM).

Electric field measurement

Electrodes are used to make contact with the ground (or, in the marine case, seawater) for the purpose of making a measurement of potential difference, which is usually then divided by the electrode separation to form an estimate of electric field. Even with high input impedance amplifiers, some small current must be drawn to make the measurement. The simplest electrode is a metal rod in contact with damp earth or water, which will act as a source or sink of electrons e^- through the reaction

$$\mathrm{M} \rightleftharpoons \mathrm{M}^+ + e^-$$

where M is the metal and M^+ is the metal cation. This is called an electrode of the first kind, and unless the metal cation is highly concentrated in the vicinity of the electrode, this reaction will not be reversible and the electrode will polarize, creating unwanted galvanic potentials and contributing to noise. This may be tolerable for high-frequency measurements, and simple metal rods are used for high-frequency MT measurements using man-made radio transmitters as the source field.

If the metal rod is immersed in a concentrated aqueous solution of a compatible salt, then the reaction is reversible and the electrode is nonpolarizing. The most commonly used electrode of this type is the copper–copper sulfate "porous pot." The "pot" containing the copper sulfate solution has a semipermeable barrier as its base, usually unglazed ceramic, which makes electrolytic contact with any moisture in the soil. Such electrodes are commonly used for "broadband" MT measurements in the frequency range of 1 mHz to 100 Hz.

Nonpolarizing electrodes of the second kind are made by coating a metal with a sparingly soluble salt and immersing in an aqueous electrolyte containing a compatible anion. The reaction here is

$$\mathrm{M} + \mathrm{A}^- \rightleftharpoons \mathrm{MA} + e^-$$

where A^- is the anion. On land the most commonly used electrode of this type is lead–lead chloride immersed in solution of potassium or sodium chloride (e.g., Petiau, 2000), and for marine work silver–silver chloride immersed in seawater, which contains about 30 g/l sodium chloride, is commonly used (e.g., Webb et al., 1985). These electrodes have lower temperature coefficients and higher stability than those of the first kind, and are preferred for long-period MT work on land and almost all MT studies in the marine environment. Silver–silver chloride electrodes are also favored for marine CSEM soundings, although polarizing, capacitively coupled, electrodes such as the carbon fiber sensors of Crona et al. (2001) have also been used.

Electrodes are separated by cables of order 100 m long on land to integrate the electric fields and provide a measurable potential difference. Contact impedance for electrodes can be as large as several kilohms, and so the input impedance of amplifiers should be large, but gain need not be very high since signals will be up to several millivolts across a 100-m antenna, and a reasonably well-designed amplifier will not contribute significantly to noise.

For marine measurements the situation is very different. The electrode contact impedance is much lower (around 10 Ω), the signal levels are much smaller (a few μV/m), and the antenna lengths are usually shorter as well (around 10 m). For long-period MT simple amplifiers with gains of about 100 are adequate, but for broadband MT and CSEM measurements at frequencies of 0.1 Hz and above,

capacitively coupled chopper amplifiers with very low noise and gains of 10,000 and higher are used (Webb et al., 1985).

Magnetic field measurement

Magnetic field sensors fall into two broad categories; total field magnetometers and vector magnetometers. Total field proton and nuclear precession devices are used for magnetic mapping, but for EM induction vector measurements are needed, either using induction coils or fluxgate sensors. Superconducting (SQUID) magnetometers have been used for induction studies, but so rarely that they will not be discussed here.

A simple loop of wire can be used as a magnetic sensor based on Faraday's Law, in which a time-varying magnetic flux of density B through area A will induce a voltage V in a loop of wire with N turns:

$$V = -NA\frac{dB}{dt} = -NA\mu\frac{dH}{dt}$$

where H is the magnetic field and μ is permeability, which for an air-cored loop is the free-space value of $4\pi \times 10^{-7}$ H/m.

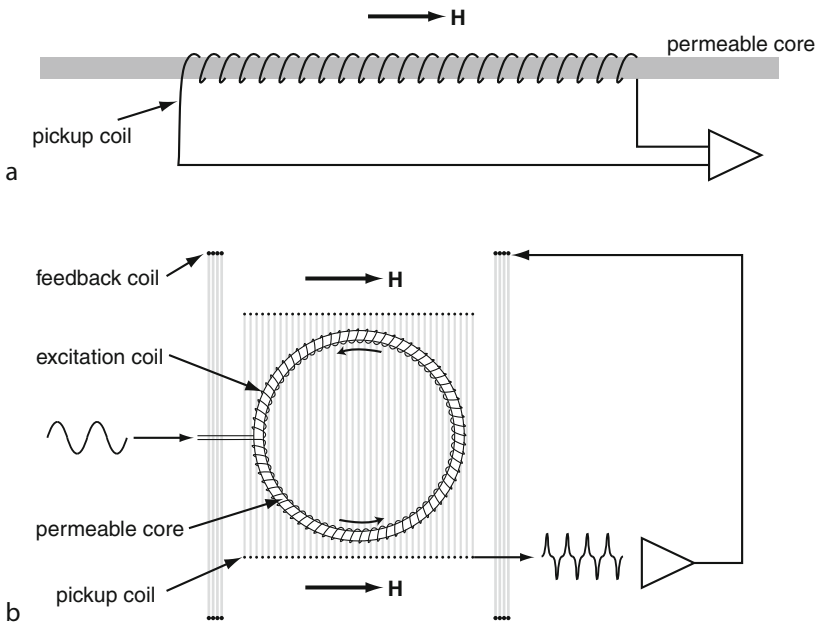
For CSEM operations at high frequencies (kilohertz range) the derivative term is large enough that an adequate signal may be measured in this way, particularly for single frequency CSEM in which lock-in amplifiers can be used. Alternatively, for time domain studies in which a large loop of wire is used as the transmitter, the area may be as much as 10^4 m^2 and it is possible to make measurements using the transmitter loop after switch-off. Normally however, induction coils (also called search coils, Figure 1a) are used, in which thousands of turns of wire are wound around a magnetically permeable core of soft ferrite or iron–nickel alloy (Permalloy or mu-metal). The core captures the magnetic field and multiplies the flux by an effective permeability, μ_e, determined by a combination of geometry and relative permeability, producing a voltage

$$V = -NA\mu\,\mu_e\frac{dH}{dt}.$$

For long, cylindrical, highly permeable cores of length L and diameter d the effective permeability depends only on geometry and is given approximately by (Tumanski, 2007):

$$\mu_e = \frac{(L/d)^2}{(\ln(2L/d) - 1)}.$$

We see that increasing length improves sensitivity, and typical broadband MT coils are of order 1 m long. The main source of noise for induction coils is the thermal (or Johnson) noise of the resistance in the wire windings, and so increasing N does not necessarily increase the signal to noise ratio (SNR) because it also increases resistance. Indeed, to a first approximation the SNR is proportional



Instrumentation, EM, Figure 1 (**a**) Schematic depicting the operation of an induction coil magnetometer. A coil of wire is wound around a permeable core. (**b**) The ring-core fluxgate magnetometer. An excitation coil drives the core into saturation in two different directions. The direction aligned with the external magnetic field will saturate sooner than the direction apposed to the field, creating a pulse of unbalanced flux twice per excitation cycle, which is detected by a synchronous amplifier tuned to twice the excitation frequency. In practice, fluxgates are often used as null sensors with the output of the pickup coil used to null the field using a feedback coil; the current in the feedback coil becomes the measurement.

to the mass of copper used in the windings. However, if weight is an issue, as it may be for marine instruments, aluminum windings can be used for a significant improvement in weight at a modest cost in resistance noise. The noise of a good-quality induction coil used for EM induction studies is around 10^{-9} nT^2/Hz at 1 Hz with a red $1/f$ type spectrum. Since motion in Earth's magnetic field can also produce a time-varying signal, on land induction coils are usually buried during operation to prevent movement. For marine MT and CSEM studies, the magnetometers must be stabilized by heavy anchors, usually made from concrete to avoid magnetic contamination.

Although integrating amplifiers or feedback can be used to flatten the frequency response of induction coils, the one-pole low-cut response of the dH/dt term is a fairly good match to the red spectrum of natural magnetic field variations. However, reasonable signal to noise performance for induction coils is still limited to periods shorter than several thousand seconds. For longer period MT and GDS studies, fluxgate magnetometers are preferred. Fluxgates use the principle that variations in magnetic flux may be generated for a fixed magnetic field by varying the permeability of the core material:

$$V = -NA\frac{dB}{dt} = -NA\mu\frac{d\mu_e}{dt}H.$$

This is achieved by winding the permeable core with an excitation coil that saturates the core in alternate directions using a sinusoidal or square wave current of around 1 kHz. A saturated core has an effective permeability of zero, since it is no longer influenced by the external field H, and so the excitation current generates the time-varying permeability required to get an inductive response.

Various core geometries can be used for fluxgates, but the most common type of geophysical sensor uses a ring-core geometry (Figure 1b). The excitation coil saturates the two sides of the core symmetrically in opposite directions; the side of the coil excited in the same direction as the external magnetic field saturates sooner, while the opposite side saturates later. This asymmetry produces a time-varying net flux at twice the excitation frequency, which is detected by a pickup coil wound around the core; the orientation of the detection coil determines the direction of sensitivity. A lock-in amplifier tuned to twice the excitation frequency will thus have an output proportional to the external field. However, this output is only linear over a range of several hundred nanotesla, and also will be sensitive to variations in μ_e caused by temperature; so most observatory quality instruments are operated as null sensors. The total magnetic field (usually the component of Earth's field aligned with the sensor) is canceled by a solenoid or Helmholtz coils. The output of the fluxgate is fed back to control the current in the nulling coils, and this current then becomes the measurement of low-frequency changes in the magnetic field.

The effective permeability of ring-core of diameter L and thickness d is approximately

$$\mu_e = \frac{L}{d}$$

for large L/d; ring-core sensors are usually about 2 cm in diameter and a millimeter or so thick, with quoted noise levels of order 10^{-4} nT^2/Hz at 0.1 Hz (Primdahl, 1979). By design, fluxgates sense the static magnetic field component in a given direction, which is typically 40 μT on Earth (possibly as large as 60 μT), and so the low-frequency response of fluxgates extends to DC. Excitation frequencies are high enough that variations up to a few tens of hertz can be recorded, but long-period MT instruments typically sample only at about 1 Hz, since the noise in fluxgates becomes higher than induction coils at a frequency of about 0.01 Hz. Also, making measurements of field variations in the presence of Earth's total field requires a large dynamic range, and even with 24-bit analog to digital conversion (ADC), the least count will be of order 0.01 nT, and resolutions of 0.1 nT are more typical. Least counts for induction coil sensors are of order 0.01 pT.

EM transmitters

Either a magnetic or electric field transmitter can be used for CSEM soundings. Electric transmitters consist of a long (hundreds of meters) wire grounded at both ends, through which a modulated current is driven. For large currents it is easier to switch square waves than generate sinusoidal waveforms, and if the switching is controlled rapidly then time domain measurements may be made. Alternatively, a swept or random sequence of binary pulses can be used and the receiver data deconvolved to generate a broadband response. For marine CSEM surveys, the high conductivity of seawater allows transmitter currents of up to a thousand amps to be used. Magnetic transmitters consist of large loops of wire through which a current is passed, and are mainly used for time domain studies on land.

Data acquisition systems

The science and technology behind EM sensors have not changed significantly for many decades, but development of modern electronics has revolutionized the way MT and CSEM data are collected. The advent of microprocessors; 24-bit ADC modules; low power, high capacity, static digital memory; and the global positioning system (GPS), which allows time to be measured with microsecond accuracy anywhere on the surface of Earth, together allow continuous, time-synchronized, measurements to be made using small, portable, low-power logging systems.

Time synchronization is important in MT data acquisition, since noise reduction in processing raw time series to frequency domain response functions depends on cross-spectra being estimated between two or more remotely separated sites. For CSEM soundings, synchronous stacking and estimation of phase lag between source and receiver both demand good timing for the transmitter and receiver. In the past this was done with wired or

wireless links between sites for both CSEM and MT. Now, all sites can be individually clocked with GPS timing for later synchronous processing, at least on land. Marine MT and CSEM receivers must have installed clocks that are reset using GPS time before deployment, and whose drift can be determined using GPS on recovery. In this way time accurate to better than a millisecond per day can be achieved. Since marine EM transmitters are attached to, and powered by, ships, GPS time can be used to control the output waveforms.

Similarly, in the past wired or wireless connections between the measurement equipment and bulky data acquisition and storage systems were needed, but now data can be stored site by site or even channel by channel, using integrated microprocessor-controlled acquisition and solid-state memory. Clearly, this has been of huge benefit to the collection of marine data, but land acquisition now also uses this approach.

Navigation

The GPS system, as its name suggests, provides highly accurate locations and makes surveying land MT and CSEM stations straightforward. For marine MT and CSEM, the positioning of seafloor receivers is more difficult, since during deployment they will drift 50–100 m from the release point as they sink down through currents in the water column. For MT surveys the release point may provide sufficient accuracy, but this is not the case for CSEM surveys. In this case accurate positions need to be derived from acoustic ranging from the survey ship, whose position is known from GPS. Short (1–10 ms), high-frequency (10–100 kHz) pulses of sound are emitted from the ship and transponders on the receiver instruments reply to these with similar "pings." The two-way travel time, along with estimates of sound speed in water derived from oceanographic tables or direct measurement, is used to calculated range. Two approaches are possible: long baseline surveys involve triangulation using ranges from three or more ship positions, and short baseline surveys use equipment that can estimate direction as well as range using multiple units or elements for the ship's transponders. For marine CSEM transmitter navigation, short baseline measurements are supplemented by depth, height, and possibly dead reckoning sensors on the transmitter.

On land the orientation of sensors is measured or set using levels and magnetic compasses during installation (fluxgate sensors can act as their own compasses). For seafloor measurements, orientation needs to be estimated using recording tiltmeters and compasses.

Summary

Recordings of natural or man-made variations in the magnetic field may be made with fluxgate sensors or induction coils, depending on the frequency of interest. In terms of signal to noise, the crossover between the two sensors is about 0.01 Hz, with induction coils used for measurements higher than 1 mHz and fluxgates for measurements lower than 1 Hz. Exact frequencies depend on the particular equipment and application. Electric field sensors are typically made using pairs of nonpolarizing electrodes to make contact with ground or water, separated and connected to amplifiers by 10–100 m of wire. At high frequencies simple metal rods may suffice as electrodes. Modern electronics and timing derived from the global positioning system have made data collection using autonomous loggers standard for both land and marine MT and CSEM data acquisition.

Bibliography

Crona, L., Fristedt, T., Lundberg, P., and Sigray, P., 2001. Field tests of a new type of graphite-fiber electrode for measuring motionally induced voltages. *Journal of Atmospheric and Oceanic Technology*, **18**, 92–99.

Petiau, G., 2000. Second generation of lead-lead chloride electrodes for geophysical applications. *Pure and Applied Geophysics*, **157**, 357–382.

Primdahl, F., 1979. The fluxgate magnetometer. *Journal of Physics E: Scientific Instrumentation*, **12**, 241–253.

Ripka, P., 1992. Review of fluxgate sensors. *Sensors and Actuators A*, **33**, 129–141.

Tumanski, S., 2007. Induction coil sensors – a review. *Measurement Science and Technology*, **18**, R31–R46.

Webb, S. C., Constable, S. C., Cox, C. S., and Deaton, T. K., 1985. A seafloor electric field instrument. *Geomagnetism and Geoelectricity*, **37**, 1115–1129.

Cross-references

Electrical Resistivity Surveys and Data Interpretation
Geomagnetic Field, Global Pattern
Geomagnetic Field, Measurement Techniques
Geomagnetic Field, Theory
Instrumentation, Electrical Resistivity
Magnetic Storms and Electromagnetic Pulsations

INTERNATIONAL GEOPHYSICAL YEAR

Ralph W. Baird
Baird Petrophysical International, Houston, TX, USA

Definition

The International Geophysical Year (IGY 1957–1958) was the most successful global effort to coordinate the measurement and collection of geophysical data from around the world during a period of anticipated maximum solar activity. It was also timed as a continuation of a series of international efforts to collect data at the poles, called the International Polar Years I and II (IPY 1882–1883 and IPY 1932–1933). The IGY actually was planned for 18 months, July 1, 1957, through December 31, 1958, and a follow-on International Geophysical Cooperation (IGC 1959) was necessary to complete the work and assemble the results, from January 1, 1959, through December 31, 1959.

During 2007–2009, there was a fourth international effort referred to as IGY-2 or IGY + 50, but instead of a coordinated singular management initiative, was organized as separate initiatives, the Electronic Geophysical Year (eGY 2007–2008), the International Polar Year (IPY 2007–2008), the International Heliophysical Year (IHY 2007–2009), and the International Year of Planet Earth (IYPE 2007–2009). To capture the work and materials from the IGY 1957–1958, organizers of the IHY created a legacy archive and recognition program entitled IGY Gold. During and after the IGY 1957–1958, the nonscience public joined the scientists with pamphlets, magazine articles, films, children's toys, games, and books, all aimed at encouraging exploration and discovery.

Introduction and concept

A lack of knowledge of natural Earth processes and the interest by thousands of Earth scientists to explore and discover relationships of the processes combined with modern communications and global transportation improvements inspired and motivated the organizers and participants of IGY in a postwar world (the early 1950s). The more measurements were studied, the more questions evolved. To get the answers, the world Earth science community needed to cooperate and collect data together and facilitate a means to share this new data to enable new Earth science discoveries.

The concept seems simple today in the twenty-first century: propose global efforts to collect physical and chemical measurements of the Earth on a semicontinuous basis over a specified period of time. In the early 1950s, this task was burdened with a lack of infrastructure and a lack of resources in all parts of the Earth, governed by separate regimes speaking hundreds of different languages, and recovering from wartime conflicts. Just getting from point A to point B was a time-consuming task. All this was about to change.

History

While in Maryland, James Van Allen (1914–2006) and his wife Abigail hosted a dinner on April 5, 1950, for British geophysicist Sidney Chapman (1888–1970), a theoretical physicist interested in the earth's magnetic phenomena and a participant in IPY II 1932–1933. Also present at dinner was Lloyd Berkner (1905–1967), a former radio-engineer who had been on Admiral Byrd's 1928–1930 Antarctic expedition. According to Van Allen, the dinner conversation ranged widely over geophysics and especially geomagnetism and ionospheric physics. Following dinner, as they were all sipping brandy in the living room, Berkner turned to Chapman and said, "Sydney, don't you think it is about time for another international polar year?" Chapman immediately embraced the suggestion, remarking that he had been thinking along the same lines himself (Van Allen, 1998).

The time was ripe. Technological improvements in instrumentation and rocketry had enabled scientists to probe much deeper into the atmosphere and deep into the Earth. In the process of enlisting support among the international scientific societies, Chapman and Berkner found a strong preference for a global program encompassing additional geographical regions and additional physical science disciplines.

Chapman first presented the idea for a third IPY to the constituent scientific unions under the International Council of Scientific Unions (ICSU). The unions, in turn, presented the proposal to the ICSU General Assembly, and ICSU, in turn, invited the World Meteorological Organization (WMO) to participate as well as the national organizations adhering to ICSU. By 1953, there were 26 countries signed up in what came to be known at the International Geophysical Year 1957–1958. The disciplines included practically all the earth, atmosphere, and oceanic sciences, covering many parts of the globe beyond the polar regions.

Scope

By 1954, the international IGY organizing committee (set up by ICSU in 1952) was known as CSAGI after its French name, Comité Special de l'Année Géophysique Internationale. The CSAGI set the priorities for IGY projects to have at least one of these characteristics:

- Problems requiring concurrent synoptic observations at many points involving cooperative observations by many nations
- Problems in the geophysical sciences whose solutions would be aided by the availability of synoptic or other concentrated work during the IGY
- Observations of all major geophysical phenomena in relatively inaccessible regions of the Earth that can be occupied during the IGY because of extraordinary effort during that interval (the Arctic and Antarctic)
- Epochal observations of slowly varying terrestrial phenomena

These were not arbitrary or unreasonable criteria. Based on this defined planning and framework, the scope of the IGY program materialized and an organization including field operations, data collection synchronization, data reporting, and assembly and archiving was begun.

Operations

When comparing the organization of IGY to today's professional and scientific working groups, the IGY was controlled by two separate bodies, the CSAGI and the ICSU. Areas of science emphasis covered:

Meteorology	Cosmic Rays	Gravity
Geomagnetism	Glaciology	Nuclear Radiation
Aurora and Air Glow	Oceanography	Latitude and Longitude
Ionosphere	Rockets and Satellites	World Days and Communication
Solar Activity	Seismology	

Participating countries represented in expeditions or contributing and sponsoring in-country data collection included:

Afghanistan	German Democratic Republic	Norway
Argentina	German Federal Republic	Pakistan
Australia	Ghana	Panama
Austria	Greece	Peru
Belgium	Guatemala	Philippines
Bolivia	Haiti	Poland
Brazil	Hawaii	Portugal
Bulgaria	Honduras	Rhodesia and Nyasaland
Burma	Hungary	Romania
Canada	Iceland	San Salvador
Ceylon	India	Saudi Arabia
Chile	Indonesia	Spain
China (Nationalist)	Iran	Sudan
China (People's Republic)	Ireland	Sweden
Colombia	Israel	Switzerland
Costa Rica	Italy	Thailand
Cuba	Japan	Tunisia
Czechoslovakia	Korea, Democratic Republic of	Turkey
Denmark	Libya	Union of South Africa
Dominican Republic	Malaya	Union of Soviet Socialist Republics
East Africa	Mexico	United Kingdom
Ecuador	Mongolia, People's Republic of	United States of America
Egypt	Morocco	Uruguay
Ethiopia	Netherlands	Venezuela
Finland	New Zealand	Vietnam, Democratic Republic of
France	Nicaragua	Vietnam, Republic of
		Yugoslavia

Achievements

General achievements of IGY were:

- Global cooperation for improved Earth physical and chemical measurements
- Improved awareness and understanding of Earth processes
- Coordinated collection and assembly of multidisciplinary data
- Archive and sustainable safekeeping of data collected and shared
- Inspiration and hope for future similar programs and initiatives

Some of the most significant IGY achievements were:

- Defining the system of mid-ocean ridges that encircle the globe, furthering our understanding of the Earth's crust and the theory of Plate Tectonics.
- Discovery of the Van Allen radiation belts. These belts surround the Earth at altitudes of hundreds and at thousands of kilometers above the surface and are significant to present-day electronic communications.
- Collection of synoptic data, a comprehensive overview of global physical phenomena. These achievements were accomplished through organization of various scientific fields under the International Council of Scientific Unions (ICSU). This union created a series of technical panels with scientific goals and facilitated international cooperation.
- Under the collection of synoptic data, special attention was given to the Antarctic Continent. Neither the race for the South Pole in the early 1900s nor the age of exploration in the 1930s brought the influx of humanity experienced during the IGY to the ice-covered continent.
- A new value for total abundance of water in the form of ice on the Antarctic continent. Ninety percent of the planet's ice is found on and around the continent, locking 68% of the world's fresh water in the Southern Hemisphere.
- Improved meteorological predictions by understanding the weather patterns of the Southern Hemisphere.
- Advancements in the theoretical analysis of glaciers.
- Seismology of the Southern Hemisphere.
- The scientific cooperation in Antarctica paved the way for the Antarctic Treaty. The treaty signed December 1, 1959, created a continent free from nuclear weapons and open to scientific research; the first truly international territory.
- Improved science and math education. Through implied competition among countries, IGY generated a new sense of the importance of math and science to competitive problem solving.
- Sputnik and satellite measurements. IGY provided the first peaceful use of previously military equipment and technology to enhance the measurement capability and later communications of people on Earth.
- World Days and Communication. During the IGY, a calendar was arranged to achieve simultaneous observations in most disciplines. During some periods, intensification in observations was considered and the World Days program was established. There were three classes of special days: Regular World Days (RWD) and World Meteorological Intervals (WMI) were picked in advance on the calendar. Special World Intervals (SWI) were designed day to day and broadcast by the World Warning agency (AGIWARN).
- World Data Centers. The IGY was predicated on full and open data exchange. The World Data Centers were created to provide equitable access for use by all qualified scientists for public good and for geophysical research as a tool for sustainability.
- World Gravity Map. Over 60,000 observations and additional 150,000 anomaly values results from the various IGY programs. The raw data were prepared for storage and further analysis on punched paper cards

for machine analysis. The final values was published as a monograph and supplied to the World Data Centers.
- Postage stamps, films, songs, and pamphlets. Outside of the scientists contributing, the new approach included outreach and generational attraction to science at all levels. These creative media and documentary projects contributed to the events during the IGY and continue through today to provide accounts and anecdotes about the progress made by participants in IGY.

Bibliography

Baird, R. W., and Hall, J. D., 2007. Introduction to this special section: eGY. *The Leading Edge*, **26**(10), 1290–1325.

Berkner, L. V., 1955. Letter to H.H. Hess. October 27. Box 4, Institute of Theoretical Physics Folder. In: *E.H. Vestine Papers*. Washington, DC: Library of Congress, Manuscripts Division.

Chapman, S., 1959. *IGY: Year of Discovery*. Ann Arbor: University of Michigan Press.

Chapman, S., 1960. From polar years to geophysical year. *Studia Geophysica et Geodaetica*, **4**, 313–324.

Dickson, P., 2001. *Sputnik: The Shock of the Century*. New York: Walker.

IGY Special Section, 1957. *Junior Scholastic*, **41**(7), 7–15. New York: Junior Scholastic.

International Council of Scientific Unions, 1957–1970. *Annals of the International Geophysical Year*. London/New York: Pergamon.

International Council of Scientific Unions, 1959. *Annals of the International Geophysical Year*. New York: Pergamon, Vol. 1.

Kelly, W., 1957. *G.O. Fizzickle*. New York: Simon and Schuster.

Korsmo, F. L., 2004. Shaping up planet earth: the International Geophysical Year (1957–1958) and communicating science through print and film media. *Science Communication*, **26**(2), 162–187.

Korsmo, F. L., 2009. The policy process and the International Geophysical Year, 1957–1958. In: *Smithsonian at the Poles: Contributions to the International Polar Year Science*. Washington, DC: Smithsonian Institution Scholarly Press, pp 23–34.

Korsmo, F. L., 2010. The origins and principles of the World Data Center System. *Data Science Journal*, **8**, IGY55–IGY65.

Marshack, A., 1958. *The World in Space: The Story of the International Geophysical Year*. New York: Thomas, Nelson.

National Academy of Sciences, 1958. *Planet Earth: The Mystery with 100,000 Clues*. Washington, DC: NAS.

National Academy of Sciences, 1965. *Report on the U.S. Program for the International Geophysical Year*. Washington, DC: NAS.

Nicolet, M., 1984. The International Geophysical Year (1957–1958): great achievements and minor obstacles. *GeoJournal*, **8**(4), 303–320.

Odishaw, H., and Ruttenberg, S. (eds.), 1958. *Geophysics and the IGY: Proceedings of the Symposium at the Opening of the International Geophysical Year*. Washington, DC: American Geophysical Union.

Schanche, D., 1957. Space beyond Sputnik lies within our grasp. *Life Magazine*, **43**(17), 26–29.

Sputnik Mania, 2007. DVD. Directed by David Hoffman. Santa Cruz: Varied Directions.

Sullivan, W., 1961. *Assault on the Unknown: The International Geophysical Year*. New York: McGraw-Hill.

Thompson, B., 2007. Preserving IGY's history and legacy. *The Leading Edge*, **26**(10), 1296–1297.

Van Allen, J. A., 1997. *Interview with B. Shoemaker, November 18, 1997. Transcript, Oral History Collection*. Columbus: Ohio State University.

Van Allen, J. A., 1998. Genesis of the International Geophysical Year. *The Polar Times*, **2**(11), 5.

Wilson, J. T., 1960. *I.G.Y.: The Year of the New Moons*. New York: Alfred A. Knopf.

Cross-references

Electronic Geophysical Year
International Polar Year 2007–2008
International Year of Planet Earth

INTERNATIONAL GRAVITY FORMULA

Hans-Jürgen Götze
Abteilung Geophysik, Institut für Geowissenschaften, Christian-Albrechts-Universität zu Kiel, Kiel, Germany

As a first approximation the Earth is a rotating sphere. As a second approximation it can be regarded as an equipotential ellipsoid of revolution. According to Moritz (1980), the theory of this equipotential ellipsoid was first given by P. Pizzetti in 1894. It was further elaborated by C. Somigliana in 1929 and served already as the basis for the International Gravity Formula adopted at the General Assembly of the International Union of Geodesy and Geophysics (IUGG) in Stockholm in 1930.

One particular ellipsoid of revolution, also called the "normal Earth" or "normal spheroid," is the one having the same angular velocity and the same mass as the actual Earth, the potential U_0 on the ellipsoid surface equal to the potential W_0 on the geoid, and the center coincident with the center of mass of the Earth. The Geodetic Reference System 1967 (GRS 67), Geodetic Reference System 1980 (GRS 80), and World Geodetic System 1984 (WGS 84) all are "normal Earth" models.

The normal or theoretical gravity is the gravity effect due to an equipotential ellipsoid of revolution. Approximate formulas are used widely even though we can calculate the exact theoretical gravity analytically. The conventionally used International Gravity Formula is obtained by substituting the parameters of the relevant reference ellipsoid. Helmert's 1901 Gravity Formula and International Gravity Formulas 1930, 1967, and 1980, correspond respectively to the Helmert 1906, International 1924, GRS 67, and GRS 80 ellipsoids. For example, the recent 1980 International Gravity Formula is (Moritz, 1980):

$$\gamma_{1980} = 978,032.7\left(1 + 0.0053024\,\sin^2\phi - 0.0000058\,\sin^2 2\phi\right)10^{-5}\ \mathrm{m/s^2}$$

where ϕ is the geodetic latitude.

The resulting difference between the 1980 International Gravity Formula and the 1930 International Gravity Formula is:

$$\gamma_{1980} - \gamma_{1930} = -16.3 + 13.7\sin\phi^2 \quad 10^{-5}\ \mathrm{m/s^2}.$$

Similarly, one can compare the 1967 formula to the 1980 formula in use today. The difference between the two is relatively small:

$$\gamma_{1980} - \gamma_{1967} = 0.8316 + 0.0782 \sin \phi^2 \quad 10^{-5}\ \mathrm{m/s^2}.$$

The first term of the International Gravity Formula is the value of gravity at the equator on the ellipsoid surface. Unfortunately, in the 1930s no one really knew what it was. The most reliable estimate at that time was based on absolute gravity measurements made by pendulums at the Geodetic Institute Potsdam in 1906. The Potsdam gravity value served as an absolute datum for worldwide gravity networks from 1909 until 1971. In the 1960s, new measurements across continents were made by precise absolute and relative gravity meters, and became the network of IGSN71 still in use today. A mean difference between the Potsdam datum and the IGSN71 reference has been found to be 14×10^{-5} m/s^2 (Woollard, 1979).

The theoretical gravity is the gravity effect due to an equipotential ellipsoid of revolution. Approximate formulas are used widely. In fact, the theoretical gravity at any position on, above, or below the ellipsoid surface can be calculated using closed-form expressions also. Such a closed-form expression for the theoretical gravity on the surface of the ellipsoid is given by the formula of Somigliana (Heiskanen and Moritz, 1967, p. 76).

The Somigliana–Pizetti formula gives the magnitude of normal gravity on the surface of a geocentric reference ellipsoid:

$$\gamma = \gamma_e \frac{1 + k \sin^2 \phi}{\sqrt{1 - e^2 \sin^2 \phi}}$$

where k is the normal gravity constant (not to be confused with the universal gravitational constant, G), γ_e is the normal gravitational acceleration at the equator and e^2 is the square of the first numerical eccentricity of the ellipsoid. Importantly, the geocentric latitude N, which is compatible with the GRS80, must be used in the Somiliana formula (Götze and Li, 2001; Hackney and Featherstone, 2003).

Parameters used to compute the normal gravity using the Somigliana–Pizetti formula are given values for the GRS80 ellipsoid from Moritz (1980) (Table 1).

International Gravity Formula, Table 1 Parameter definition and GRS80 values

a	Ellipsoid semi-major axis	6,378,137 m
b	Ellipsoid semi-minor axis	6,356,752.3141 m
γ_e	Equatorial normal gravity	9.780 326 771 5 m s^{-2}
γ_p	Polar normal gravity	9.832 186 3685 m s^{-2}
k	$(b\gamma_p/a\gamma_e) - 1$	0.001 931 851 353
e^2	$(a^2 - b^2)/a^2$	0.006 694 380 022 90
f	$(a - b)/a$	0.003 352 810 681 18
m	$(\omega^2 a^2 b)/\mathrm{GM}$	0.003 449 786 003 08
ω	Angular velocity	7292 115 $\times$ 10^{-11} rad s^{-1}
GM	Geocentric gravitational constant	3,986,005 $\times$ 108 m^3 s^{-2}

The normal gravity at any ellipsoid height h and any geodetic latitude N can also be given by a closed-form formula. Starting from the general formula of Heiskanen and Moritz (1967, p. 67–71) Lakshmanan (1991) derived the formula and published a result containing typographic errors. Later the derivation was repeated and corrected (Li and Götze, 2001).

Bibliography

Hackney, R. I., and Featherstone, W. E., 2003. Geodetic versus geophysical perspectives of the "gravity anomaly". *Geophysical Journal International*, **154**, 35–43.

Heiskanen, W. A., and Moritz, H., 1967. *Physical Geodesy.* San Francisco: W. H. Freeman.

Lakshmanan, J., 1991. The generalized gravity anomaly: endoscopic microgravity. *Geophysics*, **56**, 712–723.

Li, X., and Götze, H.-J., 1996. Effects of topography and geoid on gravity anomalies in mountainous areas: the central Andes as an example. Institut für Geologie, Geophysik und Geoinformatik, Freie Universität Berlin.

Li, X., and Götze, H.-J., 2001. Ellipsoid, geoid, gravity, geodesy and geophysics. *Geophysics*, **66**, 1660–1668.

Moritz, H., 1980. Geodetic reference system. *Bulletin Géodésique*, **54**, 395–405.

Woollard, G. P., 1979. The new gravity system – changes in international gravity base values and anomaly values. *Geophysics*, **44**, 1352–1366.

Cross-references

Gravity Anomalies, Interpretation
Gravity Field of the Earth
Gravity Method, Airborne
Gravity, Data to Anomalies

INTERNATIONAL POLAR YEAR 2007–2008

David J. Carlson
British Antarctic Survey, IPY International Programme Office, Cambridge, UK

Concept

The International Polar Year 2007–2008 (IPY) was an intensive, internationally coordinated scientific research campaign in the Arctic and the Antarctic sponsored by the International Council for Science (ICSU) and the World Meteorological Organization (WMO). Working together through international projects, IPY researchers examined all aspects of polar regions. IPY highlighted the global importance of polar processes and the urgent need to understand and track extremely rapid changes occurring at high latitudes.

History

By most accounts, the 2007–2008 IPY represented the fourth IPY: 1882–1883, 1932–1933, 1957–1958 (that event grew from a Polar Year into a Geophysical Year – the IGY), and 2007–2008. Although topics and

technologies have changed over 125 years, the challenges and difficulties of working in polar regions have not, and the most recent IPY still adheres strongly to the original concept that substantial advances in polar science require, periodically, extraordinary levels of enhanced and focused international cooperation and coordination. With news of sea ice retreat, ice sheet collapse and threats to polar species almost constantly in media headlines, the scientific need for this IPY became obvious.

Scope

The IPY Framework Document (Rapley et al., 2004) specified four goals:

- Make major advances in polar knowledge and understanding.
- Leave a legacy of new or enhanced observational systems, facilities, and infrastructure.
- Inspire a new generation of polar scientists and engineers.
- Elicit keen interest and participation from polar residents, schoolchildren, the general public, and decision-makers worldwide.

IPY emerged as one of the most ambitious international research efforts in the past 50 years. It engaged the intellectual resources of thousands of scientists – many more than expected and often from nonpolar countries – representing an unprecedented breadth of specialties, from geophysical to biological to social sciences. IPY stimulated international and interdisciplinary collaboration with over 160 endorsed science projects assembled from the ideas of researchers in more than 60 countries. IPY stimulated substantial new funding for polar science – more than US$ 400 million, which coordinated with and supplemented ongoing polar research and monitoring programs.

Achievements

Major advances in polar knowledge and understanding

Major IPY-stimulated advances in polar science will emerge in the time period 2009–2015 and beyond, as IPY researchers explore, analyze, evaluate, and share their data sets. Already, we can see increased and more pervasive evidence of change in ice sheets and sea ice; we can see and foresee alteration, migration, and depletion of polar ecosystems; we can see hints of large changes in crucial polar components (sinks and sources) of global carbon cycles; and we can see the early impacts of relentless environmental, economic, and social changes on polar communities. We can predict with confidence, as a major advance in understanding, the increasing recognition of the global importance and impact of polar change.

Infrastructure legacy

This IPY has done a superior job of observing both polar oceans, through multiple ship voyages and extensive use of buoys, drifters, and autonomous gliders. It has produced, from spaceborne radars, lidars, and gravity sensors, unprecedented coverage and views of the polar ice masses. Four separate multinational teams made difficult traverses across Antarctica, covering some regions for the first time since IGY. Field stations around the Arctic upgraded their capabilities, established new networks, and supported twice and in some cases thrice as many researchers as in previous years.

A new generation

A new Association of Polar Early Career Scientists (http://www.apecs.is/) has brought vibrant and critical energy and ideas to this IPY and to polar science. Now incorporated, internationally endorsed and engaged in all aspects of polar research, and with their own funded Directorate, APECS represents one of IPY's most significant accomplishments. With more than 1,400 members from more than 40 countries, with the ability to do online posters and career development workshops almost anywhere in the world, APECS has clearly succeeded in shaping the future of polar research.

Excite and stimulate the public

Through international collaboration and cooperation, by careful cultivation of a global community of enthusiastic volunteers, and by creative use of free technologies, this IPY stimulated active and engaged networks of hundreds of teachers, media officers, journalists, and students. Working with those partners and networks, quarterly Polar Days were organized that engaged more than 500 individual and institutional partners from 50 countries in polar activities. Teacher's workshops, summer schools, polar science weekends, and student expeditions in both hemispheres were conducted. A Polar Resource Book documenting these and many other IPY activities was published. The polar community was expanded by leaving a legacy of polar information in many languages. This IPY has demonstrated that an effective, energetic, and enthusiastic team working full-time to promote science communication serves as the key partner in a science program's success.

Challenges ahead

The variety, breadth, and volume of IPY data highlight the fact that international science data infrastructure, largely a legacy of IGY, has for the most part not kept pace with the modern expectations and needs for data access and data preservation. A few national and disciplinary data systems demonstrate admirable capabilities, and new approaches, focused on open access and an information commons approach, such as the Polar Information Commons (http://www.polarcommons.org/), may stimulate a data transformation. Identifying and sustaining the essential elements of long-term observation and monitoring systems for polar regions, as those regions undergo rapid change, will represent a severe challenge, one

without obvious technical or resource solutions. A more urgent focus on integrated prediction skills and systems will provide improved structure and rationale for maintaining polar observational capabilities. The many science specialties in this IPY, and the wide variety of data, will make the task of synthesis extremely difficult, but they will also stimulate innovative integration processes and result in products of enormous impact and value.

Summary

Often the science community seeks to organize prominent "above and beyond" events, events that transcend business as usual and accelerate scientific processes. Presented as "Experiments" or "Observation Periods" or "Years," these events share ambitious goals of enhanced research, increased resources, heightened attention and, not coincidently, re-invigorated recruitment of new researchers. The IPY 2007–2008 succeeded in stimulating international and interdisciplinary cooperation. Through widespread international cooperation and extraordinary enthusiasm, it achieved (in large part) its research, resource, recruitment, and public attention goals, and will produce a positive legacy for participants, for science in general, and for the public.

Bibliography

Rapley, C., Bell, R., and ICSU IPY 2007–2008 Planning Group, 2004. *A Framework for the International Polar Year 2007–2008*. Paris: ICSU.

Cross-references

Electronic Geophysical Year
International Geophysical Year
International Year of Planet Earth

INTERNATIONAL YEAR OF PLANET EARTH

Eduardo F. J. de Mulder[1], Wolfgang Eder[2]
[1]NGU, Geological Survey of Norway, Trondheim, Norway
[2]Department of Earth Sciences, University of Munich, Munich, Germany

Definition

The "International Year of Planet Earth" (IYPE) was a joint initiative by the International Union of Geological Sciences (IUGS) and UNESCO and ran from 2007 to 2009.

Background and activities

When, in the year 2000, the International Union of Geological Sciences (IUGS) tried to match current scientific knowledge about Planet Earth against the application of such knowledge in daily life, it found a major discrepancy. The Indian Ocean tsunami in 2004 and hurricane Katrina in 2005 only confirmed the existence of that gap; knowledge about natural hazards existed, but it was not applied – or not applied adequately – to protect vulnerable societies from devastation.

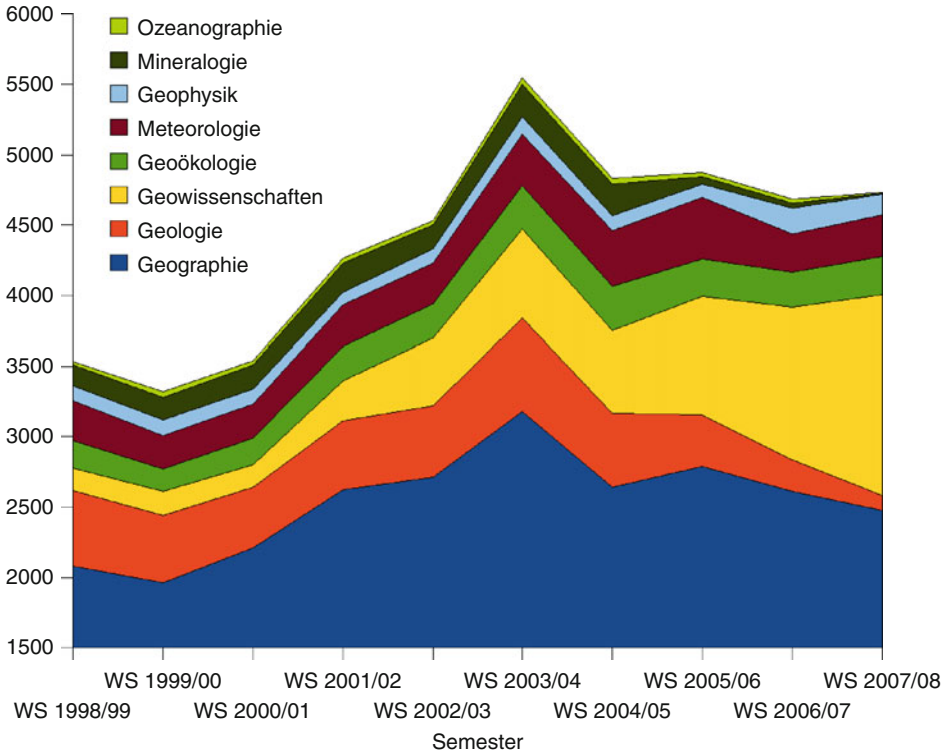
Another observation was that fewer and fewer future experts in geoscience had been recruited at universities since the early 1980s, despite the fact that society needs increasing numbers of them to help reduce the impact of such disasters and to assist governments in subsurface management issues. Today, industry needs more geo-experts to find the new earth resources for which the world is desperately searching. Those experts are the same ones that find innovative solutions to reducing the environmental impact of such necessary resource exploitation.

These observations motivated us to explore how the public and politicians could be urged to reverse these trends. We chose an international, rather than national, approach and began the long road toward the proclamation of an International Year of Planet Earth (IYPE) under the UN system. IYPE aims to capture people's imagination with the exciting knowledge about our planet, and to see that knowledge used more effectively to make the Earth a safer, healthier, and wealthier place for our children and grandchildren. That, we believe, can be achieved by building on the knowledge accumulated by the world's 400,000 Earth scientists, and is expressed by the IYPE subtitle: *Earth Science for Society.*

Two earlier events served as models. The *International Geophysical Year* (1957–1958) uncovered many facts about our planet, which was then seen for the first time from space. IGY attracted lots of public attention, and many youngsters were inspired to begin professional careers in Earth sciences. More recently, in 2002, the Germans celebrated their "Jahr der Geowissenschaften" with many public events and with an attractive logo that was later donated to the IYPE. That public exposure resulted in a much higher enrollment of German students in the Earth sciences in subsequent years (Figure 1). That information fuelled the outreach program of the International Year.

Soon after the inception of IYPE, UNESCO's (former) Earth Science Division joined in, followed by 25 Associate Partners and 11 Founding Partners. To achieve its aims and ambitions, and to collect public and political support for this issue, we then approached the United Nations system. That process went through various stages, beginning at UNESCO where the United Republic of Tanzania successfully launched the initiative in the Executive Board in April 2005, followed by the General Conference in October. In December 2005, the General Assembly of the United Nations adopted Resolution 192 proclaiming 2008 to be the *International Year of Planet Earth* (United Nations General Assembly, 2005).

Shortly afterward, the IYPE Corporation was registered as a not-for-profit 501 (c) (3) Corporation in the USA, consisting of a Board of Directors and a Secretariat based at the Norwegian Geological Survey in Trondheim. Although the International Year was proclaimed by the



International Year of Planet Earth, Figure 1 Student enrolment in the earth sciences in German universities between 1999 and 2007. Note the spike following the successful "Jahr der Geowissenschaften" (2002).

UN for 2008, its initiators expanded that into a Triennium, starting in early 2007 and closing in June 2010. A Triennium was considered the minimum needed to implement at least part of our ambitious plan.

After the UN proclamation, additional support for the IYPE grew rapidly. By December 2009, 36 International Partners were providing financial and other support. Simultaneously, geoscientific communities in individual countries began to establish IYPE National Committees (Figure 2) to a total number of 80 National and Regional Committees by the end of the IYPE Triennium. These countries are widely distributed across the globe.

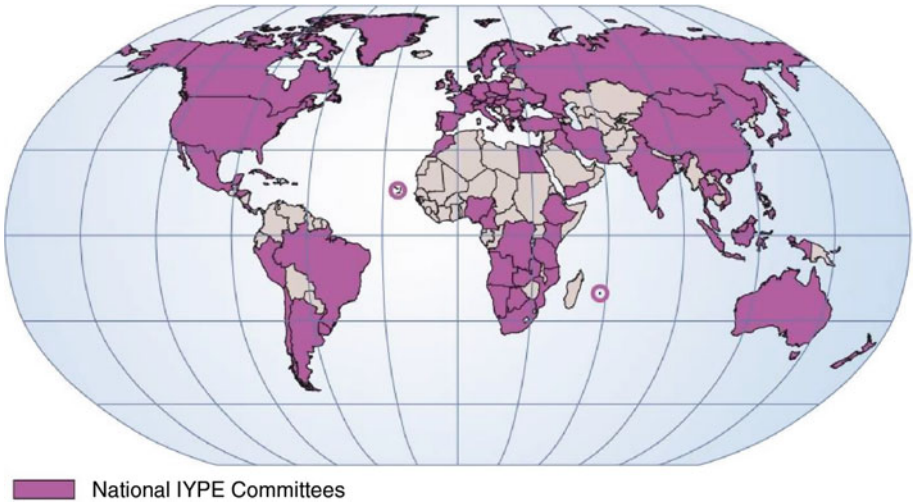
The main activities of the IYPE have been coordinated through Science and Outreach Programs. Both programs essentially operated in a "bottom-up" mode. The Science Program consisted of 10 broad, societal relevant, and multidisciplinary themes: health, climate, groundwater, ocean, soils, deep Earth, megacities, resources, hazards, and life. Brochures on each of these themes are available in hard copy or may be downloaded from IYPE's Web site: www.yearofplanetearth.org.

The Outreach Program has mainly been implemented at the national level. A global launch event was organized at UNESCO headquarters in Paris in February 2008. In addition, continent-wide launch events took place for Africa (in Arusha, Tanzania) and for Latin America (in Brasilia, Brazil). Finally, a closing event with perspectives for the future was organized in Lisbon, Portugal, in November 2009. On an international level, brochures and flyers were produced and the global IYPE web-portal was maintained. Moreover, numerous activities in many countries with IYPE National Committees were monitored and registered.

Education has been an essential element in the International Year of Planet Earth. Many of the national and international activities involved students and focused on participation by secondary and primary school pupils. An example of student involvement was the IYPE global launch event in Paris. In preparation for that event, 18–22-year-old students from around the world submitted essays and poems each dealing with one International Year theme. National Committees and the IYPE Secretariat evaluated them and selected the best. Through national and international sponsorship, some 200 winning students from almost every corner of the planet were awarded a free trip to Paris to participate actively in the launch event, including interacting with VIPs on pressing issues concerning Planet Earth. A comparable approach was taken for the Lisbon event, in 2009.

Legacy

The 80 National and Regional IYPE Committees belong to the most important legacy items of the IYPE. At



International Year of Planet Earth, Figure 2 Countries with an established national committee for IYPE (in *dark purple*). One regional committee for IYPE has been established for East and Southeast Asia.

national levels, they united key players of several, sometimes competing organizations into one joint ambition to raise awareness for the Earth sciences among decision makers and the public. An evaluation among the National Committees proved that this ambition was at least partly accomplished.

Since the onset of the IYPE Triennium in 2007, many thousands of activities all over the globe have taken place. A significant majority of these were related to education. Thousands of schools worldwide paid attention to the International Year of Planet Earth, each in their own languages and cultures.

Another IYPE legacy item was the creation of a Young Earth Science Initiative (YES). This provides a platform for young professionals in the Earth sciences. It started in 2007 and grew rapidly at the Global Launch Event of the IYPE in Paris (2008). From there the YES Initiative expanded, leading to a formal structure, a network of supporting organizations (including IYPE) and an invitation by the Chinese government hosting the first International YES Conference in October 2009 in Beijing (Figure 3). One month later, another gathering of some 200 students and young professionals from 70 countries took place in Lisbon with representation of YES.

At a political level, a flood of public statements were given by leading politicians pleading for implementation of the aims and ambitions of the International Year of Planet Earth: to base decision making more on science and knowledge of the Earth for safer, healthier, and more prosperous societies. The most common model for that has been the organization of a national launch event where politicians spoke in support of the IYPE often coupled with cultural performances by school kids highlighting the importance of the Earth and sustainable development for future generations.

OneGeology is another major IYPE legacy item. That initiative was positioned under the IYPE banner in 2007. The ambition of this project is to bring together geological data of all nations into a digital database and translate these into one single computer language. That should result into a digital map of the subsurface of the world and may eventually provide a third dimension to popular geographical Web sites, as Google Earth. By February 2010, 116 national Geological Surveys have embarked on this initiative.

Many more legacy items developed during the IYPE Triennium. For example, cross-country geological hiking tours through the Alps: ViaGeoAlpina; an elegant method to assist teachers with simple materials to explain often complex geological processes: EarthLearningIdea; the production of a series of state-of-the art scientific books on each of the ten IYPE Themes, by Springer Verlag; and an UNESCO initiative to measure the conditions of Earth sciences education in Africa and to develop tools to improve these.

It appears that student entries in the Earth sciences have been growing in a considerable number of countries since 2007. Although not all such positive developments may be attributed to the IYPE alone, the International Year of Planet Earth must have contributed at least to some extent to growth as growth is generally positively correlated with the level of IYPE activities in a specific nation.

Finally, a Report on the entire IYPE operation will be produced for the United Nations and the IYPE Initiators (IUGS and UNESCO) as a legacy item in 2010.

Perspectives

Spurred by the National and Regional IYPE Committees who considered the IYPE as an excellent opportunity to

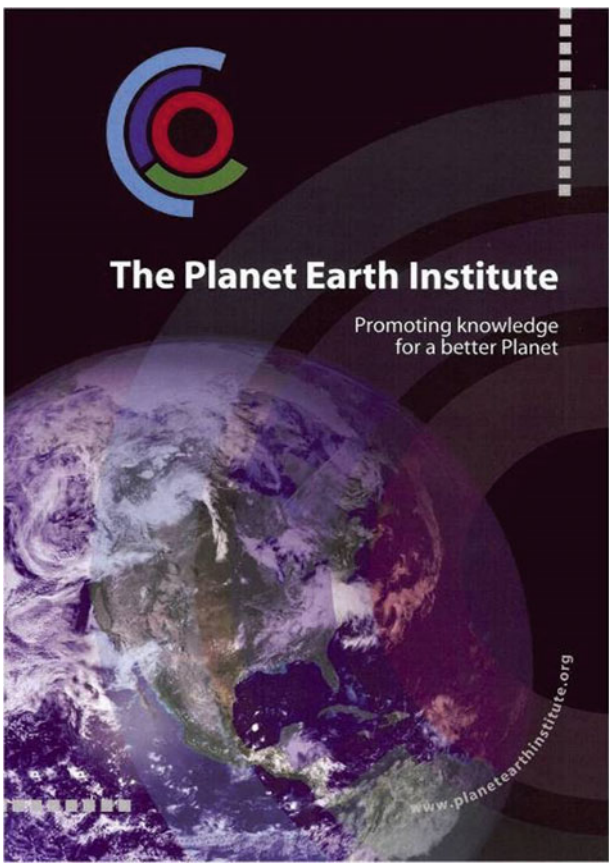
International Year of Planet Earth, Figure 3 The first international conference for Young Earth Scientists (YES) was held in Beijing, China, in October 2009.

raise public awareness to use knowledge of the Earth for better decision making and to encourage the youth to go for a professional career in the Earth sciences, and supported by the IYPE Board of Directors, some new initiatives have been discussed since early 2009. Two major new initiatives developed with the specific intention to build on the IYPE legacy are described here.

The Planet Earth Institute

The Planet Earth Institute's mission is to communicate with the general public the added value of Earth science information for society and is an international, non-governmental, non-profit organization. This (mainly virtual) Institute has been given as subtitle "Promoting knowledge for a better Planet." Implementation will be realized along three channels: through a wide variety of outreach activities, through its National Committees, and through its International Partners. Activities may include exposure through big screens in city centers, other mass media, publications, public events, education projects, cultural and artistic expressions, co-marketing, and projects (Figure 4).

National Committees shall constitute, too, the geographic backbone of the Planet Earth Institute. Current IYPE National Committees are invited to embark on the Planet Earth Institute to continue their (mainly outreach) activities. They will significantly benefit from revenues generated by international outreach activities. International Partners shall form the financial spine of the Planet Earth Institute. Current International IYPE Partners are invited to continue supporting the Planet Earth Institute activities once the IYPE Triennium has phased out. The Planet Earth Institute shall be financially self-supportive.



International Year of Planet Earth, Figure 4 Cover of the Planet Earth Institute folder.

Sponsors are invited to support the Foundation during its Initial Stage.

World-class science initiative

Another new initiative is aiming to identify or design "world-class" projects for a "Global Geoscience Initiative." This was explored by an ad-hoc group through a brainstorm meeting in July 2009, and the initiative was then presented to Earth scientific audiences at several major geoscientific gatherings around the world. It is envisaged that the theme(s) of a new world-class science initiative, its structure, and funding mechanism would have been determined by mid-2010.

Summary

The International Year of Planet Earth has been particularly successful in its outreach and education programs. As fundraising started to bear fruit quite late in the Triennium, not sufficient funds were available to properly support a (costly) science program. Moreover, several of the ten science themes of the IYPE were addressed at national levels through current university curricula and special programs. UN proclamation has been crucial for the success of the IYPE, in particular for national implementation of the aims and ambitions through the 80 National and Regional Committees. These have also been the drivers for developing a new initiative, the Planet Earth Institute.

Bibliography

de Mulder, E. F. J., et al., 2008. The international year of planet earth: why and how? *Nature*, **451**(7176), 305–305.

United Nations General Assembly, 2005. The international year of planet earth, 2008 (A/RES/ 60/192), New York.

Cross-references

Electronic Geophysical Year
International Geophysical Year
International Polar Year 2007–2008

INVERSE THEORY, ARTIFICIAL NEURAL NETWORKS

William A. Sandham[1], David J. Hamilton[2]
[1]Scotsig, Glasgow, UK
[2]Ateeda Limited, Edinburgh, UK

Introduction

Neural networks have been recognized for many years as a powerful tool for solving complex problems associated with pattern recognition, classification, and generalization, and have been used in geophysics in a large variety of different applications. The power of neural networks arises from their ability to emulate complex input/output mappings efficiently.

Comprehensive reviews may be found of all the major geophysical applications of neural networks in a number of excellent tutorial papers (Sandham and Leggett, 1998; Van der Baan and Jutten, 2000; Poulton, 2002). A number of books on the subject have also been published in the last decade (Poulton, 2001; Nikravesh et al., 2003; Sandham and Leggett, 2003).

In this present article, the fundamentals of geophysical inversion using neural networks are reviewed. In section "Geophysical inversion," the physics of geophysical inversion are described, together with a short mathematical formulation. In section "Artificial neural networks (ANNs)," the basics of artificial neural networks are discussed. This includes a review of the major types, the training strategies involved, and a discussion of issues such as data preprocessing and performance. In particular, the important case of the multilayer perceptron trained with the back-propagation algorithm is considered in some detail. In section "Geophysical inversion using artificial neural networks," geophysical inversion applications of neural networks are reviewed, including computation of velocity models, split shear-wave analysis, impedance inversion, pseudolog prediction, petrophysics, well logging, and potential field inversion.

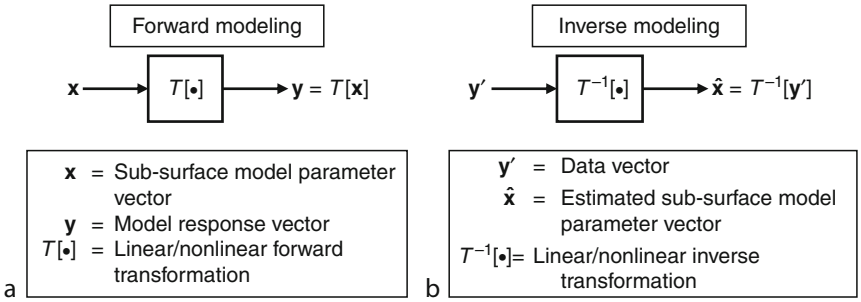
Geophysical inversion

Geophysical inversion, or inverse modeling, may be defined as the estimation of quantitative subsurface earth parameters using measured (observed) geophysical data (Treitel and Lines, 2001; Ulrych and Sacchi, 2005).

Generic definitions of forward and inverse modeling are given in Figure 1.

Mathematically, forward modeling may be described as the transformation $\mathbf{y} = T[\mathbf{x}]$, where $\mathbf{y}$ is the model response, $\mathbf{x}$ is a vector containing the set of subsurface model parameters, and $T[\bullet]$ is some transformation (linear or nonlinear) that describes an observed physical process. For example, in the seismic case, $T[\bullet]$ produces a model response in the form of a synthetic seismogram. Inverse modeling may be written as $\hat{\mathbf{x}} = T^{-1}[\mathbf{y}']$, where $\hat{\mathbf{x}} = T^{-1}[\mathbf{y}']$ is now a vector containing the set of estimated subsurface model parameters (the model space), derived from the data vector $\mathbf{y}'$ (the data space). The operator $T^{-1}[\bullet]$ then denotes the inverse transformation from data space to model space.

Unfortunately, analytical expressions for the inverse transformation seldom exist. Numerically, the inverse problem is solved using methods from mathematical optimization theory. A popular iterative solution is to assume an initial guess for a model, compute the model response $\mathbf{y}$, compare it with the observed response $\mathbf{y}'$, and then perturb the model with the hope of achieving a better fit. However, since the underlying physics in most cases is nonlinear, a straightforward inversion is, even under ideal conditions, inherently unstable. One possible solution is to try and linearize the inverse problem. Unfortunately, such methods are notoriously dependent on the choice of the initial guess for the model. Furthermore, the presence of noise in the data makes interpretation of the results



Inverse Theory, Artificial Neural Networks, Figure 1 (**a**) Forward modeling illustrating the derivation of a model response vector **y**, from the forward transformation $T[\bullet]$ of a subsurface model parameter vector **x**. (**b**) Inverse modeling illustrating the derivation of an estimated subsurface model parameter vector $\hat{\mathbf{x}}$ from the inverse transformation $T^{-1}[\bullet]$ of a data vector $\mathbf{y}'$.

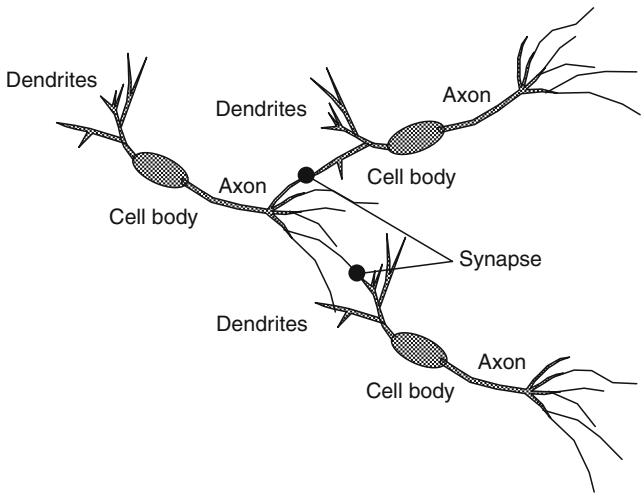
additionally difficult and circumspect. A possible alternative is to use stochastic methods like Monte Carlo (MC), simulated annealing (SA), or genetic algorithms (GA), all of which provide some insight into the statistical nature of the inverse relationships. Unfortunately, MC is extremely inefficient and computing intensive in multidimensional cases, and both SA and GA need subjective tuning parameters that determine the nature of the stochastic search. Most techniques also involve the assimilation of *a priori* knowledge during the inversion process, to constrain the result. All these techniques – linear and nonlinear – involve very expensive forward calculations and are necessarily specific to the particular data set used. Also, the solution is generally nonunique.

It is precisely these intractable features of geophysical inversion that have encouraged geophysicists to investigate solutions using artificial neural networks (ANNs).

Artificial neural networks (ANNs)

The expeditious pattern recognition and classification abilities of humans are a direct consequence of the human brain having a very large number of highly interconnected biological nerve cells or neurons. The basic anatomy of an individual neuron is shown in Figure 2.

It consists of a cell body or soma, the output of which is a long, branching fiber called the axon, which, on appropriate triggering by the cell, transmits an impulse. A human neuron "fires" when its electrical potential reaches a threshold. The output signal is transmitted by the axon and distributed to other neurons via connecting points termed synapses, located in the tree structure at the end of the axon, which modify the strength of the output signal. The input area of the cell is a set of branching fibers called dendrites. Input signals to the soma are attenuated with an increasing distance from the synapses to the soma. Since the human brain is a very dense neural network with a highly interconnected set of approximately 10^{11} neurons, the axon of each neuron is connected to around 10,000 other cells, resulting in approximately 10^{14} connections or synapses. It is this large number of interconnections, which is the basis for the human brain's immense computational power.



Inverse Theory, Artificial Neural Networks, Figure 2 Biological neural network showing three neurons, each of which consists of a cell body or soma, an output or axon, and a series of inputs or dendrites, with the connecting point between an axon and a dendrite termed the synapse.

As the name implies, an artificial neural network (ANN), or simply neural network (NN), is a mathematical or computational model that tries to emulate the structure and/or functional aspects of a biological neural network, and as such, represents a technology that is rooted in many disciplines, including neurosciences, mathematics, statistics, physics, computer science, and engineering. ANNs consist of an interconnected group of artificial neurons (programming constructs that mimic the properties of biological neurons), which process information using a connectionist approach to computation. In most cases, an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during a learning phase, realized via suitable training iterations. ANNs may either be used to gain an understanding of biological neural networks, or for solving artificial intelligence problems without necessarily creating a model of a real biological system. In the

latter case, they may be defined as nonlinear statistical data modeling tools, and can be used to model complex relationships between inputs and outputs or to find patterns in data. Indeed, the power of neural networks arises from their ability to emulate complex input/output mappings efficiently.

There are now some excellent general reviews of ANNs in the literature (Hush and Horne, 1993; Zhang, 2007; Wilamowski, 2009). A substantial number of books on the subject have also been published over the last 2 decades (McCord Nelson and Illingworth, 1991; Priddy and Keller, 2005; Haykin, 2009). The texts by Haykin (including earlier editions) are probably the most comprehensive and definitive.

The basic building block of an ANN is the (mathematical) perceptron, which was conceived almost 70 years ago by McCulloch and Pitts to mimic the behavior of a biological neuron. The perceptron is also known as an artificial neuron, neurode, processing element/unit (PE/PU), or simply unit. A schematic of a perceptron is shown in Figure 3a.

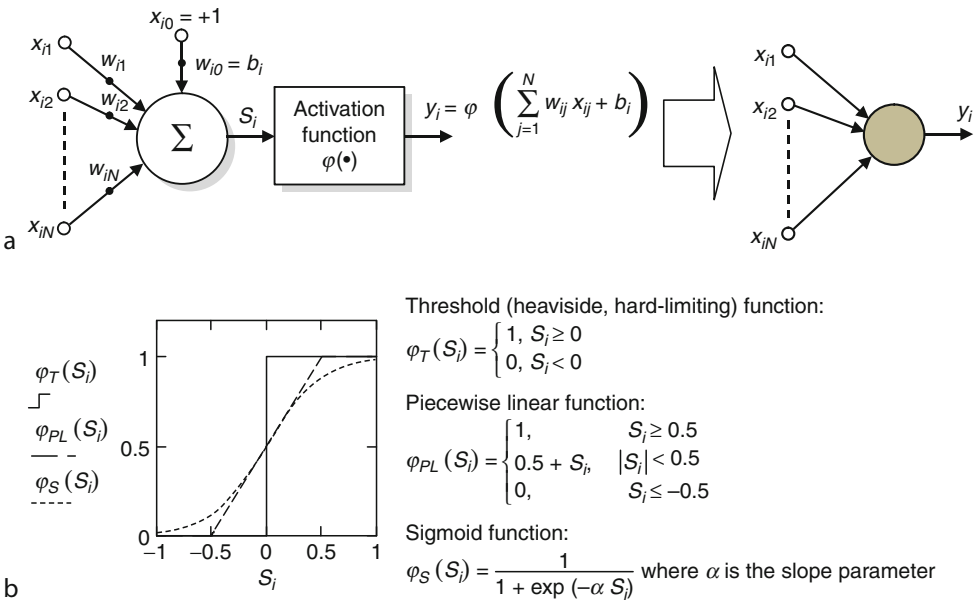
This depicts a number of inputs x_{ij}, analogous to the dendrites in a biological neuron, which are modified by corresponding weights w_{ij}, analogous to the synapses. The $w_{ij}x_{ij}$ products are then summed, along with bias b_i, and the result passed through a nonlinear activation function $\varphi(\bullet)$ to produce the final output y_i, analogous to the axon. The activation function transforms the weighted/summed input S_i to a neuron into its activation. Some major types of activation function are shown in Figure 3b. Historically, the threshold activation function was used. However, this particular function gives only a binary output (i.e., 1 or 0). Also, this particular function is not continuously differentiable and so the optimum weights were very difficult to estimate. A more popular activation function is the sigmoid, which is continuously differentiable and monotonically increasing.

An ANN is formed by connecting (usually many) PEs together in some fashion. Some of the major ANN paradigms are depicted in Figure 4.

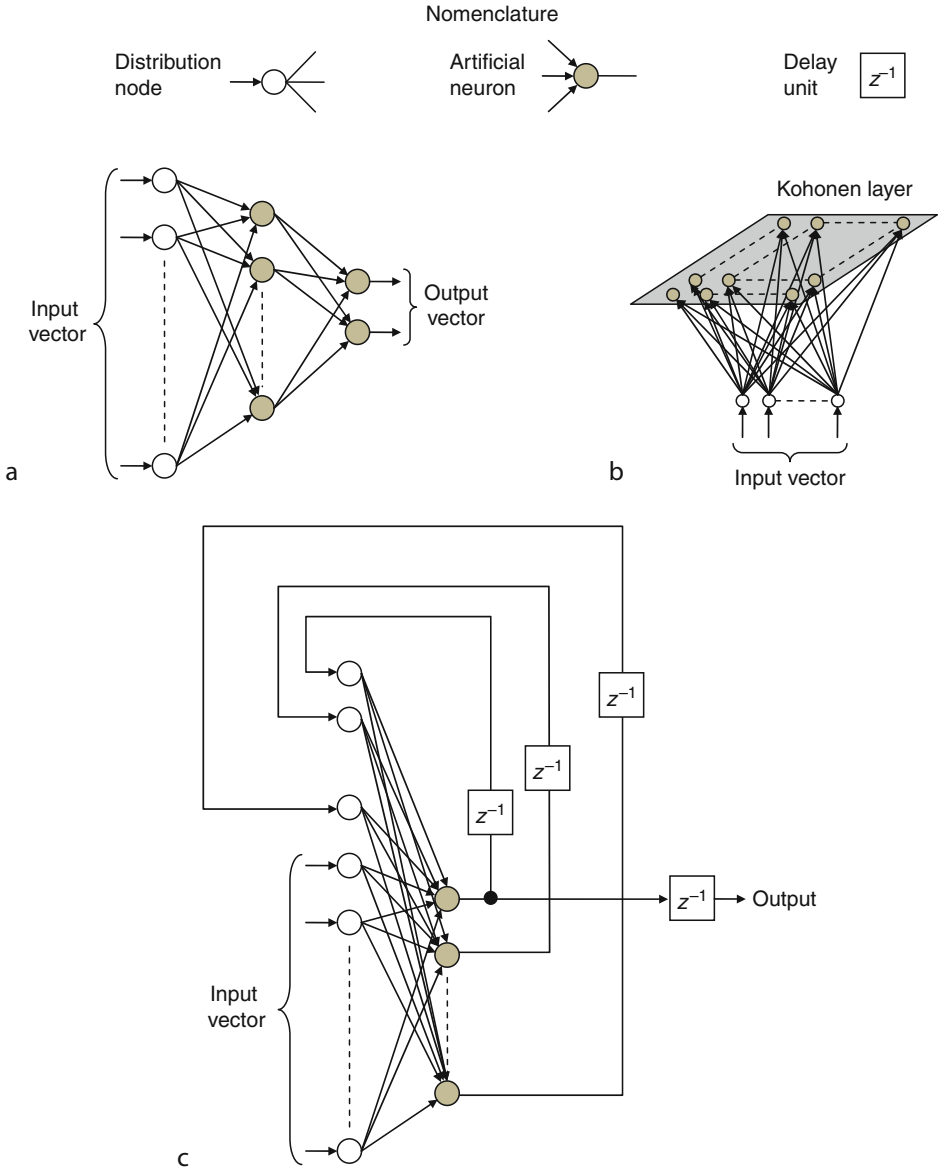
Once an ANN has been chosen, and the number of layers and neurons have been estimated, the various network weights w_{ij} have to be determined, through a network training procedure. Training can either be supervised or unsupervised – see Figure 5a and b.

Supervised training involves presenting target answers for each input pattern, whereas with unsupervised training, the network adjusts its weights in response to input patterns without target answers, effectively classifying the input patterns into similarity categories. After training has been carried out, network performance is assessed using input patterns that have not been used previously, that is, as training data – see Figure 5c. This will indicate whether the network has been over-trained and cannot generalize, and also whether the network has been trapped during training in a local minimum. Performance is dependent on the quality of data, preprocessing involved, training algorithm, network complexity, generalization, and implementation.

One of the most popular ANNs is the so-called multilayer perceptron or MLP (Figure 4a). This is a feedforward



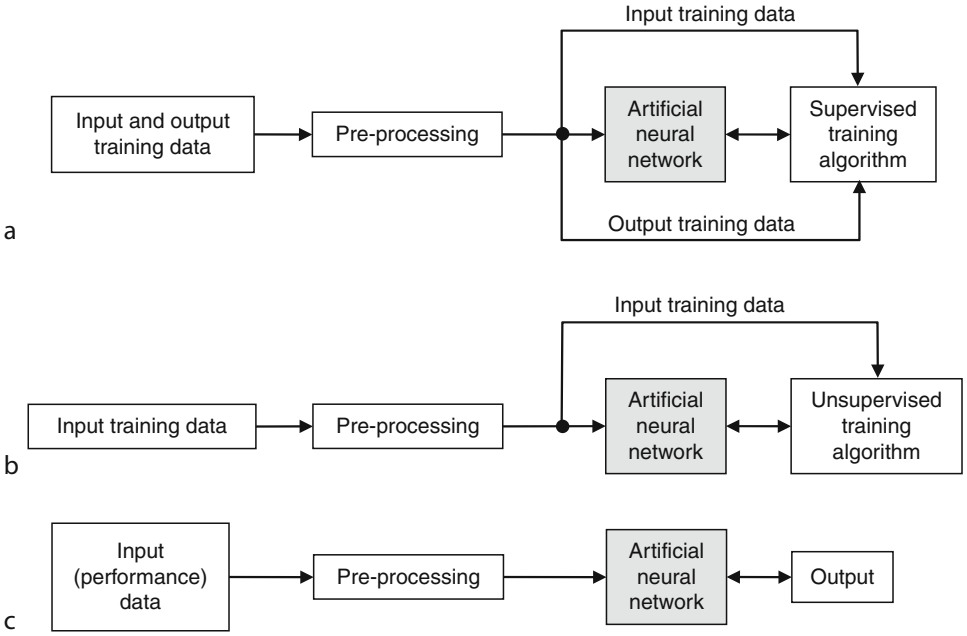
Inverse Theory, Artificial Neural Networks, Figure 3 (**a**) Schematic of an artificial neuron, which modifies inputs x_{ij} by corresponding weights w_{ij}, sums the products along with bias b_i, and passes the result through a nonlinear activation function $\varphi(\bullet)$ to produce the final output y_i. The diagram on the right is a concise way of depicting the neuron. (**b**) Common activation functions used in artificial neural networks. The range of activation functions is usually $0 \leq \varphi(S_i) \leq 1$, as defined here; however, it is sometimes desirable for the range to be $-1 \leq \varphi(S_i) \leq 1$.



Inverse Theory, Artificial Neural Networks, Figure 4 Some common neural network paradigms. (**a**) Fully connected multilayer perceptron (MLP) with one hidden layer and two output neurons. (**b**) Kohonen architecture trained using a self-organizing map (SOM) (unsupervised). (**c**) Fully connected recurrent network with one output neuron.

network consisting of a layer of input distribution nodes, a number of "hidden" layers of perceptrons, and an output layer. In a fully connected version, all distribution nodes and perceptrons are connected to perceptrons in the following layers. The network configuration should allow for an adequate description of the underlying statistical distribution of the spread in the data. Since the number of input nodes and output neurons is usually dictated by the problem, good network design relies on estimating an optimum number of hidden layers and numbers of neurons in each layer. These are usually determined through trial-and-error procedures. A supervised training algorithm called back-propagation (or back-error propagation), popularized in the mid-1980s by Rumelhart and McClelland, is usually used to train a MLP. This iterative, steepest-descent algorithm is outlined below.

Step 1: All w_{ij} set to small random values.
Step 2: Input presented together with target outputs t_i.
Step 3: Actual outputs y_i computed.
Step 4: Errors computed for all nodes in a layer.
Output Layer: $\delta_i = (t_i - y_i)\varphi'(S_i)$
Hidden Layer: $\delta_i = (t_i - y_i)\varphi'(S_i)$



Inverse Theory, Artificial Neural Networks, Figure 5 The two major network training strategies. (**a**) Supervised training. (**b**) Unsupervised training. (**c**) Post-training performance assessment.

Step 5: Update weights.

$$w_{ij}(t+1) = w_{ij}(t) + \eta\delta_i y_j + \alpha\left[w_{ij}(t) - w_{ij}(t-1)\right]$$

Step 6: Return to Step 4 or Step 2.

The weight update in Step 5 includes learning rate (η) and momentum (α) terms to improve performance of the algorithm. The selection of a learning rate is of critical importance in finding the true global minimum of the error distance. Back-propagation training with too small a learning rate will make agonizingly slow progress. Too large a learning rate will proceed much faster, but may simply produce oscillations between relatively poor solutions. Both of these conditions are generally detectable through experimentation and sampling of results after a fixed number of training epochs. Ideally, the largest learning rate should be used that still converges to the minimum solution. Use of a momentum term can be helpful in speeding algorithm convergence and avoiding local minima. This gives the system a certain amount of inertia since the weight vector will tend to continue moving in the same direction unless opposed by the gradient term. The momentum term (i) smooths the weight changes and suppresses cross-stitching, that is, cancels side-to-side oscillations across the error valley, (ii) amplifies the learning rate when all weight changes are in the same direction, causing a faster convergence, and (iii) enables escape from small local minima on the error surface. Typical values for the learning rate and momentum parameter are $0 < \eta, \alpha < 1$. The momentum term will allow a larger learning rate and thus speed convergence and avoid local minima. On the other hand, a learning rate of 1 with no momentum will be much faster when no problem with local minima or non-convergence is encountered. Other popular training algorithms include Quickprop and Cascade Correlation.

A very popular network that uses unsupervised training is Kohonen's Self-Organizing Feature Map or SOM (Figure 4b). This is a two-layered feedforward network consisting of a layer of input distribution nodes, and an output layer usually arranged as a 2-D map, which shows the natural relationships among the patterns presented to the network. Essentially, neighboring neurons compete in their activation by means of mutual lateral interactions, and develop adaptively into specific detectors of different input patterns. The unsupervised training algorithm is outlined below.

Step 1: All w_{ij} set to small random values.
Step 2: Input presented.
Step 3: Euclidian distances δ_i between each input node (j) and output node (i) computed.

$$\delta_i = \sum \left[x_j(t) - w_{ij}(t)\right]^2$$

Step 4: Select output node i^* with minimum distance.
Step 5: Update weights to node i^* and neighbors.

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t)\left[x_j(t) - w_{ij}(t)\right], \mathrm{i} \in NE_{i*}(t)$$

Step 6: Return to Step 2.

Recurrent networks, which involve the inclusion of feedback loops such as the one depicted in Figure 4c, have a profound impact on the learning capability of the network and its performance.

Many other types of ANN exist, including Hopfield recurrent networks, those based on the adaptive resonance theory of Carpenter and Grossberg, probabilistic, modular, and radial-basis function ANNs, support vector machines, and more recently, networks including spiking ANNs and wavelet ANNs.

Geophysical inversion using artificial neural networks

ANNs offer many advantages to geophysicists for inversion applications, and this has made them an intensely active research area. Indeed, many ANN inversion algorithms are now sufficiently mature and are routinely employed in their respective application area. A number of good reviews of geophysical inversion using ANNs are available in the literature (Röth and Tarantola, 1994; Nunnari et al., 2001).

The basis of ANN geophysical inversion methods is illustrated in Figure 6.

First, a geophysical model is used to derive many simulated model response vectors $\mathbf{y}^{(i)}$ from a large number N of subsurface model parameter vectors $\mathbf{x}^{(i)}$, where $1 \leq i \leq N$. Next, the response vectors act as input data training sets to an ANN. In supervised training the model parameter vectors are used as the target vectors. Finally, after appropriate training, a new survey data vector $\mathbf{y}'$ is input to the ANN, and an estimated subsurface model parameter vector $\hat{\mathbf{x}}$ obtained, that is, geophysical inversion has been achieved. There are many variations of this basic technique.

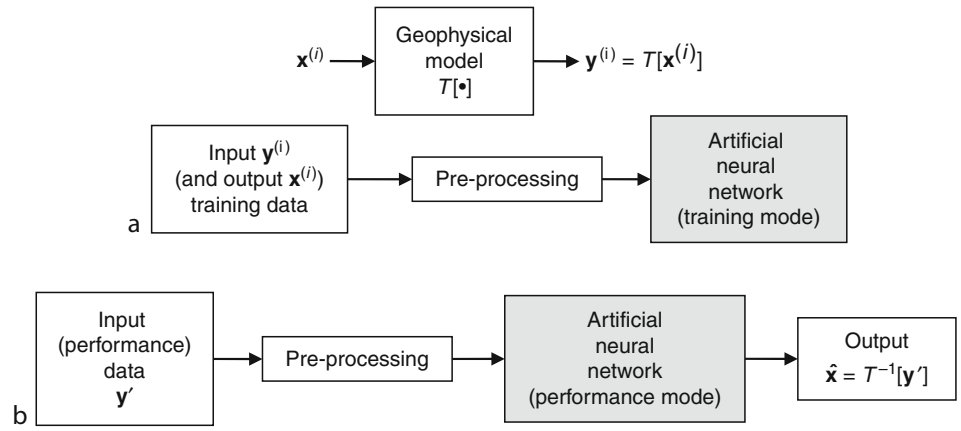
It is important to emphasize that geophysical inversion using ANNs will only be successful if proper consideration and diligence are provided during the preprocessing and training phases. This is to ensure the prevention of over-training and local minima entrapment (Maiti and Tiwari, 2010).

There are many examples in the literature of ANNs being employed across the complete spectrum of geophysical inversion applications, and these publications, many of which are cited below under their respective application domains, should be consulted for further information.

SEISMIC
General: (Michaels and Smith, 1997; Xiangjun et al., 2006)
Seismic Attributes: (Dumitrescu, 2009)
Porosity: (Wang et al., 1997; Ecoublet et al., 1998; Burnett et al., 2004)
Impedance: (Fu, 2004)
Facies Analysis: (Castro de Matos et al., 2007)
Refraction Statics: (Taner et al., 1988)
Seismic Velocities: (Calderón-Macias et al., 1998, 2000)
Amplitude Variation with Offset (AVO): (Mogensen and Link, 2001; Kuzma and Rector, 2004)
Split Shear Wave: (Dai and MacBeth, 1994)
Seismic Tomography: (Tselentis et al., 2007)
WELL-LOGS: (Liu and Liu,1998; Hampson et al., 2001; Zhang and Zhou, 2002; Aristodemou et al., 2005)
RESERVOIR GEOPHYSICS: (Saggaf et al., 2003; Aminzadeh and Brouwer, 2006)
ELECTROMAGNETIC: (Winkler, 1994; Seiberl et al., 1998; Spichak and Popova, 2000; Ahl, 2003)
RESISTIVITY: (Teles and Do Cormo,1999; El-Qady and Ushijima, 2001; Neyamadpoura et al., 2009)
POTENTIAL FIELD INVERSION
General: (Guo et al., 1992; Zhang and Paulson, 1997)
Gravity: (Kaftan and Şalk, 2009)
Magnetic: (Al-Garni, 2009)
ARCHAEOLOGY: (Bescoby et al., 2006)
REMOTE SENSING: (Del Frate et al., 2007; Tso and Mather, 2009)

Summary

Geophysical inversion is an essential tool to help geophysicists solve practical environmental, engineering, or exploration problems. This is particularly relevant for situations



Inverse Theory, Artificial Neural Networks, Figure 6 Basis of geophysical inversion using ANNs. (**a**) A geophysical model is used to derive many simulated model response vectors $\mathbf{y}^{(i)}$ from subsurface model parameter vectors $\mathbf{x}^{(i)}$. The response vectors then act as input data training sets to an ANN. In supervised training the model parameter vectors are also used. (**b**) After appropriate training, a new survey data vector $\mathbf{y}'$ is input to the ANN, and an estimated subsurface model parameter vector $\hat{\mathbf{x}}$ obtained.

that require detailed quantitative subsurface information and data. Traditional approaches to inversion are many and varied, and include model perturbation/comparison, linearization usually of a nonlinear geophysical model, together with a variety of stochastic methods including Monte Carlo, simulated annealing, and genetic algorithms. These approaches do not perform well in the presence of noise, are very inefficient and computing intensive, require subjective tuning parameters and a priori knowledge, and/or produce results that are nonunique.

Alternative geophysical inversion methods, based on artificial neural networks (ANNs), have been investigated over the past 15–20 years, to try and overcome the disadvantages of the traditional approaches. These ANN inversion methods have now found considerable success across many different geophysical domains, including seismic attributes, velocity fields, and tomography, to potential field inversion, archaeology, and remote sensing.

With the continuous introduction of improved ANN paradigms and training algorithms, coupled with the ever-increasing power of digital computing hardware (including Field Programmable Gate Arrays [FPGAs]), there is no doubt that ANNs have a very positive future in the exciting and challenging world of geophysical inversion.

Bibliography

Ahl, A., 2003. Automatic 1D inversion of multifrequency airborne electromagnetic data with artificial neural networks: discussion and a case study. *Geophysical Prospecting*, **51**, 89–97.

Al-Garni, M. A., 2009. Interpretation of some magnetic bodies using neural networks inversion. *Arabian Journal of Geosciences*, **2**, 175–184.

Aminzadeh, F., and Brouwer, F., 2006. Integrating neural networks and fuzzy logic for improved reservoir property prediction and prospect ranking. In *Proceedings SEG Annual Conference and Exposition*, New Orleans, October 1–6, 2006.

Aristodemou, E., Pain, C., de Oliveira, C., Goddard, T., and Harris, C., 2005. Inversion of nuclear well-logging data using neural networks. *Geophysical Prospecting*, **53**, 103–120.

Bescoby, D. J., Cawley, G. C., and Chroston, P. N., 2006. Enhanced interpretation of magnetic survey data from archaeological sites using artificial neural networks. *Geophysics*, **71**, H45–H53.

Burnett, M. D., Castagna, J. P., Camargo, G., Chen, H., Sanchez, J. J., Santana, A., and Hernandez, E. M., 2004. Synergistic porosity mapping in the upper cretaceous of the Chiapas region using spectral decomposition and neural network inversion. In *Proceedings SEG Annual Conference and Exposition,* Denver, October 10–15, 2004.

Calderón-Macias, C., Sen, M. K., and Stoffa, P. L., 1998. Automatic NMO correction and velocity estimation by a feedforward neural network. *Geophysics*, **63**, 1696–1707.

Calderón-Macias, C., Sen, M. K., and Stoffa, P. L., 2000. Artificial neural networks for parameter estimation in geophysics. *Geophysical Prospecting*, **48**, 21–47.

Castro de Matos, M., Osorio, P. L. M., and Johann, P. R. S., 2007. Unsupervised seismic facies analysis using wavelet transform and self-organizing maps. *Geophysics*, **72**, P9–P21.

Dai, H., and MacBeth, C., 1994. Split shear-wave analysis using an artificial neural network. *First Break*, **12**, 605–613.

Del Frate, F., Pacifici, F., Schiavon, G., and Solimini, C., 2007. Use of neural networks for automatic classification from high-resolution images. *IEEE Transactions on Geoscience and Remote Sensing*, **45**, 800–809.

Dumitrescu, C. C., 2009. Exploration of Lower Cretaceous sands in the Leland Area, Alberta, using seismically derived rock properties. *First Break*, **27**, 53–60.

Ecoublet, P. O., Symes, W. W., and Levin, S. A., 1998. Porosity prediction from seismic inversion using artificial neural networks. In *Proceedings EAGE Annual Conference and Exposition*, Leipzig, June 8–12, 1998.

El-Qady, G., and Ushijima, K., 2001. Inversion of DC resistivity data using neural networks. *Geophysical Prospecting*, **49**, 417–430.

Fu, L.-Y., 2004. Joint inversion of seismic data for acoustic impedance. *Geophysics*, **69**, 994–1004.

Guo, Y., Hansen, R., and Harthill, N., 1992. Feature recognition from potential fields using neural networks. In *Proceedings SEG Annual Conference and Exposition*, New Orleans, October 10–15, 1992.

Hampson, D. P., Schuelke, J. S., and Quirein, J. A., 2001. Use of multiattribute transforms to predict log properties from seismic data. *Geophysics*, **66**, 220–236.

Haykin, S., 2009. *Neural Networks and Learning Machines*. Upper Saddle River: Pearson Education.

Hush, D. R., and Horne, B. G., 1993. Progress in supervised neural networks – what's new since Lippmann? *IEEE Signal Processing Magazine*, **10**, 8–39.

Kaftan, I., and Şalk, M., 2009. Determination of structure parameters on gravity method by using radial basis functions networks case study: Seferihisar geothermal area (Western Turkey). In *Proceedings SEG Annual Conference and Exposition*, Houston, October 25–30, 2009.

Kuzma, H. A., and Rector, J. W., 2004. Non-linear AVO inversion using support vector machines. In *Proceedings SEG Annual Conference and Exposition*, Denver, October 10–15, 2004.

Liu, Z., and Liu, J., 1998. Seismic-controlled nonlinear extrapolation of well parameters using neural networks. *Geophysics*, **63**, 2035–2041.

Maiti, S., and Tiwari, R. K., 2010. Automatic discriminations among geophysical signals via the Bayesian neural networks approach. *Geophysics*, **75**, E67–E78.

McCord Nelson, M., and Illingworth, W. T., 1991. *A Practical Guide to Neural Nets*. Reading: Addison-Wesley.

Michaels, P., and Smith, R. B., 1997. Surface wave inversion by neural networks (radial basis functions) for engineering applications. *Journal of Environmental and Engineering Geophysics*, **2**, 65–76.

Mogensen, S., and Link, C. A., 2001. Artificial neural network solutions to AVO inversion problems. In *Proceedings SEG Annual Conference and Exposition*, San Antonio, September 9–14, 2001.

Neyamadpoura, A., Taibb, S., and Abdullaha, W. A. T. W., 2009. Using artificial neural networks to invert 2D DC resistivity imaging data for high resistivity contrast regions: A MATLAB application. *Computers & Geosciences*, **35**, 2268–2274.

Nikravesh, M., Aminzadeh, F., and Zadeh, L. A. (eds.), 2003. *Soft Computing and Intelligent Data Analysis in Oil Exploration. Developments in Petroleum Science*. Amsterdam: Elsevier.

Nunnari, G., Bertucco, L., and Ferrucci, F., 2001. A neural approach to the integrated inversion of geophysical data of different types. *IEEE Transactions on Geoscience and Remote Sensing*, **39**, 736–748.

Poulton, M. M. (ed.), 2001. *Computational Neural Networks for Geophysical Data Processing*. London: Elsevier. Handbook of Geophysical Exploration: Seismic Exploration.

Poulton, M. M., 2002. Neural networks as an intelligence amplification tool: a review of applications. *Geophysics*, **67**, 979–993.

Priddy, K. L., and Keller, P. E., 2005. *Artificial Neural Networks: An Introduction*. Bellingham: The International Society for Optical Engineering.
Röth, G., and Tarantola, A., 1994. Neural networks and inversion of seismic data. *Journal of Geophysical Research*, **99**(B4), 6753–6768.
Saggaf, M. M., Toksöz, M. N., and Mustafa, H. M., 2003. Estimation of reservoir properties from seismic data by smooth neural networks. *Geophysics*, **68**, 1969–1983.
Sandham, W. A., and Leggett, M., 1998. Geophysical applications of artificial neural networks and fuzzy logic. *First Break*, **16**, 58–60.
Sandham, W., and Leggett, M. (eds.), 2003. *Geophysical Applications of Artificial Neural Networks and Fuzzy Logic*. Dordrecht: Kluwer. Modern Approaches in Geophysics.
Seiberl, W., Ahl, A., and Winkler, E., 1998. Interpretation of airborne electromagnetic data with neural networks. *Exploration Geophysics*, **29**, 152–156.
Spichak, V., and Popova, I., 2000. Artificial neural network inversion of magnetotelluric data in terms of three-dimensional earth macro-parameters. *Geophysical Journal International*, **142**, 15–26.
Taner, T., Lu, L., and Baysal, E., 1988. Unified method for 2-D and 3-D refraction statics with first break picking by supervised learning. In *Proceedings SEG Annual Conference and Exposition*, Anaheim, October 19–24, 1988.
Teles, T. N., and Do Carmo, E. A., 1999. DC resistivity data inversion using neural networks. In *Proceedings EAGE Annual Conference and Exposition*, Helsinki, June 7–11, 1999.
Treitel, S., and Lines, L., 2001. Past, present, and future of geophysical inversion – a Y2K analysis. *Geophysics*, **66**, 21–24.
Tselentis, G.-A., Serpetsidaki, A., Martakis, N., Sokos, E., Paraskevopoulos, P., and Kapotas, S., 2007. Local high-resolution passive seismic tomography and Kohonen neural networks – application at the Rio-Antirio Strait, central Greece. *Geophysics*, **72**, B93–B106.
Tso, B., and Mather, P., 2009. *Classification Methods for Remotely Sensed Data*. Boca Raton: CRC Press.
Ulrych, T. J., and Sacchi, M. D., 2005. *Information-Based Inversion and Processing with Applications*. Amsterdam: Elsevier.
Van der Baan, M., and Jutten, C., 2000. Neural networks in geophysical applications. *Geophysics*, **65**, 1032–1047.
Wang, B., Pann, K., Schuelke, J. S., Shirley, T., and Ferguson, B., 1997. View of neural network training as constrained optimisation and applications to rock porosity prediction. In *Proceedings SEG Annual Conference and Exposition*, Dallas, October 25–30, 1997.
Wilamowski, B. M., 2009. Neural network architectures and learning algorithms. *IEEE Industrial Electronics Magazine*, **3**, 56–63.
Winkler, E., 1994. Inversion of electromagnetic data using neural networks. In *Proceedings EAGE Annual Conference and Exposition*, Vienna, June 6–10, 1994.
Xiangjun, Z., Jianjun, X., and Qinxue, L., 2006. Rock parameter inversion using well and depth domain seismic data by SRM neural networks. In *Proceedings SEG Annual Conference and Exposition*, New Orleans, October 1–6, 2006.
Zhang, G. P., 2007. Avoiding pitfalls in neural network research. *IEEE Transactions on Systems, Man, and Cybernetics. Part C: Applications and Reviews*, **37**, 3–16.
Zhang, Y., and Paulson, K. V., 1997. Magnetotelluric inversion using regularized Hopfield neural networks. *Geophysical Prospecting*, **45**, 725–743.
Zhang, Z., and Zhou, Z., 2002. Real time quasi–2-D inversion of array resistivity logging data using neural network. *Geophysics*, **67**, 517–524.

Cross-references

Deep Seismic Reflection and Refraction Profiling
Earth, Density Distribution
Electrical Resistivity Surveys and Data Interpretation
Geomagnetic Field, Measurement Techniques
Geophysical Well Logging
Gravity Data, Advanced Processing
Gravity Method, Airborne
Gravity Method, Satellite
Gravity Method, Surface
Instrumentation, Electrical Resistivity
Instrumentation, EM
Inverse Theory, Linear
Inverse Theory, Monte Carlo Method
Inverse Theory, Singular Value Decomposition
Magnetic Methods, Airborne
Magnetic Methods, Satellite
Magnetic Methods, Surface
Magnetotelluric Data Processing
Magnetotelluric Interpretation
Poroelasticity
Propagation of Elastic Waves: Fundamentals
Remote Sensing, Applications to Geophysics
Sedimentary Basins
Seismic Anisotropy
Seismic Data Acquisition and Processing
Seismic Velocity-Density Relationships
Seismic Velocity-Temperature Relationships
Seismic, Viscoelastic Attenuation
Seismic, Waveform Modeling and Tomography
Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting
Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat
Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity
Traveltime Tomography Using Controlled-Source Seismic Data
Wavelet Analysis

INVERSE THEORY, GLOBAL OPTIMIZATION

Mrinal K. Sen, Paul L. Stoffa
Department of Geological Sciences, UT Institute for Geophysics, Jackson School of Geosciences, The University of Texas at Austin, Austin, TX, USA

Definition

Inversion of geophysical data involves searching for optimal set(s) of earth model parameters that can be used to compute synthetic data that match with observations. The data misfit (or fitness) is often measured by a suitably defined objective function that is generally non-convex. Optimization methods search for a minimum (optimal) value of an objective function. In general, there can be solutions that are locally optimal.

Global optimization refers to finding the best set of parameters corresponding to the global minimum of a function of multiple parameters.

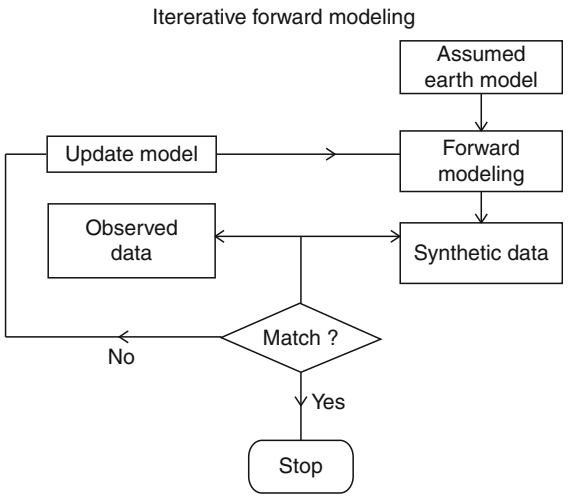
Introduction

The primary goal of geophysics is to determine the properties of the earth's interior from a finite set of remotely sensed measurements from instruments placed on the surface of the earth with perhaps a few in boreholes. The forward problem involves computation of theoretical data for a given set of earth model parameters. Typically, this requires numerical solution of a partial differential equation. The inverse problem whose goal is to estimate earth model parameters from observations is nontrivial due to the fact that data are almost always inadequate, inconsistent, and insufficient (Jackson, 1972). The direct inversion methods are formulated based on the physics of the forward problem by recognizing or designing a mathematical operator, which is applied to the observed data (sometimes recursively) to derive a model of the earth. Such methods are generally unstable due to the typical data problems of inadequacy.

In a *model-based inversion* approach, synthetic data are generated for an assumed model and compared with the observed data. If the match between observed and synthetic data is acceptable, the model is accepted as the solution, otherwise, the model is changed. The synthetics are recomputed and again compared with the observations. This iterative forward modeling procedure is repeated until an acceptable match is obtained between data and synthetics. Thus in this approach, inversion is viewed as an *optimization* process in which a model is sought that best explains the observations (Figure 1). Therefore, searching for a model that explains the field observation the best, comprises one of the major tasks of inversion.

Inversion is not restricted to just finding the so-called best-fit model since there can be multiple solutions. In other words, often we have nonunique solutions. Thus it is very important for a practitioner to have a clear understanding of the forward problem and to impose constraints and restrict the search for an optimal realistic solution within regions that are geologically meaningful. This can be addressed by a proper choice of an objective function. One approach to inversion may be termed "exploitation of model space" in which one simply seeks a best-fit model. Another approach may be termed "exploration of model space," which casts the inverse problem in Bayesian Framework (Tarantola, 1987) and constructs a posteriori probability density function (PPD) in model space.

Details of geophysical inverse theory with some examples of application can be found in the texts by Menke (1984), Tarantola (1987), Sen and Stoffa (1995), Aster et al. (2005), and Sen (2006). In the following, we will provide a summary of global optimization methods with some examples of application to geophysical inversion.



Inverse Theory, Global Optimization, Figure 1 Principle of model-based inversion by optimization.

Method

Background

The essential elements of a model-based inversion algorithm include data, model, forward problem, objective function, and optimization method. We will restrict ourselves to discrete data **d** and model **m** defined by the following vectors

$$\begin{aligned} \mathbf{d} &= [d_1\, d_2\, d_3 \ldots d_N]^T, \\ \mathbf{m} &= [m_1\, m_2\, m_3 \ldots m_M]^T. \end{aligned} \tag{1}$$

In general $N \neq M$. The forward problem can be represented by the following equation

$$d^{syn} = g(m), \tag{2}$$

where g is a nonlinear forward modeling operator that operates on the model vector to generate synthetic data vector $\mathbf{d}^{syn}$. The next step is to define an objective function that measures a misfit between the observed and synthetic data using a suitably defined norm. The data residual is given by

$$\mathbf{\Delta d} = \mathbf{d}^{obs} - \mathbf{d}^{syn}, \tag{3}$$

where $\mathbf{d}^{obs}$ is the observed data vector and the data misfit norm can be written as

$$\|\mathbf{\Delta d}\|_p = \left[\sum_{i=1}^{N} |\Delta d_i|^p\right]^{1/p}, \tag{4}$$

where p is an integer representing the order of the norm. The most commonly used norm is an L_2 norm for which $p = 2$ resulting in a minimization of the mean square error. This corresponds to the popular least squares minimization approach.

To address the issues of stability and nonuniqueness, constraints are incorporated by adding a model norm term, which can be used to impose smoothness, positivity, and bounds. A detailed description of this issue is beyond the

scope of this article. A fairly general objective function with such constraints using an L_2 norm can be written as

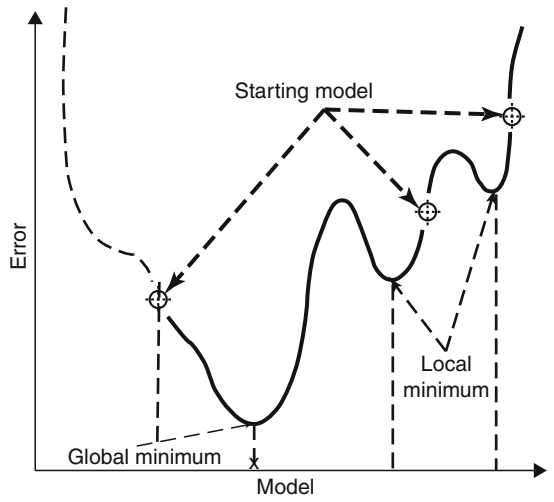
$$F(\mathbf{m}) = \left(\mathbf{d}^{obs} - \mathbf{d}^{syn}\right)^T \mathbf{W}_d \left(\mathbf{d}^{obs} - \mathbf{d}^{syn}\right) + \alpha \left(\mathbf{m} - \mathbf{m}^{prior}\right)^T \mathbf{W}_m \left(\mathbf{m} - \mathbf{m}^{prior}\right), \quad (5)$$

where $\mathbf{W}_d$ and $\mathbf{W}_m$ are the data and model weighting matrices, $\mathbf{m}^{prior}$ is the prior model, and α is a regularization weight. The misfit function given by Equation 5 is also called the objective function, cost function, error function, energy function, etc., in different applications.

Optimization

The goal of an optimization algorithm is to search for the minimum of an objective function such as the one defined by Equation 5. Note that the operator g in Equation 2 is, in general, a nonlinear operator. Therefore the objective function (Equation 5) is expected to be non-convex with multiple peaks and troughs (Figure 2). In other words, the objective function can have multiple local minima; the minimum of all the local minima is called a "global" minimum (e.g., Gill et al., 1981). In most applications, once an objective function has been defined, the goal is to find the global minimum of the objective function.

Optimization methods vary depending on the method of search for the minimum. They can be broadly classified into two categories: local optimization and global optimization. The local optimization methods generally use calculus-based rules to find an update in their search. In other words, they make use of local gradient and/or curvature of the objective function to compute an update. Most of these methods are deterministic in nature and the success of these methods is largely dictated by the choice of the starting model. Only if the starting model is close to the globally optimal model can we expect to reach the global minimum using local optimization methods. These techniques cannot generate or even use the global information needed to find the global minimum for a function with multiple local minima.



Inverse Theory, Global Optimization, Figure 2 A hypothetical error function as a function of model (the model is assumed to have only one model parameter) showing several minima. The iterative gradient method will find the global minimum only when starting at position 1. Others will end up in secondary minima of the error function.

The global optimization methods, on the other hand, are mostly based on stochastic rules and use more global information about the misfit surface to update their current position. The interaction between computer science and optimization has yielded new practical solvers for global optimization problems, called *meta-heuristics*. The structures of meta-heuristics are mainly based on simulating nature and artificial intelligence tools. Meta-heuristics mainly invoke exploration and exploitation search procedures in order to diversify the search all over the search space and intensify the search in some promising areas. Therefore, meta-heuristics cannot easily be entrapped in local minima. However, meta-heuristics are computationally expensive due to their slow convergence. One of the main reasons for their slow convergence is that they may fail to detect promising search directions especially in the vicinity of local minima due to their random constructions.

The convergence of these methods to the globally optimal solution is not guaranteed for all the algorithms. Only for some of the algorithms under certain conditions is convergence to the globally optimal solution statistically guaranteed. Also, with real observational data it is never possible to know whether the derived solution corresponds to the global minimum or not. However, our experience indicates that we are able to find many good solutions starting with only poor initial models using global optimization methods (Sen and Stoffa, 1995). They are less greedy than the well-known local optimization methods in that during iterative optimization worse solutions are occasionally accepted, which allow these algorithms to avoid local minima. There are several variants of global optimization methods; detailed descriptions of some of these approaches as applied to geophysical inversion can be found in Sen and Stoffa (1995). Here we will provide brief descriptions of a few commonly used global optimization methods, namely the *Simulated Annealing* (SA), *Genetic Algorithms* (GA), *Neighborhood Algorithms* (NA), and *Particle Swarm Optimization* (PSO).

Simulated annealing

Simulated annealing was first proposed by Kirkpatrick et al. (1983). It is analogous to the natural process of crystal annealing when a liquid gradually cools to a solid state. The SA technique starts with an initial model $\mathbf{m}_0$, with associated error or energy $E(\mathbf{m}_0)$. It draws a new model $\mathbf{m}_{\text{new}}$ from a flat distribution of models within the predefined limits. Note that each model parameter can be bounded by different limits. The associated energy (objective function value) $E(\mathbf{m}_{\text{new}})$ is then computed, and compared against $E(\mathbf{m}_0)$. If the energy of the new state is less

than the initial state, the new state is considered to be good. In this case, the new model is accepted and replaces the initial model unconditionally. However, if the energy of the new state is larger than the initial state, $\mathbf{m}_{\text{new}}$ is accepted with the probability of $\exp(-(E(\mathbf{m}_{\text{new}}) - E(\mathbf{m}_0))/T)$, where T is a control parameter called annealing temperature that controls if the "bad" model should be carried over to the new model. This completes one iteration. It is the rule of accepting with a probability that makes it possible for SA to be able to jump out of the local minima. The same process is repeated for a large number of times, with the annealing temperature gradually decreasing according to a predefined scheme. Hopefully with a carefully defined cooling schedule, a global minimum can be found. One may choose a linear, or a logarithmically decreasing cooling scheme. The tradeoff here is between the computation cost and the accuracy of the result. Fast cooling will fail to produce a crystal (the algorithm gets stuck in a local minimum), slow cooling takes a long time but may eventually find the global minimum.

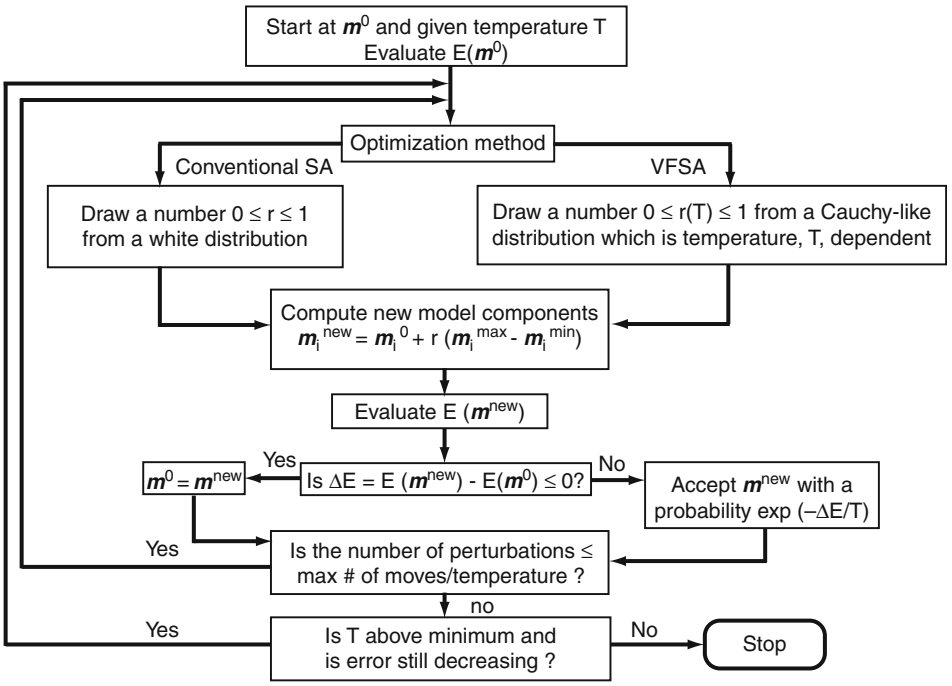
To speed up the annealing process without much sacrifice in the solution, a variant of SA, called Very Fast Simulated Annealing (VFSA) was proposed by Ingber (1989). VFSA differs from SA in the following ways. The new model is drawn from a temperature dependent Cauchy-like distribution centered on the current model (Figure 3). This change has two fundamental effects. First it allows for larger sampling of the model space at the early stages of the inversion when the temperature is high, and much narrower sampling in the model space as the inversion converges when the temperature decreases. Second, each model parameter can have its own cooling schedule and model space-sampling scheme. Therefore it allows for the individual control for each parameter, and the incorporation of a priori information. For many geophysical inversion applications, VFSA has been demonstrated to have excellent performance (Sen and Stoffa, 1995).

Genetic algorithm

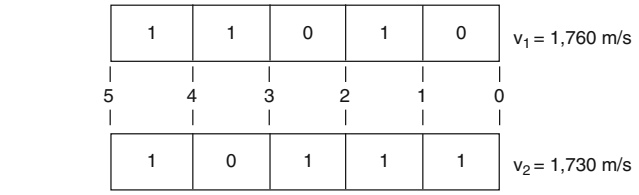
Genetic algorithms are so named because they emulate the biological processes of evolution and are based on the principle of survival of the fittest (e.g., Goldberg, 1989). In a simple GA, the model parameters are coded in binary form (Figure 4). The algorithm starts with a randomly chosen population of models called *chromosomes*. The second step is to evaluate the fitness values of these models. Note that unlike our preceding discussion on optimization in which we attempt to minimize an objective function, a GA searches for the maximum of a suitably defined fitness function that measures similarities between the observed and synthetic data. After the initial selection process, the three genetic processes of selection, crossover, and mutation are performed upon the models in sequence.

In *selection*, models are copied in proportion to their fitness values based on a probability of selection. Thus during this process, models with higher fitness values (lower error values) have a higher probability of getting selected.

Crossover acts on the selected pairs of models. This operation picks a crossover site within the paired strings and exchanges, based on a crossover probability, the bits

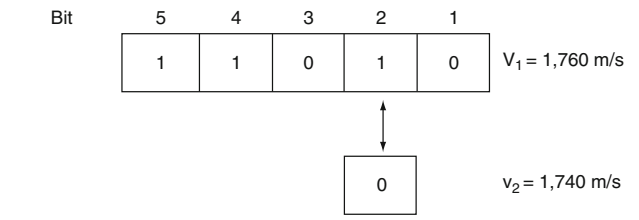


Inverse Theory, Global Optimization, Figure 3 A flowchart for global optimization by simulated annealing. Note that the VFSA draws models from a Cauchy-like distribution whose shape changes with iteration.



a

Crossover position	v_1 m/s	v_2 m/s
0	1,760	1,730
1	1,770	1,720
2	1,770	1,720
3	1,810	1,680
4	1,730	1,760
5	1,730	1,760



b

Mutation bit	v_1 m/s	v_2 m/s
1	1,760	1,770
2		1,740
3		1,800
4		1,680
5		1,600

Inverse Theory, Global Optimization, Figure 4 Genetic algorithms are based on binary coding of model parameters (**a**) and then processes of crossover(**a**) and mutation (**b**). Figure (**a**) shows bit representation of velocity values and the new values resulting from crossover at different crossover positions. Figure (**b**) demonstrates how the velocity changes with mutation at different mutation bits.

between the two models to the right of the crossover site (Figure 4). Thus the crossover process results in two new models (or children).

Mutation, which involves changing a bit at random based on a mutation probability, is applied to the models to maintain diversity.

After execution of these three processes, the new models are compared to the previous generation and accepted based on an update probability (Stoffa and Sen, 1991). The procedure is repeated until convergence is reached, that is, when the fitness of all the models becomes close to one another.

Note that the repeated applications of selection, crossover, mutation, and update do not necessary guarantee convergence to the global maximum of the fitness function. Nonetheless, the algorithm has been found to be successful in many optimization problems including those from Geophysics.

Neighborhood algorithm (NA)

NA was introduced by Sambridge (1999). The principal idea behind this algorithm is to generate a set of models, at each generation, whose sampling density is constructed from all previous models (Sambridge and Mosegaard, 2002). In other words, the proposal distribution to draw the models from is updated at each iteration. One other unique feature of the algorithm is that the algorithm involves portioning of the model parameter space into nearest-neighbor or Voronoi cells about each of the models, and thus the information in the previous model samples drives the search for new models. At regular intervals the approximation is updated, and sampling can concentrate in multiple regions.

Particle swarm optimization (PSO)

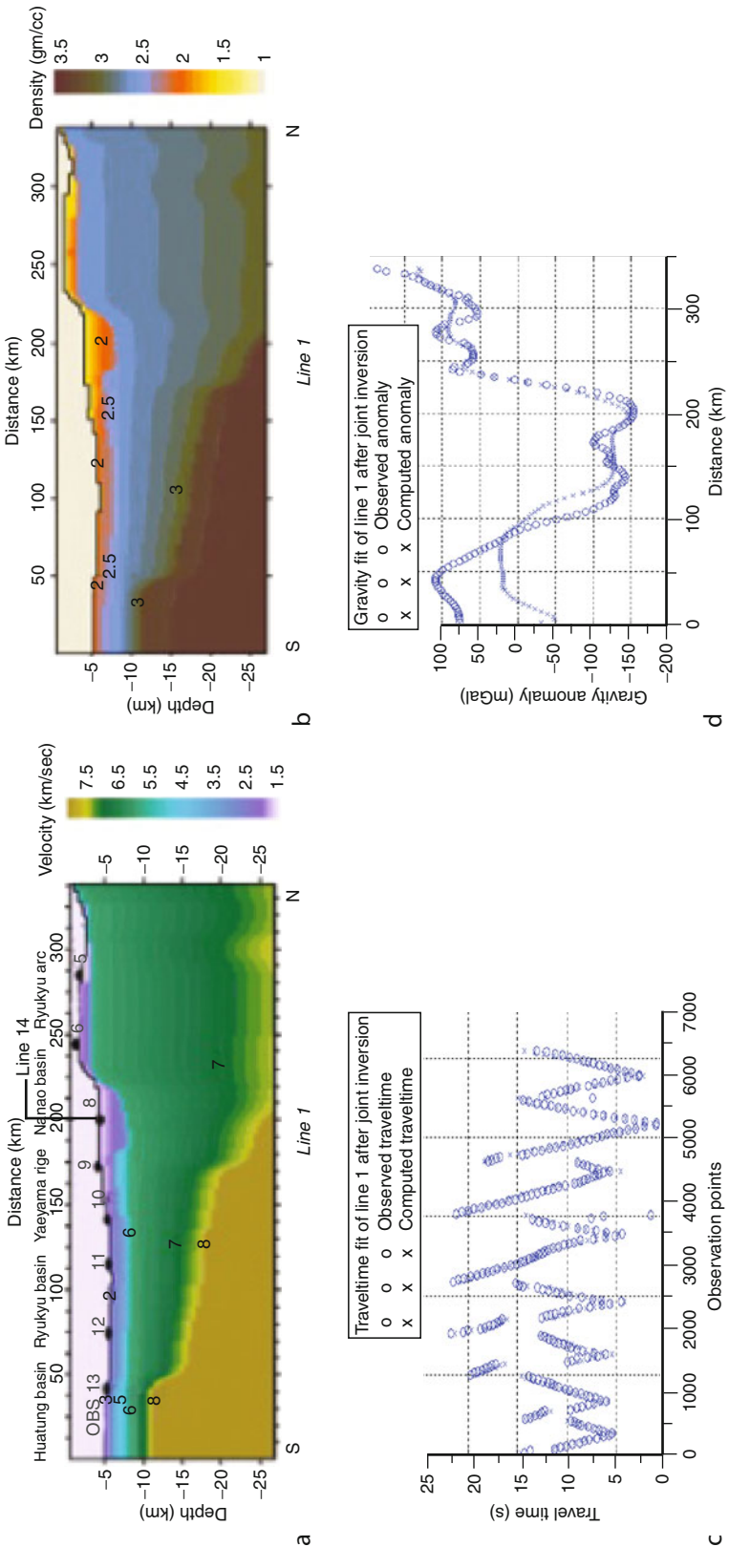
The particle swarm is a stochastic evolutionary computation technique (Kennedy and Kennedy and Eberhart, 1995) used in optimization, which is inspired by observation of social behavior of individuals (called particles) in nature, such as bird flocking and fish schooling. This has been successfully used in many different science and engineering fields including Geophysics (Shaw and Srivastava, 2007). PSO defines a swarm of particles (models) in an M-dimensional space. Each particle maintains memory of its previous best position, $\mathbf{p}_i$ and velocity, $\mathbf{v}_i$. At each iteration, a velocity adjustment of the particle is determined jointly by the previous best position occupied by the particle and the best position of the swarm. The new velocity is then used to compute a new position for the particle. The component of the velocity adjustment determined by the individual's previous best position has been termed as the *cognition*, and the component influenced by the best in the population is the *social* part. The velocity update formula is given by

$$\begin{aligned} \mathbf{v}_i^k &= \mathbf{v}_i^{k-1} + b.ran(.)\left(\mathbf{m}_i^l - \mathbf{m}_i^k\right) + c.ran(.)\left(\mathbf{m}_g - \mathbf{m}_i^k\right) \\ \mathbf{m}_i^{k+1} &= \mathbf{m}_i^k + a\mathbf{v}_i^k, \end{aligned} \tag{6}$$

where $\mathbf{m}_i^k$ is the current location; $\mathbf{v}_i^k$ is the current velocity; $\mathbf{m}_i^l$ is the best location so far; $\mathbf{m}_g$ is the best location achieved by the swarm prior to the kth iteration; *ran*(.) is a random number generator; and a, b, and c are constants. This algorithm is very intuitive and easy to program.

Applications

Several applications of global optimization to geophysical data can now be found in the literature. Due to space limitation, only two examples are included here, which are

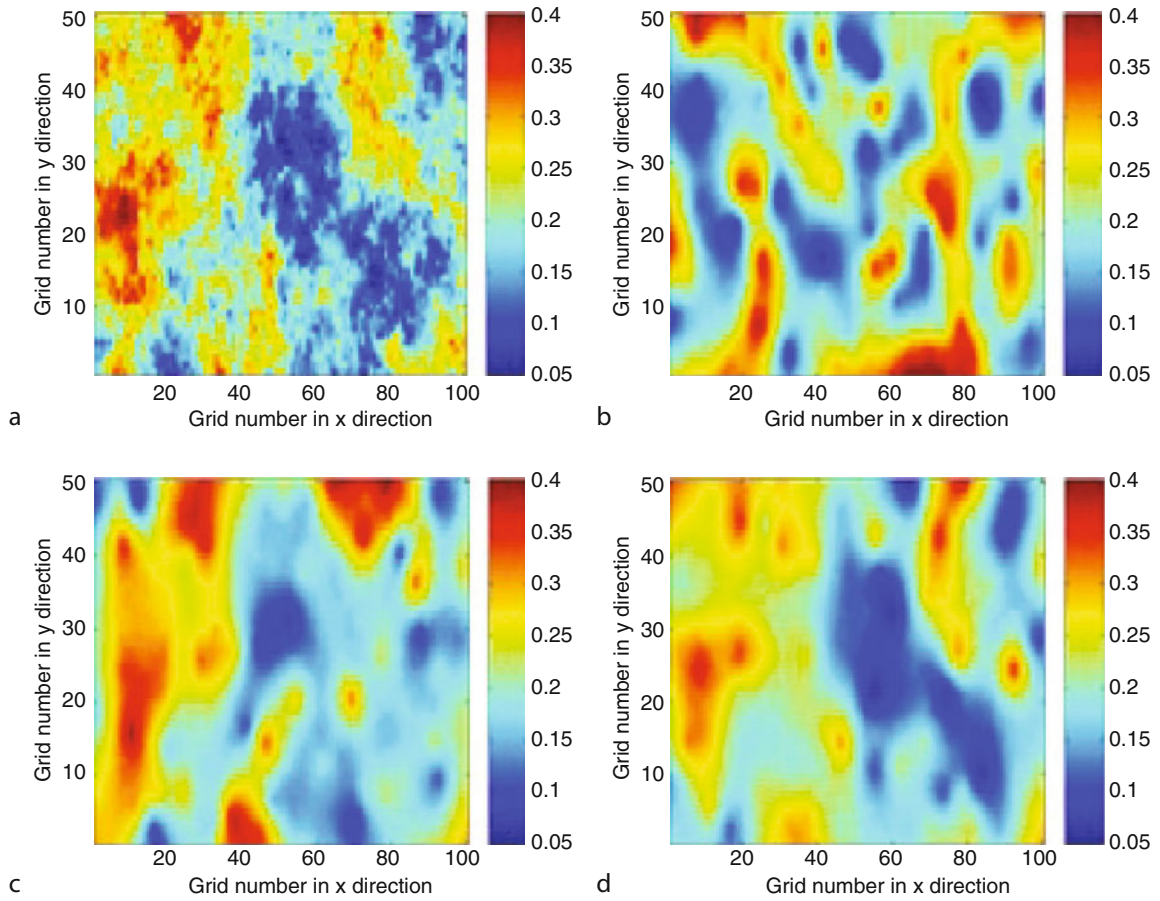


Inverse Theory, Global Optimization, Figure 5 Example of joint inversion of first arrival seismic travel time and gravity data (Modified from Roy et al., 2005): (**a**) the best-fit P-wave velocity (*top*) and (**b**) density (*bottom*) models for a 2D line marked line 1; (**c**) comparison between observed and predicted travel time at different OBS locations and (**d**) comparison between observed and predicted gravity anomalies at different locations (Figure reprinted with permission from IOP Publishing).

particularly intriguing since both of these applications address the issue of joint inversion of disparate datasets.

Roy et al. (2005) implemented a joint inversion of first arrival travel time and gravity data with application to field data from a geologically complex subduction zone. They employed a layer-based model description, in which interfaces (which may also be called iso-velocity lines) were defined by a summation of arc-tangent functions. Within each layer, the velocity is assumed to vary linearly with depth at each surface location. Because of the nonuniqueness of the gravity inversion, they used prior knowledge to relate velocity to density values. The nonlinear optimization problem was solved by a very fast simulated annealing (VFSA) technique. At each iteration, travel times were generated by the solution of the Eikonal equation while the gravity anomalies are computed using a standard formula. The objective function consists of two parts: one measures the misfit in travel time and the other measures the misfit of gravity anomalies. The technique was applied to field data collected over the Ryukyu subduction zone offshore Taiwan during an ocean bottom seismometer (OBS) experiment (called TAICRUST) conducted in the year 1995. An example of results from a profile along NS direction is shown in Figure 5. The left panel in Figure 5 shows the velocity and density models derived by joint inversion of travel time and gravity data. The fit between observed and synthetic data is shown in the right panel of Figure 5.

The second application (taken from Jin et al., 2009) is that of joint inversion of flow and seismic data for reservoir parameters. It is a challenging task in that these disparate datasets are sensitive to different physics and model resolutions for the forward problem. To overcome some of these challenges, Jin et al. (2009) developed a global optimization method based on very fast simulated annealing (VFSA) and a pilot-point-based model parameterization scheme. Reservoir simulation was used to create the saturation and pressure distribution with time. The simulation results were converted to seismic properties using an appropriate rock physics model. Seismic modeling was used to create the seismic response. The objective function was defined as a weighted sum of data misfit and model misfit and VFSA was used to derive optimal model parameters. Results from application to a synthetic dataset



Inverse Theory, Global Optimization, Figure 6 Joint inversion of time-lapse seismic and well production data obtained by pilot-point parameterization and very fast simulated annealing (Modified from Jin et al., 2009):true porosity (**a**); inversion result using only well data (**b**); inversion result using only seismic data (**c**); and inversion result using both seismic data and well data (**d**).

are shown in Figure 6. A comparison of results from using well data alone (top right panel), seismic data alone (bottom left panel), seismic and well data (bottom right) with the true model parameters (top left panel; a 2D porosity slice) shows marked improvement with the use of time-lapse seismic and well production data.

Summary

Most geophysical inverse problems are nonlinear in nature and require searching for the minimum of an objective function that is often multimodal. Although linearization of the forward problems and use of local optimization methods have been fairly successful in many applications, many complex problems show suboptimal results using these methods. The global optimization methods are not sensitive to the choice of the starting models and make no assumption on the shape of the objective function. These methods, however, require numerous forward model evaluations and are therefore, computationally expensive. With the advent of cheap fast computers with large memory, applications of these methods are becoming increasingly popular. In some applications, hybrid methods based on local and global optimization (e.g., Chunduru et al., 1997; Porsani et al., 2000) can be designed that can take advantage of both the approaches.

Bibliography

Aster, R. C., Brochers, B., and Thurber, C. H., 2005. *Parameter Estimation and Inverse Problems*. London: Elsevier.

Chunduru, R. K., Sen, M. K., and Stoffa, P. L., 1997. Hybrid optimization methods for geophysical inversion. *Geophysics*, **62**(4), 1196–1207.

Gill, P. E., Murray, W., and Wright, M. H., 1981. *Practical Optimization*. London: Academic.

Goldberg, D. E., 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading: Addison Wesley.

Ingber, L., 1989. Very fast simulated annealing. *Mathematical and Computer Modelling*, **12**(8), 967–993.

Jackson, D. D., 1972. Interpretation of inaccurate, insufficient and inconsistent data. *Geophysical Journal of the Royal Astronomical Society*, **28**, 97–100.

Jin, L., Stoffa, P. L., Sen, M. K., and Seif, R., 2009. Pilot point parameterization in stochastic inversion for reservoir properties using time-lapse seismic and production data. *Journal of Seismic Exploration*, **18**, 1–20.

Kennedy, J., and Eberhart, R., 1995. Particle swarm optimization. *Proceedings of the IEEE international conference on neural networks*, **IV**, 1942–1948.

Kirkpatrick, S., Gelatt, C. D., Jr., and Vecchi, M. P., 1983. Optimization by simulated annealing. *Science*, **220**, 671–680.

Menke, W., 1984. *Geophysical Data Analysis: Discrete Inverse Theory*. New York: Academic.

Porsani, M. J., Stoffa, P. L., Sen, M. K., and Chunduru, R. K., 2000. Fitness functions, genetic algorithms and hybrid optimization in seismic waveform inversion. *Journal of Seismic Exploration*, **9**(2), 143–164.

Roy, L., Sen, M. K., Stoffa, P. L., McIntosh, K., and Nakamura, Y., 2005. Joint inversion of first arrival travel time and gravity data. *Journal of Geophysics and Engineering*, **2**, 277–289.

Sambridge, M., 1999. Geophysical inversion with a neighbourhood algorithm -I. Searching a parameter space. *Geophysical Journal International*, **138**, 479–494.

Sambridge, M., and Mosegaard, K., 2002. Monte Carlo Methods in geophysical inverse problems, Reviews of Geophysics, 40(3), article no. 1009.

Sen, M. K., and Stoffa, P. L., 1995. *Global Optimization Methods in Geophysical Inversion*. Amsterdam: Elsevier.

Sen, M. K., 2006. Seismic inversion, Society of Petroleum Engineers Press.

Shaw, R. K., and Srivastava, S., 2007. Particle swarm optimization: a new tool to invert geophysical data. *Geophysics*, **72**(2), V75–F83.

Stoffa, P. L., and Sen, M. K., 1991. Nonlinear multiparameter optimization using genetic algorithms: Inversion of plane wave seismograms. *Geophysics*, **56**(11), 1794–1810.

Tarantola, A., 1987. *Inverse Problem Theory, Methods of Data Fitting and Model Parameter Estimation*. Amsterdam: Elsevier.

Cross-references

Inverse Theory, Artificial Neural Networks
Inverse Theory, Linear
Inverse Theory, Monte Carlo Method
Inverse Theory, Singular Value Decomposition

INVERSE THEORY, LINEAR

Pravin K. Gupta
Department of Earth Sciences, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India

Introduction

The activity of solving inverse problems is all pervading. We perform this activity throughout our life right since birth. As a new born baby the first inverse problem we solve, albeit through pattern recognition, is identification of mother in a group of persons. In this vein, the geophysical inverse problems are just a member of the class of inverse problems encountered and studied in science. Formally, the inverse problem can be defined as an estimation of system parameters together with their uncertainties from the observed system response to a given source excitation. In contrast, the forward problem is defined as computation of response of the system for a given model of its properties and for a given source excitation. Most of the geophysical inverse problems are nonlinear by virtue of the system response being a nonlinear function of system parameters. However, a widely employed class of methodologies solves the nonlinear inverse problems through quasi-linearization whereby the nonlinear problem is recast into a sequence of linear problems that are solved iteratively using linear methodologies. Further, since bulk of the nonlinearity signatures in a geophysical inverse problem arise from geometry of the distribution of physical property, the problem can be recast, in some cases, to a linear problem when the model geometry is considered known. For example, in gravity studies, if the domain of study is divided into small cubes of known volumes and the unknowns are only the density of cubes, the corresponding inverse problem is linear. Similarly, in case of a layered earth model, the problem can be cast into

a linear one when the layer thickness is assumed to be known according to some criterion, for example, constant travel time in case of seismics, constant thickness in DC resistivity, or constant skin depth in case of magnetotellurics (Kunetz, 1972; Gupta et al., 1996, 1997). So, there exists an enormous interest in methodologies for solving the linear inverse problems. This article deals with general aspects of linear inverse theory, the description of well- and ill-posed problems, the least square, minimum norm, and generalized inverses of a matrix, the data and parameter resolution matrices, and the ways of handling noise in data. Lastly, the seminal method proposed by Backus and Gilbert (1967, 1968, 1970) for solving linear inverse problems is discussed in some detail. For rigorous readings on inversion theory one can refer the books by Aki and Richards (2002), Indira and Gupta (1998), Lines and Treitel (1984), Menke (1984), Scales et al. (2001), Sneider and Trampert (1999), Tarantola (2005), Tikhanov and Arsenin (1977), Twomey (1977), and Zhdanov (2002), while for numerical aspects of inverse problems one can refer the books by Bjorck (1996), Lawson and Hanson (1995), and Rao and Bhimasankaram (1992) and also the works of Rojas and Sorenson (2002) and Saltelli et al. (2006). For applications and case studies, one can refer Aki and Richards (2002) and Tarantola (2005) for seismological inverse problems, Borcea (2002) for electrical impedance tomography, Habashy and Abubakar (2004) for electromagnetic inversion, Mallat (2009) for signal processing, and Rawlinson et al. (2010) for seismic tomography. For general geophysical applications one can refer to Treitel and Lines (2001) and Oldenburg and Li (2004).

Linear inverse problems

The common linear inverse problems can be presented in the form of Fredholm's integral equations of first kind,

$$g(y) = \int_a^b K(x,y) f(x) \mathrm{d}x, \quad c < y < d. \tag{1}$$

Here, $K(x, y)$ is the kernel function that transforms the system parameter function, $f(x)$, into the observable function $g(y)$. A special class of linear problems is the convolution problem where the kernel function has the special form $K(y–x)$.

The problem of obtaining a function, $f(t)$, from its Fourier transform function, $F(\omega)$, can be looked upon as an inverse problem as stated below

$$F(\omega) = \int_{-\infty}^{\infty} f(t) \mathrm{e}^{-i\omega t} \mathrm{d}t. \tag{2}$$

Oldenburg (1976) studied this linear inverse problem using Backus–Gilbert method.

Another example, the gravity anomaly problem can be stated as

$$g_z(\mathbf{r}) = \iiint_{\mathbf{v}} \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} d^3\mathbf{r}'. \tag{3}$$

Here, $\rho(\mathbf{r}')$ is the unknown density distribution and the remaining factor defines the kernel function. The anomaly $g_z(\mathbf{r})$ is a linear function of density $\rho(\mathbf{r}')$. It may be noted that it is a convolution type problem.

Similarly, in DC resistivity method, the apparent resistivity function, $\rho_a(r)$, is linearly related to the resistivity transform function, $T(\lambda)$, through the kernel function as

$$\rho_a(r) = \int_0^{\infty} T(\lambda) J_n(\lambda r) d\lambda. \tag{4}$$

Ghosh (1971a, b) transformed this problem to a convolution problem by transformation of independent variables.

The discrete equivalent of linear problems can be cast in the form of a matrix equation as

$$\mathbf{Ax} = \mathbf{b}. \tag{5}$$

Here, $\boldsymbol{A}$ is an $m \times n$ matrix, $\boldsymbol{b}$ is the m-dimensional observation data vector and $\boldsymbol{x}$ is the n-dimensional parameter variable vector to be determined.

The traditional solution of this problem exists only for the case where $m = n$ and the coefficient matrix is nonsingular so that its exact inverse exists. However, in such special cases also there exist serious problems. Consider the simple example

$$x + y = 2$$
$$x + 1.0001y = 2.0001$$

The solution of this problem is $x = y = 1$. However, if the right-hand side vector is an outcome of some observations where accuracy is only up to two decimal places, statement of the problem may become

$$x + y = 2$$
$$x + 1.0001y = 2.00$$

The solution will now be $x = 2$, $y = 0$. Such drastic change in the solution vector for a very small change in data vector is a reflection of the problem being an *unstable* one. Similarly, if our coefficient is an outcome of some computations where the accuracy is only up to two decimal places, assuming the data vector to be error free, the problem becomes

$$x + y = 2$$
$$x + 1.00y = 2.0001$$

In this case we find that the two equations are *inconsistent*.

Lastly, if in the previous case the data is also erroneous, the resulting equations are

$$x + y = 2$$
$$x + 1.00y = 2.00$$

Now there exist infinite solutions and the problem is *nonunique*.

These problems of instability, inconsistency and nonuniqueness invariably crop up while solving inverse problems. If a problem has any one of these features it is termed as an ill-posed problem. Hadamard (1902) formalized the definition of a well-posed problem in terms of the three conditions:

1. The solution exists.
2. The solution is unique.
3. The solution is stable.

Since in applied mathematics one is interested in obtaining the solution, another condition was subsequently added to this list of conditions for a problem to be termed as well posed, that is,

4. The solution is constructible.

A linear or linearized inverse problem, when cast in a matrix form, can be solved as a least square problem. Here, we shall be discussing various ways of solving the matrix equation in its generality. The various classifications of matrix equation 5 can be

1. Case when $m > n = k$, is known as strictly overdetermined.
2. Case when $n > m = k$, is known as strictly underdetermined.
3. Case when $m = n = k$, is known as evenly determined.
4. Case when $k < min(m,n)$ is known as underdetermined.

Here, k is the rank of matrix $\boldsymbol{A}$.

In any real problem the observed data is, in general, inadequate and inaccurate. As a result the resulting inverse problem is inconsistent, underdetermined, and most of the time unstable. Under such adverse conditions when the exact solution does not exist, the basic task in solving the matrix Equation 5 pertains to obtaining a solution that meets the desired objectives. Backus (1970, 1996), Backus and Gilbert (1967, 1968, 1970), Jackson (1972, 1979), and others studied these problems and developed methodologies to obtain a unique solution. It is inevitable that in such cases there would be a trade-off between the propagation of data error to parameter estimation and the parameter resolution. The inevitable nonuniqueness is removed by incorporating the available a priori information about the unknown parameter vector $\boldsymbol{x}$. The various methods differ in handling the a priori knowledge about data error and model and in controlling the trade-off between error propagation and parameter resolution.

To be able to solve Equation 5 one designs a matrix $\boldsymbol{A}^{\boldsymbol{g}}$, termed as generalized inverse, such that

$$\mathbf{x} = A^{\mathbf{g}}\mathbf{b}. \tag{6}$$

It is obvious that matrix $\boldsymbol{A}^{\boldsymbol{g}}$ will not possess all the properties of the exact matrix inverse, $\boldsymbol{A}^{-1}$, which can be defined only for non-singular matrix. For various cases mentioned above, the generalized inverse will possess only a subset of these properties. One of the important properties of exact inverse is $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$. In case of a matrix $\boldsymbol{A}$ of order $m \times n$, the generalized inverse matrix, $\boldsymbol{A}^{\boldsymbol{g}}$, will lead to the product matrices, $\boldsymbol{S} = \boldsymbol{A}\boldsymbol{A}^{\boldsymbol{g}}$ of order $m \times m$ and $\boldsymbol{R} = \boldsymbol{A}^{\boldsymbol{g}}\boldsymbol{A}$ of order $n \times n$ and these, in general, will not be identity matrices. The former product matrix $\boldsymbol{S}$ is termed as *data resolution matrix* or *information density matrix*, while the latter product matrix $\boldsymbol{R}$ is known as *parameter resolution matrix*. These product matrices play an important role in the evaluation of quality of solution. Norms of the deviation of matrices $\boldsymbol{S}$ and $\boldsymbol{R}$ from respective Identity matrices provide quantitative measures of quality of information contained in data vector and of the degree of parameter resolution, respectively.

The matrix $\boldsymbol{A}^{\boldsymbol{g}}$ can be synthesized in several ways for a given type of coefficient matrix $\boldsymbol{A}$. The important approaches widely employed to solve the Equation 5 lead to

1. Least square inverse
2. Minimum norm inverse
3. Damped least square (DLS) inverse
4. Rao–Mitra generalized inverse
5. Moore–Penrose generalized inverse based on singular value decomposition (SVD)

All these approaches are briefly presented below followed by Backus–Gilbert method which also provides a solution of the linear inverse problem stated as Equation 1.

Generalized inverses

Least square inverse

The least square (LS) solution of a strictly overdetermined problem was first proposed by Gauss in 1795. The LS solution exists even for inconsistent system of equations. It is obtained by minimizing the L2 norm of misfit error vector, that is,

$$\left\{\begin{array}{c} \text{Minimize } \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} = (\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}) \\ \text{Solution: } \boldsymbol{x} = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{b} \\ \text{Inverse: } \boldsymbol{A}^{\boldsymbol{g}} = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}} \\ \text{Information density matrix: } \boldsymbol{S} = \boldsymbol{A}\boldsymbol{A}^{\boldsymbol{g}} \\ = \boldsymbol{A}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}} \neq \boldsymbol{I}_{\boldsymbol{m}} \\ \text{Parameter resolution matrix: } \boldsymbol{R} = \boldsymbol{A}^{\boldsymbol{g}}\boldsymbol{A} \\ = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = I_n \end{array}\right\} \tag{6}$$

The expressions of matrices $\boldsymbol{S}$ and $\boldsymbol{R}$ reveal that the data quality is not optimal while all the parameters are resolved exactly.

Minimum norm inverse

The minimum norm (MN) solution of a strictly underdetermined problem is obtained through constrained minimization of the L2 norm of the unknown vector $\boldsymbol{x}$.

$$\left\{\begin{array}{c}\text{Minimize } \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\\ \text{Subject to the constraint } \boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}=\boldsymbol{0}\\ \text{Solution: } \boldsymbol{x}=\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}\boldsymbol{b}\\ \text{Inverse: } \boldsymbol{A}^{\boldsymbol{g}}=\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}\\ \text{Information density Matrix: } \boldsymbol{S}=\boldsymbol{A}\boldsymbol{A}^{\boldsymbol{g}}\\ =\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}=\boldsymbol{I}_{\boldsymbol{m}}\\ \text{Parameter resolution matrix: } \boldsymbol{R}=\boldsymbol{A}^{\boldsymbol{g}}\boldsymbol{A}\\ =\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}\boldsymbol{A}\neq\boldsymbol{I}_{\boldsymbol{n}}\end{array}\right\} \tag{7}$$

The expressions of matrices $\boldsymbol{S}$ and $\boldsymbol{R}$ suggest that the data quality is optimal while all the parameters are not resolved exactly and there exists linear dependence amongst some components of the parameter vector.

Damped least square inverse

The DLS solution of an overdetermined problem when rank $k<min(m,n)$ was concurrently developed by Marquardt, Phillips, Tikhanov, and Twomey. It is obtained by minimizing the two objective functions defining norms of error $\boldsymbol{e}$ and of unknown vector $\boldsymbol{x}$, subject to the constraint of $\boldsymbol{x}$ satisfying the matrix equation 5.

$$\left\{\begin{array}{c}\text{Minimize } \boldsymbol{e}^{T}\boldsymbol{e}=(\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}) \text{ and } \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\\ \text{Subject to the constraint } \boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}=\boldsymbol{0}\\ \text{Solution: } \boldsymbol{x}=(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}+\lambda^{2}\boldsymbol{I})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{b}\\ =\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}+\mu^{2}\boldsymbol{I})^{-1}\boldsymbol{b}\\ \text{Inverse: } \boldsymbol{A}^{\boldsymbol{g}}=(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}+\lambda^{2}\boldsymbol{I})^{-1}\boldsymbol{A}^{\mathrm{T}}\\ \text{or} \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}+\mu^{2}\boldsymbol{I})^{-1}\\ \text{Information density matrix: } \boldsymbol{S}=\boldsymbol{A}\boldsymbol{A}^{\boldsymbol{g}}\neq\boldsymbol{I}_{\boldsymbol{m}}\\ \text{Parameter resolution matrix: } \boldsymbol{R}=\boldsymbol{A}^{g}\boldsymbol{A}\neq\boldsymbol{I}_{\boldsymbol{n}}\end{array}\right\} \tag{8}$$

Here, the parameter λ is a control parameter that represents relative weighting given to the two objective functions. The expressions of matrices $\boldsymbol{S}$ and $\boldsymbol{R}$ suggest that neither the data quality nor resolution of parameters is optimal.

Weighted damped least square inverse

To incorporate any available information about the error characteristics of data and/or about smoothness characteristics of parameter vector, the two objective functions optimized in case of DLS inverse, can be redefined in terms of weighting matrices $\boldsymbol{W}_{\boldsymbol{e}}$ and $\boldsymbol{W}_{\boldsymbol{m}}$. $\boldsymbol{W}_{\boldsymbol{e}}$ is the data error covariance matrix and $\boldsymbol{W}_{\boldsymbol{m}}$ is the matrix defining smoothness characteristics of the desired model. $\boldsymbol{W}_{\boldsymbol{m}}$ can be constructed from the low order differences of parameter vector components.

$$\left\{\begin{array}{c}\text{Minimize } \boldsymbol{e}^{\mathrm{T}}\boldsymbol{W}_{\boldsymbol{e}}=(\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x})^{\mathrm{T}}\\ \boldsymbol{W}_{\boldsymbol{e}}(\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}) \text{ and } \boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}_{\boldsymbol{m}}\boldsymbol{x}\\ \text{Subject to the constraint } \boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}=\boldsymbol{0}\\ \text{Solution: } \boldsymbol{x}=(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}_{\boldsymbol{e}}\boldsymbol{A}+\lambda^{2}\boldsymbol{W}_{\boldsymbol{m}})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}_{\boldsymbol{e}}\boldsymbol{b} \text{ or}\\ \boldsymbol{x}=\boldsymbol{W}_{\boldsymbol{m}}^{-1}\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{W}_{\boldsymbol{m}}^{-1}\boldsymbol{A}^{\mathrm{T}}+\lambda^{2}\boldsymbol{W}_{\boldsymbol{e}}^{-1})^{-1}\boldsymbol{b}\\ \text{Inverse: } \boldsymbol{A}^{\boldsymbol{g}}=(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}_{\boldsymbol{e}}\boldsymbol{A}+\lambda^{2}\boldsymbol{W}_{\boldsymbol{m}})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}_{\boldsymbol{e}} \text{ or}\\ \boldsymbol{A}^{g}=\boldsymbol{W}_{\boldsymbol{m}}^{-1}\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{W}_{\boldsymbol{m}}^{-1}\boldsymbol{A}^{\mathrm{T}}+\lambda^{2}\boldsymbol{W}_{\boldsymbol{e}}^{-1})^{-1}\\ \text{Information density matrix : } \boldsymbol{S}=\boldsymbol{A}\boldsymbol{A}^{\boldsymbol{g}}\neq\boldsymbol{I}_{\boldsymbol{m}}\\ \text{Parameter resolution matrix : } \boldsymbol{R}=\boldsymbol{A}^{\boldsymbol{g}}\boldsymbol{A}\neq\boldsymbol{I}_{\boldsymbol{n}}\end{array}\right\} \tag{9}$$

All the above inverses can be considered as special cases of the generalized inverse defined by Rao and Mitra (1971) which is discussed next.

Rao–Mitra inverse

Rao and Mitra (1971) put forth a definition of generalized inverse, $\boldsymbol{A}^{\boldsymbol{g}}$, as any matrix that provides a solution of matrix equation 5 as

$$\boldsymbol{x}=\boldsymbol{A}^{\boldsymbol{g}}\boldsymbol{b}. \tag{10}$$

It is evident that this inverse satisfies the relation

$$\boldsymbol{A}\boldsymbol{A}^{\boldsymbol{g}}\boldsymbol{A}=\boldsymbol{A}. \tag{11}$$

This condition leads to two more properties

$\mathbf{A}\mathbf{A}^{\mathbf{g}}$ is idempotent,

$\mathbf{A}^{\mathbf{g}}\mathbf{A}$ is idempotent.

It may be mentioned that idempotent matrices are those for which any power of the matrix is equal to itself.

Rao and Bhimasankaram (1992) presented a suit of algorithms for evaluating the generalized inverse using Gaussian elimination like steps based on elementary row and column operations that reduce the coefficient matrix to a Hermite canonical form.

An earlier definition of generalized inverse, $\boldsymbol{A}^{+}$, given by Moore (1920) and rediscovered by Penrose (1955), was based on SVD of matrix $\boldsymbol{A}$. It satisfies the conditions

$$\left\{\begin{array}{l}\boldsymbol{A}\boldsymbol{A}^{+}\boldsymbol{A}=biA\\ \boldsymbol{A}^{+}\boldsymbol{A}\boldsymbol{A}^{+}=\boldsymbol{A}^{+}\\ \boldsymbol{A}\boldsymbol{A}^{+} \text{ is idempotent}\\ \boldsymbol{A}^{+}\boldsymbol{A} \text{ is idempotent}\end{array}\right\} \tag{12}$$

Moore–Penrose inverse (MPI) is always unique while Rao–Mitra inverse (RMI) is, in general, not. In fact, MPI is one member of the set of all possible RMIs. To elaborate this point, it may be stated that for a given SVD of matrix $\boldsymbol{A}$,

$$\boldsymbol{A}=\boldsymbol{U}\Lambda\boldsymbol{V}^{\mathrm{T}}. \tag{13}$$

Here $\boldsymbol{U}$ and $\boldsymbol{V}$ are the modal matrices of orders $m\times m$ and $n\times n$, respectively and Λ is an $m\times n$ matrix whose leading k diagonal elements correspond to the nonzero singular values and all other elements are zero.

The MPI is given by

$$\boldsymbol{A}^{+} = \boldsymbol{V}\Lambda^{+}\boldsymbol{U}^{\mathrm{T}} \tag{14}$$

Here Λ^{+} is the $n \times m$ matrix whose first k diagonal elements correspond to the inverse of nonzero singular values and all other elements are zero.

In contrast, the RMI is given by

$$\boldsymbol{A}^{\boldsymbol{g}} = \boldsymbol{V}\Lambda^{\boldsymbol{g}}\boldsymbol{U}^{\mathrm{T}}. \tag{15}$$

Here matrix $\Lambda^{\boldsymbol{g}}$ is an $n \times m$ matrix whose leading $k \times k$ submatrix block is diagonal with nonzero diagonal elements corresponding to the inverse of nonzero singular values and all the other submatrices are arbitrary. This freedom of having arbitrary elements outside the main $k \times k$ block leads to the nonuniqueness of MRI. This, in turn, leads to nonunique solutions of matrix Equation 5, which can be derived from a given solution, $\boldsymbol{x}^{(1)}$, of Equation 5 as

$$\boldsymbol{x} = \boldsymbol{x}^{(1)} + (\boldsymbol{A}^{\boldsymbol{g}}\boldsymbol{A} - \boldsymbol{I})\boldsymbol{\xi} \tag{16}$$

$\boldsymbol{\xi}$ being an arbitrary vector. One possible vector $\boldsymbol{x}^{(1)}$ can be $\boldsymbol{x}^{+}$ the solution obtained using MPI.

Backus–Gilbert method

Backus and Gilbert (1967, 1968, 1970) studied in detail the problem of synthesizing solutions to a linear inverse problem expressible as a Fredholm's equation of first kind like Equation 1. The main feature of this methodology is the possible detailed study of the trade-off between parameter resolution and error propagation. The Backus–Gilbert method (BGM) is briefly described here both for continuous and discrete cases.

Continuous inverse problem

The solution of Equation 1 can be represented as

$$f(x) = \int_{c}^{d} G(x,y)g(y)dy, \quad a < x < b \tag{17}$$

For Equations 1 and 17 to coexist, it is essential that the functions $K(x,y)$ and $G(x,y)$ satisfy the relations

$$\int_{a}^{b} G(x,y)K(x,y')dx = \delta(y - y'), \tag{18}$$

and

$$\int_{c}^{d} G(x,y)K(x',y)dy = \delta(x - x'). \tag{19}$$

The function $\delta(r-r')$ is the transcendental Dirac-delta function.

Backus and Gilbert brought into focus the infinitely undetermined nature of the real continuous linear inverse problems represented by Equation 1. They emphasized that from the finite observations, it would never be possible to exactly determine the infinite unknowns constituting the sought model property function $f(x)$. They circumvented the problem by seeking not the exact solution but its approximation as a linear superposition of the observed data values.

Let the finite observation set comprise m values $g_i = g(y_i)$ with $c \le y_1 \le y_2 \le \ldots \le y_{m-1} \le y_m \le d$. Further let the unknown function $f(x)$ need be estimated at the n argument values $a \le x_1 \le x_2 \le \ldots \le x_{n-1} \le x_n \le b$. Now Equations 1 and 17 can be rewritten as

$$g_i = g(y_i) = \int_{a}^{b} K_i(x)f(x)dx, \quad \forall 1 < i < m \tag{20}$$

and

$$f_j = f(x_j) = \int_{c}^{d} G_j(y)g(y)dy, \quad \forall 1 < j < n \tag{21}$$

where

$$K_i(x) = K(x, y_i) \text{ and } G_j(y) = G(x_j, y). \tag{22}$$

Backus and Gilbert sought an approximation of Equation 21 as

$$f_j \cong \hat{f}_j = f(x_j) = \sum_{1}^{m} c_i(x_j)g_i \tag{23}$$

Substituting for g_i from Equation 20, we get

$$\hat{f}_j = \sum_{1}^{m} c_i(x_j) \int_{a}^{b} K_i(x)f(x)dx$$

For K_i and f being square integrable, the order of integration and summation can be interchanged and the equation can be rewritten as

$$\hat{f}_j = \int_{a}^{b} \left[\sum_{1}^{m} c_i(x_j)K_i(x)\right] f(x)dx = \int_{a}^{b} A(x,x_j)f(x)dx \tag{24}$$

where

$$A(x,x_j) = \sum_{1}^{m} c_i(x_j)K_i(x), \tag{25}$$

is termed by Backus and Gilbert as averaging kernel and is also known in literature as resolving kernel or scanning function.

From Equation 24, it is evident that for f_j to be exactly estimated, the averaging kernel should be the Dirac-delta

function $\delta(x–x_j)$. This equality is, in general, not achievable. In order to make $A(x,x_j)$ as close to $\delta(x–x_j)$ as possible, the coefficients c_i's are chosen to satisfy a judiciously chosen δ-ness criterion. One such criterion can be

$$q_j = q(x_j) = \int_a^b J(x,x_j)\left[A(x,x_j) - D(x,x_j)\right]^2 dx. \quad (26)$$

Several pairs of functions J,D have been used by different workers. Backus and Gilbert used $J(x,x_j) = 12(x–x_j)^2$ and $D(x,x_j) = \delta(x–x_j)$. They interpreted q_j as the half-width or spread of the resolving kernel A. Hence, c_i's are so chosen that these minimize q_j subject to the condition that A is unimodular, that is,

$$\int_a^b A(x,x_j)dx = 1 \quad (27)$$

Expressing Equations 24, 26, and 27 in matrix form, we get

$$\left\{\begin{array}{l} \hat{f}_j = \mathbf{c^T g} \\ q_j = \mathbf{c^T M c} + \mathbf{a}^{\mathrm{T}}\mathbf{c} + b \\ \mathbf{u}^{\mathrm{T}}\mathbf{c} = 1 \end{array}\right\} \quad (28)$$

with

$$M_{ij} = \int_a^b K_i(x)K_j(x)dx$$

$$a_i = -2\int_a^b K_i(x)J(x,x_j)D(x,x_j)dx$$

$$b = \int_a^b D^2(x,x_j)dx$$

$$u_i = \int_a^b K_i(x)dx$$

Using Lagrange's method of undetermined multipliers, the solution of Equation 28 is found to be

$$\boldsymbol{c} = \frac{\boldsymbol{M}^{-1}\boldsymbol{u}}{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{u}} \quad (29)$$

Substituting these c-values in Equation 23, one gets the estimate of unknown physical property function f_j at argument value x_j. All values $f_1,f_2,\ldots,f_n$ are obtained by repeating the above steps for each $x_1,x_2,\ldots,x_n$.

The quality of solution can be evaluated by analyzing the characteristics of the averaging kernel $A(x,x_j)$, which ideally should be the Dirac-delta function. Hence, the deviation of the center of $A(x,x_j)$ from x_j and the finite range of x over which $A(x,x_j)$ is nonzero may be taken as pointers to the quality of estimation.

The discussion till now was confined to error-free data. However, if the data is erroneous and its error covariance matrix is $\boldsymbol{E}$, the procedure need to be modified to get a solution. In such cases, the superposition coefficient vector $\boldsymbol{c}$ is obtained by minimizing besides q_j given by Equation 26 and the error propagation measure e_j given by

$$e_j = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{c} \quad (30)$$

The two objective functions are combined into one as

$$r_j = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{N}\boldsymbol{c},$$

where

$$\boldsymbol{N} = \lambda\boldsymbol{M} + (1-\lambda)\boldsymbol{E}, 0 \le \lambda \le 1. \quad (31)$$

The procedure described above can be followed after replacing matrix $\boldsymbol{M}$ by $\boldsymbol{N}$ and the resulting estimate of coefficient vector will then be

$$\boldsymbol{c} = \frac{\boldsymbol{N}^{-1}\boldsymbol{u}}{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{N}^{-1}\boldsymbol{u}}. \quad (32)$$

Equations 20 and 21 make it obvious that for $\lambda = 1$, spread gets minimized but error propagation is uncontrolled while for $\lambda = 0$, the spread is unlimited and error propagation is minimized.

Discrete inverse problem

The implementation of BGM for the discrete case represented by Equation 5 was given by Gupta (1998) where it is shown that, for error-free case, if the unimodularity constraint is not applied the BG estimate is same as the minimum norm estimate. However, when this constraint is applied the BG estimate has an additional correction term.

The solution of matrix equation 5 is synthesized as

$$\boldsymbol{x} = \boldsymbol{A^g b}. \quad (33)$$

To synthesize $\boldsymbol{A^g}$ some norm of $(\boldsymbol{A^g A – I})$ is minimized. In particular, for L2 norm, we can minimize

$$\boldsymbol{u}^{\mathrm{T}}(\{\boldsymbol{A^g A} - \boldsymbol{I})(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A^{g^{\mathrm{T}}}} - \boldsymbol{I})\boldsymbol{u}.$$

Here $\boldsymbol{u}^{\mathrm{T}} = \mathbf{(1\ 1\ \ldots\ 1)}$. Minimization yields

$$\boldsymbol{A^g} = \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}. \quad (34)$$

This will lead to the minimum norm solution $\boldsymbol{x_{min}}$.

When the unimodularity condition, asking for each row vector of the resolution matrix $\boldsymbol{R = A^g A = A^T(AA^T)A}$ to be of unit length, is imposed as constraint, the solution $\boldsymbol{x_{BG}}$ is found to be given as

$$\boldsymbol{x_{BG}} = \boldsymbol{x_{min}} + \frac{\boldsymbol{u} - \mathbf{Ru}}{\boldsymbol{u}^{\mathrm{T}}\mathbf{Ru}}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{m_{min}}) \quad (35)$$

Trade-off between resolution and error propagation

In Backus–Gilbert (BG) method Equation 31 demonstrates the trade-off between the two objective functions. For different values of λ one gets different solutions. Hence, it is expected that some value of λ will provide optimal trade-off solution. BG demonstrated in their paper of 1970 that when the values of objective functions q_j and r_j for each λ are plotted against each other, one gets an L-shaped curve. This curve reveals that for large value of λ, error propagation is uncontrolled and the spread is optimized while for small value of λ, error propagation is optimized and the spread is uncontrolled. The tip of the L-curve provides the optimal value of λ.

In case of DLS method, the same exercise can be carried out to obtain the optimal value of λ. Lot of research has been put in to devise the strategies to find optimal λ values. Most of these works analyze the dependence of solution on λ through eigenanalysis. For details, one can refer Calvetti et al. (2000), Hansen (1992), Reichel and Sadok (2008), and Rezghi and Hosseini (2009).

Summary

This article presents salient aspects of linear inverse problems. The LS formulations and the Backus–Gilbert formulation are discussed in brief. The information density and parameter resolution matrices and their importance are discussed. Finally, the methods for finding the optimal multi-objective optimization parameter λ are briefly mentioned. For more details on SVD one can refer to the article by A. Manglik, while for global optimization techniques for nonlinear inverse problems one can refer to the articles by M. K. Sen and by W. Sandham published in this series.

Bibliography

Aki, K., and Richards, P. G., 2002. *Quantitative Seismology*. Sausalito: University Science Book.

Backus, G. E., 1970. Inference from inadequate and inaccurate data. *Proceedings of the National Academic Science*, **65**, 1–7, 281–287; **67**, 282–289.

Backus, G. E., 1996. Trimming and procrastination as inversion techniques. *Physics of the Earth and Planetary Interiors*, **98**, 101–142.

Backus, G. E., and Gilbert, J. F., 1967. Numerical applications of a formalism for geophysical inverse problems. *Geophysical Journal of the Royal Astronomical Society*, **13**, 247–273.

Backus, G., and Gilbert, J. F., 1968. Numerical applications of a formalism for geophysical inverse problems. *Geophysical Journal of the Royal Astronomical Society*, **16**, 169–205.

Backus, G., and Gilbert, J. F., 1970. Uniqueness in the inversion of inaccurate gross earth data. *Philosophical Transactions of the Royal Society, London A*, **266**, 123–192.

Bjorck, A., 1996. *Numerical Methods for Least Squares Problems*. Philadelphia: SIAM.

Borcea, L., 2002. Electrical impedance tomography – topical review. *Inverse Problems*, **18**(6), R99–R136.

Calvetti, D., Morigi, S., Reichel, L., and Sgallari, F., 2000. Tikhonov regularization and the L-curve for large discrete ill-posed problems. *Journal of Computational and Applied Mathematics*, **123**, 423–446.

Ghosh, D. P., 1971a. The application of linear filter theory to the direct interpretation of geoelectrical resistivity sounding measurements. *Geophysical Prospecting*, **19**, 192–217.

Ghosh, D. P., 1971b. Inverse filter coefficients for the computation of resistivity standard curves for a horizontally stratified earth. *Geophysical Prospecting*, **19**, 769–775.

Gupta, P. K., 1998. The Backus Gilbert method. In Indira, N. K., and Gupta, P. K. (eds.), Inverse Methods: General Principles and Applications in Earth System Sciences. Narosa Publishing House, New Delhi, Chap. 7, pp. 60–68.

Gupta, P. K., Sri Niwas, and Gaur, V. K., 1996. Straightforward inversion scheme (SIS) for one-dimensional magnetotelluric data. *Proceedings of the Indian Academy of Science (EPS)*, **105**(4), 413–429.

Gupta, P. K., Niwas, S., and Gaur, V. K., 1997. Straightforward inversion of vertical electrical sounding data. *Geophysics*, **62**(3), 775–785.

Habashy, T. M., and Abubakar, A., 2004. A General framework for constraint minimization for the inversion of electromagnetic measurements. *Progress In Electromagnetics Research, PIER*, **46**, 265–312.

Hadamard, J., 1902. Sur les problemes aux derives partielles et leur signification physique. *Princeton University Bulletin*, **13**, 49–52. Reprinted in his Oeuvres, Vol. III, Centre Nat. Recherche Sci., Paris, 1968, 1099–1105.

Hansen, P. C., 1992. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review*, **34**, 561–580.

Indira, N. K., and Gupta, P. K. (eds.), 1998. *Inverse Methods: General Principles and Applications to Earth System Sciences*. New Delhi: Narosa.

Jackson, D. D., 1972. Interpretation of inaccurate, insufficient, and inconsistent data. *Geophysical Journal of the Royal Astronomical Society*, **28**, 97–109.

Jackson, D. D., 1979. The use of a-priori data to resolve non-uniqueness in linear inversion. *Geophysical Journal of the Royal Astronomical Society*, **57**, 137–158.

Kunetz, G., 1972. Processing and interpretation of magnetotelluric soundings, *Geophysics*, **37**, 1005–1021.

Lawson, C. L., and Hanson, R. J., 1995. *Solving Least Squares Problems*. Philadelphia: SIAM.

Lines, L. R., and Treitel, S., 1984. Tutorial: a review of least-squares inversion and its application to geophysical problems. *Geophysical Prospecting*, **32**, 159–186.

Mallat, S., 2009. *A Wavelet Tour of Signal Processing: The Sparse Way*. Amsterdam: Elsevier.

Menke, W., 1984. *Geophysical Data Analysis: Discrete Inverse Theory*. Orlando: Academic.

Moore, E. H., 1920. On the reciprocal of the general algebraic matrices. *Bulletin of American Mathematical Society*, **26**, 394–395.

Oldenburg, D. W., 1976. Calculation of Fourier transform by the Backus-Gilbert method, *Geophysical Journal of the Royal Astronomical Society*, **44**(2), 413–431.

Oldenburg, D. W., and Li, Y., 2004. Inversion for Applied Geophysics – A Tutorial, presented at EMI workshop held at Hyderabad.

Penrose, R., 1955. A generalized inverse for matrices. *Proceedings of the Cambridge Philosophical Society*, **51**, 406–413.

Rao, A. R., and Bhimasankaram, P., 1992. *Linear Algebra*. New Delhi: Tata McGraw-Hill.

Rao, C. R., and Mitra, S. K., 1971. *Generalized Inverse of Matrices and Its Applications*. New York: Wiley.

Rawlinson, N., Pozgay, S., and Fishwick, S., 2010. Seismic tomography: a window into deep Earth – review. *Physics of the Earth and Planetary Interiors*, **178**, 101–135.

Reichel, L., and Sadok, H., 2008. A new L-curve for ill-posed problems. *Journal of Computational and Applied Mathematics*, **219**, 493–508.

Rezghi, M., and Hosseini, S. M., 2009. A new variant of L-curve for Tikhonov regularization. *Journal of Computational and Applied Mathematics*, **231**, 914–924.
Rojas, M., and Sorensen, D. C., 2002. A trust region approach to the regularization of large scale discrete forms of ill-posed problems. *SIAM Journal of Scientific Computing*, **23**(6), 1843–1861.
Saltelli, A., Ratto, M., Tarantola, S., and Campolongo, F., 2006. Sensitivity analysis practices: strategies for model based inferences – review. *Reliability Engineering and Safety*, **91**, 1109–1125.
Scales, J. A., Smith, M. L., and Treitel, S., 2001. *Introductory Geophysical Inverse Theory*. Golden, White River Junction: Smizdat.
Sneider, R., and Trampert, J., 1999. Inverse problems in geophysics. In Wirgin, A. (ed.), *Wavefield Inversion*. New York: Springer, pp. 119–190.
Tarantola, A., 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Philadelphia: SIAM.
Tikhanov, A. N., and Arsenin, V. Y., 1977. *Solutions Ill-Posed Problems*. New York: Wiley.
Treitel, S., and Lines, L., 2001. Past, present, and future of geophysical inversion – a new millennium analysis. *Geophysics*, **66**(1), 21–24.
Twomey, S., 1977. *Introduction to the Mathematics of Inversion in Remote Sensing and Indirect Measurements*. Amsterdam: Elsevier Scientific.
Zhdanov, M. S., 2002. *Geophysical Inverse Theory and Regularization Problems*. New York: Elsevier.

Cross-references

Inverse Theory, Artificial Neural Networks
Inverse Theory, Global Optimization
Inverse Theory, Singular Value Decomposition

INVERSE THEORY, MONTE CARLO METHOD

Malcolm Sambridge[1], Kerry Gallagher[2]
[1]Seismology and Mathematical Geophysics, Research School of Earth Sciences, The Australian National University, Canberra, Australia
[2]UMR 6118- Géosciences Rennes, Gêosciences, Université de Rennes 1, Rennes Cedex, France

Definition

Monte Carlo method. A computational technique making use of random numbers to solve problems that are either probabilistic or deterministic in nature. Named after the famous Casino in Monaco.
Monte Carlo inversion method. A method for sampling a parameter space of variables representing unknowns, governed by probabilistic rules.
Markov chain Monte Carlo (McMC). A probabilistic method for generating vectors or parameter variables whose values follow a prescribed density function.

Introduction

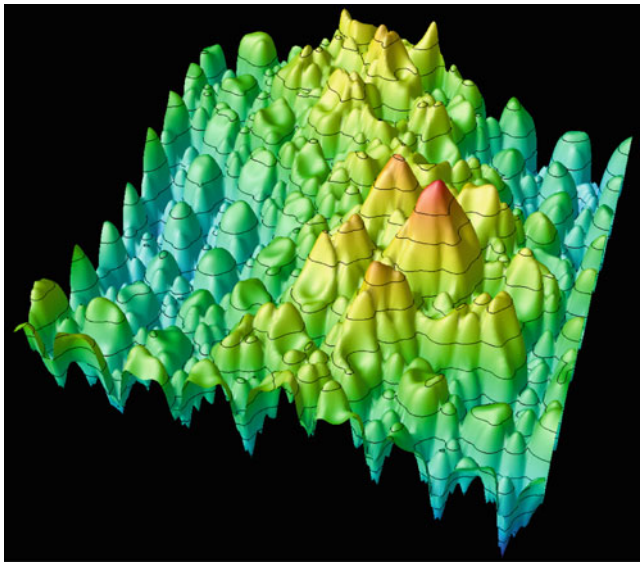
Because geophysical observations are made at (or very near) the Earth's surface, all knowledge of the Earth's interior is based on indirect inference. There always exists an inverse problem where models of physical properties are sought at depth that are only indirectly constrained by the available observations made at the surface. Geophysicists have been dealing with such problems for many years, and in doing so have made substantial contributions to the understanding of inverse problems.

Pioneering work on linear inverse problems in the 1960s arose out of the need to understand how to use new surface observables from seismology to constrain radial variations in geophysical properties at depth within the Earth. Data were few in number and attention was focused on the mathematical structure of the inverse problem and the ways in which reliable information could be recovered. This resulted in a series of important papers beginning with Backus and Gilbert (1967, 1968, 1970). Since that time the geosciences, like many other fields, have moved into a data-rich environment with increasing availability of computational power. Considerable progress has been made over 30 years utilizing the class of linear (typically least squares) parameter estimation algorithms, which are common to many areas of the physical sciences (Aster et al., 2005). In many of the inverse problems encountered the dependence of data on models is nonlinear and this must be taken into account for meaningful solutions. Often this is achieved by performing a local linearization and using *Inverse Theory, Linear*. As the mathematical relationship between data and unknowns becomes complex then linearized methods fail because they depend heavily on having a starting model for the iterative process which must be close enough to the solution for convergence.

Over the last 30 years there has been considerable progress in the solution of highly nonlinear inverse problems involving a limited number of parameters so that thorough exploration can be made of the character of models. Many algorithms have been devised, most of which make use of random numbers to make decisions, that is, in how to generate a set of values of the unknowns whose predictions can be compared to the available data. The original description of Monte Carlo methods by Hammersley and Handscomb (1964) is "the branch of experimental mathematics which is concerned with experiments on random numbers." By this definition all inversion techniques that make use of random numbers are Monte Carlo methods. Many of the *Inverse Theory, Global Optimization* inversion methods fall within the class. The particular approach known as Markov chain Monte Carlo is the primary focus of the present article.

Nonlinearity and multimodal fitness functions

Figure 1 shows a fitness surface from an inverse problem that arises in the analysis of infrasound array data. (Infrasound arrays are used by the United Nations Comprehensive Test Band Treaty Organization to monitor international adherence to the nuclear test ban treaty.) The height of the surface represents the degree of



Inverse Theory, Monte Carlo Method, Figure 1 A multimodal data fit surface arising from the mismatch between two oscillatory fields in the infrasound inversion problem (Kennett et al., 2003).

agreement between two oscillatory fields. There are just two unknowns in this case, which represent tuning parameters in the infrasound array. The object of the exercise is to best tune the array for sensitivity to incoming atmospheric signals, which means finding the point on the surface where the fitness is maximum. We see a curved rim of local maxima from background to foreground, with a broader valley of low fit to the left of the central maximum (red). The physics of the forward problem, that is, calculating the oscillatory field corresponding to a pair of tuning parameters, as well as the nature of the data itself result in a complicated 2-D fitness function.

The set of unknowns that gives the best fit (i.e., smallest misfit) to data corresponds to the global maximum of the multimodal function and to find it one must employ *Inverse Theory, Global Optimization* techniques such as model space search. In this example the global maximum (at the red central peak) was efficiently found with the neighborhood algorithm of Sambridge (1999), which utilizes ideas from the field of computational geometry. Optimization techniques based on local *Inverse Theory, Linear* would only be suitable once a trial solution is found within the vicinity of the global peak (shaded red). Adaptive Monte Carlo-based direct search approaches like genetic algorithms, simulated annealing, and the neighborhood algorithm (see *Inverse Theory, Global Optimization*) are able to solve this (two unknown parameters) problem relatively easily due to their ability to detect the variation of the fit and concentrate sampling where there is most benefit.

This example demonstrates the complexity of inverse problems in cases where the data are highly oscillatory waveforms, a common situation in fields such as acoustics and seismology, where the dimension of the problem is often much higher.

For the 2-D example in Figure 1, the objective is to tune a particular instrument for maximum sensitivity, and it is appropriate to seek a global maximum. More generally in inverse problems the fitness landscape would represent the difference between observations and predictions made by a mathematical model. In this case simply finding the best-fit solution is inadequate. One needs to characterize the uncertainty in the solution, for example, assess how noise in the data lead to errors and trade-offs in the estimated model. Linearized techniques (see *Inverse Theory, Linear*) could be used, but all uncertainty estimates are then based on the assumption of local linearity and do not truly reflect the global nature of the data constraint.

Another issue that often arises in inverse problems is that of nonuniqueness (see *Inverse Theory, Singular Value Decomposition*). In this case it is not possible to fully constrain the unknowns from the data. The model is unbounded and so best data fit solutions do not exist, and extra assumptions or independent information must be introduced to achieve a single optimal solution. In linearized inversion some form of regularization is used. An example is damping a solution back to some reference set of values (or model) (Aster et al., 2005). It is well known that in this case the details of the solution depend on the nature of regularization used. In addition, uncertainty estimates produced by linearized theory often reflect the choice of regularization. Typically, the least well-constrained components of the solution require the most regularization and resulting uncertainty estimates are severe underestimates of the real errors (Aster et al., 2005 for an example) potential leading to overconfidence in the results.

Bayesian inference

An alternative approach to inversion is Bayesian inference. Many textbooks and review papers are available. Discussions within a geophysical context can be found in Tarantola and Valette (1982), Duijndam (1988a, b), Sambridge and Mosegaard (2002), and Mosegaard and Sambridge (2002). In Bayesian inference, all information on the unknowns is represented in terms of probability density functions (PDF). Within this framework it is accepted that all inference is relative. What one learns from the data gets added to what is known prior to collecting the data and represented in terms of an a posteriori PDF. The most commonly used form of Bayes' rule is given below

$$p(m|d) = kp(d|m)p(m) \tag{1}$$

where $p(m|d)$ is the PDF of the model vector, m, containing the unknowns, given the data vector, d, containing the data; $p(d|m)$ is the likelihood function measuring the probability of the data, d, being observed given the model m; $p(m)$ is the a priori PDF on the model (which is known or assumed about m before the data are

collected), and k is a constant of proportionality. In a Bayesian framework, all information on the unknown variables in the model is represented by the posterior PDF, $p(m|d)$ and one usually sets about trying to generate an ensemble of candidate solutions to the inverse problem whose density is distributed according to this function. This is termed sampling the posterior PDF.

The posterior PDF is the product of the likelihood and the prior PDF. Only the former contains the data vector, d. The likelihood increases as the model fits the data better relative to the noise in the data. The form of the likelihood depends on the statistical character of the data errors. A simple example is a multidimensional Gaussian function characterized by a mean and a covariance matrix, both of which are usually known or assumed. The prior PDF represents information known about the model before collecting the data represented in a probabilistic manner, and may take a variety of forms. Again a multidimensional Gaussian is the most simple, but rarely is real information in this convenient form. Prior PDFs can be the most controversial component of Bayesian inference as there is always a degree of subjectivity in any choice, and the only way to represent no prior information is to not have a prior PDF. Comparisons of Bayesian and alternate approaches can be found in Malinverno and Parker (2005).

We see then that instead of seeking a single optimal solution, in a Bayesian framework many samples are sought. Assessment of the constraints placed on the model is achieved by examining collective properties. Typically, this is done by plotting the distribution of samples as a function of one or more subsets of unknowns, calculating credible intervals to represent uncertainty and covariance matrices to examine the trade-offs between parameters.

The main task to be carried out is then to generate random samples that follow the multidimensional posterior PDF $p(m|d)$arising from the inverse problem. McMC methods are practical tools for dealing with complicated probability distributions. Used correctly they result in (quasi)-independent samples whose density follows any target PDF. They have been the subject of much research in fields from Theoretical Physics to Computational statistics. For summaries, see Smith (1991), Smith and Roberts (1993), and Bernardo and Smith (1994). Below we describe the McMC method briefly and provide a simple illustrative example.

Markov chain Monte Carlo

Fixed dimension approach

McMC can be regarded as a combination of random Monte Carlo sampling and a Markov chain random walk strategy around the model space. The aim is to produce an ensemble of models from a probability distribution, that is, the posterior PDF, using only function evaluations. The basic approach was developed from the work of Metropolis et al. (1953), placed in a Bayesian framework by Hastings (1970), and a useful overview is given in Gilks et al. (1996). The practical applications lagged behind theoretical developments as a consequence of the need for many simulations. However, the increase in computing power over the last 15 years or so has led to a rapid increase in use of this methodology in geophysics and other fields of Earth Sciences (e.g., Mosegaard and Tarantola, 1995; Malinverno, 2002; Sambridge et al., 2006; Gallagher et al., 2009).

The algorithm is as follows: first we choose an initial model from the prior distribution, and calculate its likelihood. Then we generate a new model by making a random perturbation (Monte Carlo) to the current model. This new model is known as the proposed model and depends only on the values of the current model (Markov chain). The final stage is to decide whether we replace the current model with the proposed model, or stay at the current model and repeat the whole process. This important step is determined from the acceptance criterion, which is defined below

$$\alpha = \min\left\{1, \frac{p(m')p(d|m')q(m|m')}{p(m)p(d|m)q(m'|m)}\right\} \tag{2}$$

where m' and m are the proposed and current models, respectively, $q(a|b)$ is the probability of proposing model a, given a current model b, and the other distributions are as defined earlier. The decision to accept or reject a proposed model is made by comparing the value of α (which is always between 0 and 1) to a uniform (between 0 and 1) random number, u. If $u < \alpha$ then we replace the current model with the proposed model, if not we discard the proposed model and stay at the current model. We then continue the sampling process (perturb the new current model and so on) for many iterations.

The choice of the proposal function is not critical to the correctness of the sampler, but does affect the efficiency, performance, and convergence. A typical choice might be a normal distribution, centered on the current model, and then we need to tune the performance through the scale parameter of this distribution (e.g., the variance). If we choose too small a scale parameter, the proposed model will be very similar to the current model, their likelihoods will be similar, and we will almost always accept the proposed model. If we choose too large a scale parameter, the proposed model will tend to be very different to the current model, and lead to large changes in the likelihood, which are more likely to be rejected. In practice, both situations mean that we tend to move slowly around the model space. The proposal functions need then to be tuned for particular problems to achieve a reasonable balance between accepting and rejecting the proposed models. A reasonable rate of acceptance is around 30–40%. Generally, we can choose proposal functions that are symmetrical, so that $q(a|b) = q(b|a)$, so these terms cancel out in the acceptance criterion. Also, if we choose uniform prior distributions, then the prior terms also cancel. The acceptance criterion then reduces to the original Metropolis et al. (1953) algorithm.

After an initial period sampling (known as the burn-in), the current model from each iteration is taken as representing a sample from the posterior distribution (the sampling chain is then stationary). If the model space has N dimensions, and we are interested in the distribution on one of the parameters, m_i, for example, then formally we need to solve the following integral

$$p(m_i) = \int p(m_i, m_j) dm_j; \quad j = 1, .., i-1, i+1, ..., N \tag{3}$$

that is we need to integrate out the variation in all parameters except m_i. This is known as marginalizing and $p(m_i)$is the marginal probability distribution of m_i. Using the McMC samples, we can just plot all value of m_i as a histogram as the sampling effectively deals with the integration.

Also, it is straightforward to calculate estimates of the expected (or average) value for any parameter. Formally, we have the expected value for parameter m_i defined as

$$E(m_i) = \int p(m_i) m_i dm_i \tag{4}$$

Using the McMC samples, we simply average over all the samples accepted for that parameter, that is,

$$E(m_i) = \frac{1}{N_a} \sum_{j=1}^{N_a} m_i^j \tag{5}$$

where N_a is the number of samples accepted (post-burn-in) for model parameter m_i

Transdimensional approach

A major issue concerning most inverse problems, and the approaches used to solve them, is how best to balance the twin desires of fitting the observations and avoiding introduction of unjustified complexity in the resulting models. Green (1995) introduced a transdimensional form of McMC (referred to as Reversible Jump), in which the inversion procedure involves the inference of the model complexity (see also birth–death McMC Geyer and Møller, 1994). For finite dimension models (with a fixed number of unknowns) this then typically becomes a question of determining the dimension of the model. If we are dealing with two models with dimensions k and k', then acceptance criterion can be written as

$$\alpha = \min\left\{1, \frac{p(\mathrm{k}')p(m'|\mathrm{k}')p(d|m', \mathrm{k}')q(m|m')}{p(\mathrm{k})p(m|\mathrm{k})p(d|m, \mathrm{k})q(m'|m)}\right\} \tag{6}$$

Here we separate the prior on the number of dimension, $p(k)$, from the model parameter prior, $p(m|k)$. The proposal function $q()$ becomes more complex as we now want to propose models with different dimensions. Moreover, we need to allow for the transformation from one model to another to ensure that theoretical probability requirements are maintained. In dealing with a situation where we are simply increasing or decreasing the number of parameters, we can write

$$\alpha = \min\left\{1, \frac{p(\mathrm{k}')p(m'|\mathrm{k}')p(d|m', \mathrm{k}')g(u^{\mathrm{k}'})}{p(\mathrm{k})p(m|\mathrm{k})p(d|m, \mathrm{k})g(u^{\mathrm{k}})} |\mathrm{J}|\right\} \tag{7}$$

Here $u^{k'}$ and u^k are vectors of random numbers of length r' and r, respectively, and used to transform from one model to another, such that $r+k = r'+k'$, and $g(.)$ is the probability distribution used to generate these random numbers. The term $|J|$ is the Jacobian, and allows for the transformation between the two models, that is,

$$|\mathrm{J}| = \frac{\partial(m', u^{\mathrm{k}'})}{\partial(m, u^{\mathrm{k}})} \tag{8}$$

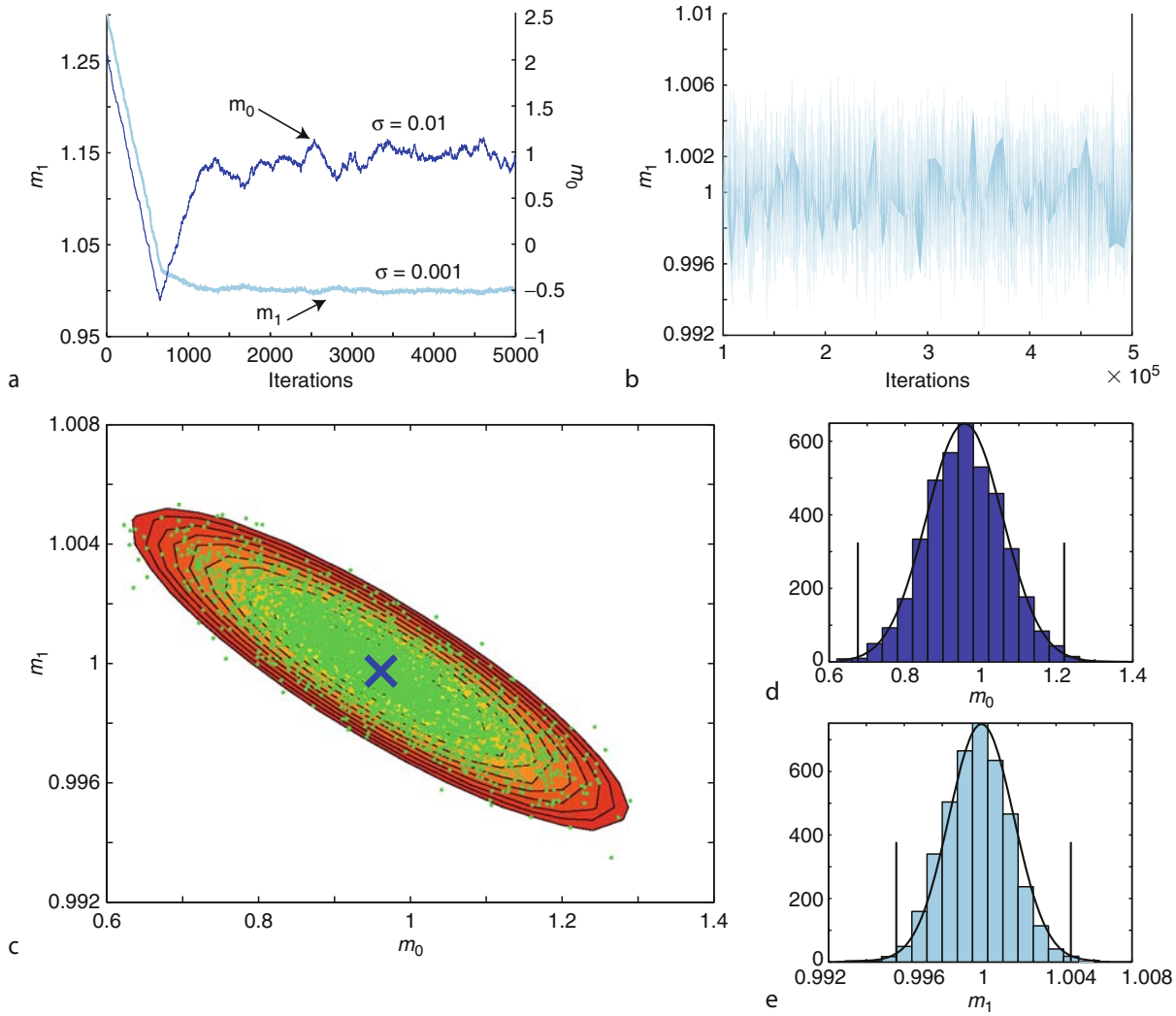
The last equation for α is actually a general form for the acceptance criterion, although for fixed-dimensional problems the Jacobian is generally 1, and the proposal functions are of the form as described earlier. The details of the reversible jump acceptance criterion are discussed in more detail by Green (2003), Malinverno (2002), and Sambridge et al. (2006) and examples of the implementation algorithms for variable dimension problems are given in Jasra et al. (2006), Bodin and Sambridge (2009), Charvin et al. (2009), and Hopcroft et al. (2009). One important characteristic of the Bayesian transdimensional formulation is that it is naturally parsimonious. For two models that fit the data equally well, it will tend to favor simpler models over complex ones as a consequence of the posterior probability effectively being penalized through the addition of more terms to the prior distribution.

A simple example

To demonstrate McMC in action, we choose a simple two parameter linear regression problem, that is,

$$y_i = m_0 + m_1 x_i + \varepsilon_i, i = 1, ..., N \tag{9}$$

where m_0 and m_1 are the model parameters, y is an observed/measured value, and ε is the data error. In a Bayesian formulation, this problem has an analytical solution for a uniform prior on the model parameters, assuming the data error is known (Lee, 1989, p. 180). We chose $m_0 = m_1 = 1$ and generated 100 synthetic data (y) for random values of x between 0 and 100, and added noise ($\varepsilon = 0.5$). We used Gaussian proposal distributions, with different scale parameters ($\sigma_{m_0} = 0.01$, and $\sigma_{m_1} = 0.001$). In Figure 2 we show the sampling for the two parameters, starting from a randomly selected model. Figure 2a shows that the sampler has not reached the stationary state until at least 1,500–2,000 iterations. Figure 2b shows the sampling for parameter m_1 for later iterations which is clearly stationary. The sampling resembles a white noise spectrum about the mean (or expected) value, lacking any internal structure as a function of



Inverse Theory, Monte Carlo Method, Figure 2 (**a**) Intitial 5,000 iterations for sampling of two parameters for the linear regression problem. (**b**) Post-burn-in iterations for parameter m_1. (**c**) The *green points* show the post-burn-in sampling, and the contours are the log likelihood function. The *blue cross* is the best solution (equivalent to the analytical maximum likelihood or least squares solution). (**d**) Marginal distribution for parameter m_0. The histogram is constructed from the post-burn-in accepted samples, and the curve is the analytical solution for the marginal distribution. The two vertical bars mark the upper and lower bounds on the 95% credible interval. (**e**) As (**d**) but for parameter m_1.

iteration. It is also clear that the sampler manages to move toward the upper and lower extreme values of the parameter range (determined by the posterior PDF and the proposal function scale parameter). These are diagnostic (but qualitative) characteristics of stationarity.

Figure 2c shows the 2-D distribution of samples of the post-burn-in accepted samples (here we thinned the chain taking every 100th sample), compared to the log likelihood function (which is proportional to the log of the posterior distribution as we use uniform priors). The density of the sampling increases around the high likelihood values, but there are still some samples from the lower likelihood regions. Figure 2d and e shows the marginal distributions for the two parameters as a frequency histogram, and also the analytical solutions (scaled to the same peak height). These are constructed simply by taking all the accepted values for a given model parameter, as the sampler effectively integrates out the other parameters. Also shown are the 95% credible interval ranges for each parameter. These are constructed by sorting all the samples for a given parameter in ascending order, and finding the indices for the credible values such that 2.5% of the samples are less than the lower credible value, and 2.5% of the values are greater than the upper credible value. Again, we can see that the sampler has managed to distribute itself across the distribution according to the posterior probability and these histograms are good representations of the marginal distributions.

Summary

Monte Carlo sampling, relying on random numbers, has been used in Geophysics for over 40 years, although the increase in computing power has seen a commensurate increase in applications in the last 15 years or so. This approach avoids the use of gradients, is robust to local minima, and so is suitable for nonlinear inverse problems which often have complex misfit (or fitness) surfaces in high dimensions. McMC, particularly when used in a Bayesian formulation, provides a means of sampling a model space according to the (unknown) posterior distribution for the model parameters. Transdimensional (or reversible jump) Markov chain Monte Carlo generalizes this approach to allow models of different dimensions to be considered, and provides a means of choosing between models of differing complexity. Quantifying the posterior distribution with McMC is then a solution to the inverse problem and various types of inference can be made from this distribution (e.g., expected values, marginal distributions, credible intervals) to characterize the model space.

Acknowledgments

We would like to thank FAST (French-Australia Science and Technology exchange program) for their support during the preparation of this article. This project is supported by the Commonwealth of Australia under the *International Science Linkages program*.

Bibliography

Aster, R., Borchers, R., and Thurber, C. H., 2005. *Parameter Estimation and Inverse Problems*. International Geophysics Series, 90. Amsterdam: Elsevier.

Backus, G. E., and Gilbert, J. F., 1967. Numerical applications of a formalism for geophysical inverse problems. *Geophysical Journal of the Royal Astronomical Society*, **13**, 247–276.

Backus, G. E., and Gilbert, J. F., 1968. The resolving power of gross Earth data. *Geophysical Journal of the Royal Astronomical Society*, **16**, 169–205.

Backus, G. E., and Gilbert, J. F., 1970. Uniqueness in the inversion of inaccurate gross Earth data. *Philiosophical Transactions of the Royal Society of London A*, **266**, 123–192.

Bernardo, J. M., and Smith, A. F. M., 1994. *Bayesian Theory.* Chichester: Wiley.

Bodin, T., and Sambridge, M., 2009. Seismic tomography with the reversible jump algorithm. *Geophysical Journal International*, **178**, 1411–1436.

Charvin, K., Gallagher, K., Hampson, G., and Labourdette, R., 2009. A Bayesian approach to infer environmental parameters from stratigraphic data 1: methodology. *Basin Research*, **21**, 5–25.

Duijndam, A. J. W., 1988a. Bayesian estimation in seismic inversion part I: principles. *Geophysical Prospecting*, **36**, 878–898.

Duijndam, A. J. W., 1988b. Bayesian estimation in seismic inversion part II: uncertainty analysis. *Geophysical Prospecting*, **36**, 899–918.

Gallagher, K., Charvin, Nielsen, S., Sambridge, M., and Stephenson, J., 2009. Markov chain Monte Carlo (McMC) sampling methods to determine optimal models, model resolution and model choice for Earth science problems. *Marine and Petroleum Geology*, **26**, 525–535.

Geyer, C. J., and Møller, J., 1994. Simulation procedures and likelihood inference for spatial point processes. *Scandinavian Journal of Statistics*, **21**, 369–373.

Gilks, W. R., Richardson, S., and Spiegalhalter, D. J., 1996. *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

Green, P. J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Green, P. J., 2003. Trans-dimensional McMC. In Green, P. J., Hjort, N., and Richardson, S. (eds.), *Highly Structured Stochastic Systems*. Oxford Statistical Sciences Series, Oxford: Oxford University Press, Chap. 6, pp. 179–196.

Hammersley, J. M., and Handscomb, D. C., 1964. *Monte Carlo Methods*. London: Chapman & Hall.

Hastings, W. K., 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Hopcroft, P., Gallagher, K., and Pain, C. C., 2009. A Bayesian partition modelling approach to resolve spatial variability in climate records from borehole temperature inversion. *Geophysical Journal International*, **178**, 651–666.

Jasra, A., Stephens, D. A., Gallagher, K., and Holmes, C. C., 2006. Analysis of geochronological data with measurement error using Bayesian mixtures. *Mathematical Geology*, **38**, 269–300.

Kennett, B. L. N., Brown, D. J., Sambridge, M., and Tarlowski, C., 2003. Signal parameter estimation for sparse arrays. *Bulletin. Seismological Society of America*, **93**, 1765–1772.

Lee, P. M., 1989. *Bayesian Statistics: an Introduction*. New York/Toronto: Edward Arnold.

Malinverno, A., 2002. Parsimonious Bayesian Markov chain Monte Carlo inversion in a non linear geophysical problem. *Geophysical Journal International*, **151**, 675–688.

Malinverno, A., and Parker, R. L., 2005. Two ways to quantify uncertainty in geophysical inverse problems. *Geophysics*, **71**, 15–27.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E., 1953. Equations of state calculations by fast computing machines. *The Journal of Chemical Physics*, **21**, 1087–1092.

Mosegaard, K., and Sambridge, M., 2002. Monte Carlo analysis of inverse problems. *Inverse Problems*, **18**, R29–R54.

Mosegaard, K., and Tarantola, A., 1995. Monte Carlo sampling of solutions to inverse problems. *Journal of Geophysical Research*, **100**, 12431–12447.

Sambridge, M., 1999. Geophysical inversion with a neighbourhood algorithm -I. Searching a parameter space. *Geophysical Journal International*, **138**, 479–494.

Sambridge, M., and Mosegaard, K., 2002. Monte Carlo methods in Geophysical inverse problems. *Reviews of Geophysics*, **40**, 3.1–3.29.

Sambridge, M., Gallagher, K., Jackson, A., and Rickwood, P., 2006. Trans-dimensional inverse problems, model comparison and the evidence. *Geophysical Journal International*, **167**, 528–542.

Smith, A. F. M., 1991. Bayesian computational methods. *Philiosophical Transactions of the Royal Society of London A*, **337**, 369–386.

Smith, A. F. M., and Roberts, G. O., 1993. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society B*, **55**, 3–23.

Tarantola, A., and Valette, B., 1982. Inverse problems = quest for information. *Journal of Geophysics*, **50**, 159–170.

Cross-references

Inverse Theory, Artificial Neural Networks
Inverse Theory, Global Optimization
Inverse Theory, Linear
Inverse Theory, Singular Value Decomposition

INVERSE THEORY, SINGULAR VALUE DECOMPOSITION

Ajay Manglik
National Geophysical Research Institute, Council of Scientific and Industrial Research (CSIR), Hyderabad, India

Definition

Condition number. Ratio of the largest singular value to the smallest singular value of a matrix. For a singular matrix it is infinite and it is very large for an ill-conditioned matrix.

Description

Linear geophysical inverse theory (*Inverse Theory, Linear*) deals with solving a system of linear equations, represented in matrix form as

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{1}$$

where $\mathbf{A}$ is a $m \times n$ Jacobean matrix, $\mathbf{x}$ a $n \times 1$column vector of perturbations in unknown model parameters with respect to a reference model, and $\mathbf{b}$ the corresponding $m \times 1$ column vector of perturbations in observed responses. For a full rank square matrix ($m = n = p$: the rank of matrix $\mathbf{A}$), it is easy to find a unique inverse $\mathbf{A}^{-1}$ and estimate the model parameters as $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$. However, geophysical inverse problems mostly deal with inversion of rectangular matrix $\mathbf{A}$ wherein m is either greater than n (over-determined system representing more number of observations than the unknown parameters, also known as least squares solution) or less than n (under-determined system representing less number of observations compared to the unknown parameters, also called minimum norm solution). For many geophysical problems the rank p of the matrix is less than m and n. In such situations, linear geophysical inverse theory aims at finding out a pseudo-inverse $\mathbf{A}^{-}$of $\mathbf{A}$ such that

$$\hat{\mathbf{x}} = \mathbf{A}^{-}\mathbf{b}, \tag{2}$$

where $\hat{\mathbf{x}}$ represents the column vector of the estimated model parameters. Moore (1920) and Penrose (1955), using a spectral decomposition form of matrix $\mathbf{A}$, showed that there exists a unique pseudo-inverse $\mathbf{A}^{-}$ of the matrix $\mathbf{A}$ such that (1) $\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}$, (2) $\mathbf{A}^{-}\mathbf{A}\mathbf{A}^{-} = \mathbf{A}^{-}$, (3) $(\mathbf{A}\mathbf{A}^{-})^{T} = \mathbf{A}\mathbf{A}^{-}$, and (4) $(\mathbf{A}^{-}\mathbf{A})^{T} = \mathbf{A}^{-}\mathbf{A}$, where superscript Tdenotes the transpose of matrix. Further details can be found in Inverse theory, linear.

Singular Value Decomposition (SVD) is a powerful tool to invert matrices that are either singular or numerically very close to singular (ill-conditioned system). Standard matrix equation solvers can fail in such situations. The advantage of SVD is that one need not a priori know the nature of the problem, that is, whether it is over determined, under determined, or partially determined. SVD is very general in the sense that it can be applied to any arbitrary size matrix. The technique was discovered independently by Beltrami in 1873 and by Jordan in 1874 as a tool to solve square matrices and was extended to rectangular matrices by Eckart and Young in the 1930s (Stewart, 1993). Its use as a computational tool dates back to the 1960s (Golub and Kahan, 1965; Golub and Reinsch, 1970). Golub and van Loan (1996) demonstrated its usefulness and feasibility in a wide variety of applications.

Following SVD, a matrix $\mathbf{A}$ of size $m \times n$ is factorized in the form

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{T}, \tag{3}$$

where $\mathbf{U}$ is a $m \times m$ unitary matrix, $\mathbf{\Lambda}$ is a $m \times n$ diagonal matrix, and $\mathbf{V}$ is a $n \times n$ unitary matrix. The columns of $\mathbf{V}$ form a set of orthonormal vectors called parameter eigenvectors and the columns of $\mathbf{U}$ form a set of orthonormal basis vectors called data eigenvectors. For a special case of $m = n$, the matrix $\mathbf{\Lambda}$ has the form

$$\mathbf{\Lambda} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n \end{bmatrix}, \tag{4}$$

where $\sigma_1, \cdots, \sigma_n$ are called the singular values of matrix $\mathbf{A}$. The number of nonzero singular values is equal to the rank of $\mathbf{A}$. In the SVD representation, the singular values are arranged in the order of decreasing amplitude, that is,

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n. \tag{5}$$

For a real matrix, the columns of matrices $\mathbf{U}$ and $\mathbf{V}$ are orthogonal, that is, $\mathbf{U}^{T}\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^{T}\mathbf{V} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix and $T$is transpose of matrix. For complex matrices transpose is replaced by complex conjugate.

Inverse of A

The expression for the inverse of $\mathbf{A}$ can be obtained by first substituting Equation 3 into Equation 1 to get

$$\mathbf{U}\mathbf{\Lambda}\mathbf{V}^{T}\mathbf{x} = \mathbf{b}. \tag{6}$$

and then applying the following orthogonal transformation

$$\mathbf{b} = \mathbf{U}\mathbf{b}'; \qquad \mathbf{x} = \mathbf{V}\mathbf{x}', \tag{7}$$

and multiplying Equation 6 by $\mathbf{U}^{T}$, which gives

$$\mathbf{\Lambda}\mathbf{x}' = \mathbf{b}'. \tag{8}$$

Since $\mathbf{\Lambda}$ is a diagonal matrix, we can easily get

$$\mathbf{x}' = \mathbf{\Lambda}^{-1}\mathbf{b}'. \tag{9}$$

From Equation 7 and Equation 9, and orthogonality property of data and parameter eigenvectors, we get

$$\mathbf{x} = \mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{U}^T\mathbf{b}, \tag{10}$$

or the inverse of $\mathbf{A}$ can be written as

$$\mathbf{A}^{-1} = \mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{U}^T. \tag{11}$$

Rank-deficient matrix and its inverse

The advantages of above spectral decomposition can be seen in the case of $\mathbf{A}$ being a singular matrix. Let us assume that only p of n singular values of $\mathbf{A}$ are nonzero. In this case, the data and parameter eigenvectors can be split into two sub-spaces representing nonzero eigenspace and null-space, respectively, that is,

$$\begin{aligned} &\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_p & 0 \\ 0 & 0 \end{bmatrix}; \qquad \mathbf{U} = \left[\mathbf{U}_p, \mathbf{U}_0\right]; \\ &V = \left[V_p, V_0\right], \end{aligned} \tag{12}$$

then

$$\mathbf{A} = \left[\mathbf{U}_p, \mathbf{U}_0\right] \begin{bmatrix} \boldsymbol{\Lambda}_p & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{V}_p \\ \mathbf{V}_0 \end{bmatrix}^T, \tag{13}$$

which gives

$$\mathbf{A} = \mathbf{U}_p\boldsymbol{\Lambda}_p\mathbf{V}_p^T. \tag{14}$$

This indicates that matrix $\mathbf{A}$ can be reconstructed only from nonzero singular values and corresponding vector spaces. An inverse of a rank-deficient matrix can be written as:

$$\mathbf{A}^- = \mathbf{V}_p\boldsymbol{\Lambda}_p^{-1}\mathbf{U}_p^T, \tag{15}$$

by substituting Equation 12 into Equation 11.

Ill-condition matrix

Equation 15 assumes that the singular values of the system are either zero or positive scalars. If some of the singular values are very small then it is easy to see that the inverse of such values can be very large and, in the case of error in data, can lead to unstable solutions. Stability of the matrix inverse can be defined in terms of ratio of largest to smallest singular value, also called the condition number

$$\eta = \frac{\sigma_1}{\sigma_{\min}}. \tag{16}$$

The matrix is ill-conditioned if the condition number is very large. In such situations, the system can be stabilized by specifying a cutoff threshold η_0 for the condition number and ignoring all singular values that lead to the condition number larger than this threshold. Let there be p nonzero singular values arranged in decreasing order of their amplitudes of which only the first q singular values have the condition number smaller than η_0. Then $(q+1, \cdots, p)$ singular values can be reset to zero thus reducing the rank of the matrix to q. This can be understood as ignoring finer details of the parameter space, which are associated with small singular values. Geophysical inverse problems normally fall under this category.

For least square inverse problems, where an inverse of $\mathbf{A}$ is given by the expression

$$\mathbf{A}^- = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T, \tag{17}$$

Marquardt (1963) suggested that an ill-conditioned matrix $\mathbf{A}^T\mathbf{A}$ may be stabilized by adding an arbitrary damping factor λ to its diagonal elements, that is,

$$\mathbf{A}^- = \left(\mathbf{A}^T\mathbf{A} + \lambda^2\mathbf{I}\right)^{-1}\mathbf{A}^T. \tag{18}$$

In terms of singular value decomposition, it amounts to boosting the small singular values of the matrix $\mathbf{A}^T\mathbf{A}$. This can be shown by expressing Equation 18 in terms of singular values decomposition of the matrix $\mathbf{A}$, which gives

$$\mathbf{A}^- = \mathbf{V}\mathrm{diag}\left(\frac{\sigma}{\sigma^2 + \lambda^2}\right)\mathbf{U}^T, \tag{19}$$

where diag(...) represents the elements of a diagonal matrix. Here, it can be seen that adding λ to small singular values stabilizes the inverse.

Sensitivity to errors in data

Field observations normally contain noise superimposed on the actual response. If we assume a noise vector $\boldsymbol{\varepsilon}$ then we observe $\mathbf{b} + \boldsymbol{\varepsilon}$ instead of $\mathbf{b}$. This leads to an error $\delta\mathbf{x}$ in the estimated model parameters. Therefore, for noisy data Equation 1 may be written as:

$$\mathbf{A}(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \boldsymbol{\varepsilon}. \tag{20}$$

Substituting from Equation 1, we get

$$\mathbf{A}\delta\mathbf{x} = \boldsymbol{\varepsilon}. \tag{21}$$

Expressing matrix $\mathbf{A}$ in spectral form and applying orthogonal transformation $\delta\mathbf{x}' = \mathbf{V}\delta\mathbf{x}$ and $\boldsymbol{\varepsilon}' = \mathbf{U}\boldsymbol{\varepsilon}$, Equation 21 gives

$$\delta\mathbf{x}' = \boldsymbol{\Lambda}^{-1}\boldsymbol{\varepsilon}. \tag{22}$$

This shows that in the case of small singular values small error in data yields large error in estimated model parameters. This problem of instability of estimated model parameters can be overcome by either truncating the small singular values or boosting these by adding a Marquardt damping factor, as discussed in the previous section. Such treatment nevertheless results in loss of finer details of the model parameters.

Resolution matrices

In linear inverse theory, the concepts of parameter resolution and data information density analyses are used to analyze the resolution of estimated model parameters and the

regions of data contributing maximum to these model estimates. These resolution matrices can be expressed in terms of data and parameter eigenvectors. Parameter resolution matrix for a rank-deficient system can be obtained by substituting Equations 1, 14, and 15 into Equation 2, which gives

$$\Re = \mathbf{V}_p \mathbf{V}_p^T. \tag{23}$$

For a well-resolved system $\Re$ is an identity matrix $\mathbf{I}$. For a rank-deficient system this matrix provides information about the equivalence of various model parameters.

Similarly, data information density matrix can be written as

$$\Im = \mathbf{U}_p \mathbf{U}_p^T, \tag{24}$$

by substituting Equations 2, 14, and 15 into Equation 1.

Jupp and Vozoff (1975) provided an elegant description of the concepts of stability of solution and resolution of model parameters for geophysical data contaminated by noise and Vozoff and Jupp (1975) showed its application to Direct Current (DC) electrical resistivity and Magnetotelluric (MT) data interpretation.

Summary

SVD is a powerful tool to solve geophysical linear inverse problems, which often deal with rank-deficient singular matrices. The advantage of SVD is that one need not a-priori know the nature of the problem, that is, whether it is over determined, under determined, or partially determined. It is very general in the sense that it can be applied to any arbitrary size matrix.

Bibliography

Golub, G. H., and Kahan, W., 1965. Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics: Series B*, **2**, 205–224.

Golub, G. H., and Reinsch, C., 1970. Singular value decomposition and least squares solutions. *Numerische Mathematik*, **14**, 403–420.

Golub, G. H., and Van Loan, C. F., 1996. *Matrix Computations*, 3rd edn. Baltimore: Johns Hopkins.

Jupp, D. L. B., and Vozoff, K., 1975. Stable iterative methods for the inversion of geophysical data. *Geophysical Journal of the Royal Astronomical Society*, **42**, 957–976.

Lawson, C. L., and Hanson, R. J., 1974. *Solving Least Squares Problems*. Englewood Cliffs: Prentice-Hall.

Lines, L. R., and Treitel, S., 1984. A review of least-squares inversion and its application to geophysical problems. *Geophysical Prospecting*, **32**, 159–186.

Marquardt, D. W., 1963. An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, **11**, 431–441.

Moore, E. H., 1920. On the reciprocal of the general algebraic matrices. *Bulletin of the American Mathematical Society*, **26**, 394–395.

Penrose, R., 1955. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, **51**, 406–413.

Press, W. H., Teukolsky, S. A., Vellerling, W. T., and Flannery, B. P., 1993. *Numerical Recipes in C: The Art of Scientific Computing*. Indian Edition (New Delhi): Cambridge University Press.

Stewart, G. W., 1993. On the early history of the singular value decomposition. *SIAM Review*, **35**, 551–566.

Tarantola, A., 1987. *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation*. The Netherlands: Elsevier Science.

Vozoff, K., and Jupp, D. L. B., 1975. Joint inversion of geophysical data. *Geophysical Journal of the Royal Astronomical Society*, **42**, 977–991.

Cross-references

Inverse Theory, Linear

ISOSTASY

Anthony B. Watts
Department of Earth Sciences, University of Oxford, Oxford, UK

Synonyms

Isostacy – (Dutton, 1882)

Definition

Isostasy. A principle or general law (Heiskanen, 1931). Isostasy considers there is a certain surface within the Earth, known as the depth of compensation, on which the vertical stresses due to an overlying column of rock are equal (isos = Greek ισος "equal," stasis = Greek στάσις "a standing still"). Isostasy implies a state of hydrostatic equilibrium such that the Earth's crust and mantle float on their substrate and light regions have a greater elevation than dense regions.

Isostatic equilibrium. An idealized state that the outermost layers of the Earth tend toward following their disturbance by the addition or removal of loads associated, for example, with the waxing and waning of ice sheets, the growth and decay of volcanoes and the deposition, sliding and slumping of sediments.

Local isostasy. A hypothesis that the outermost layers of the Earth are weak such that the deformation caused by loading and unloading is localized, intense, and may involve faulting.

Regional (or flexural) isostasy. A hypothesis that the outermost layers of the Earth have an intrinsic strength such that they resist the deformation associated with loading and unloading and bend (or flex) over a broad area rather than break.

Thermal isostasy. A hypothesis that the outermost layers of the Earth form a cooling thermal boundary layer and that regional differences in topography are controlled by differences in the temperature structure such that hot regions have a greater elevation than cold regions.

Historical background

Isostasy is a principle that has played a major role in our understanding of the structure, morphology, and tectonic evolution of the Earth's outermost layers. The term was first coined by Dutton (1889), but there is evidence (Delaney, 1940; Vai, 2006) that the equilibrium state of the Earth's crust was being considered much earlier by, for example, the Italian scientists Leonardo da Vinci (1452–1519), who pondered the occurrence of marine fossils at high elevations in mountain ranges, and Luigi Fernando Marseli (1658–1730), who was puzzled about the juxtaposition of ocean deeps and mountains. J. F. W. Herschel (1792–1871) was one of the first to link subsidence in one area to uplift in another by some sort of subcrustal mantle flow and both C. Lyell (1797–1875) and C. Babbage (1790–1871) appreciated that the same area of the crust could experience both subsidence and uplift.

The breakthrough that led to the formulation of the principle of isostasy came following George Everest's pioneering geodetic work in India. Airy (1855) and then Pratt (1855) used Everest's deflection of the vertical data in northern India to address the question of how the Himalayan mountains were supported at depth. Fisher (1881), who favored Airy's hypothesis that the Himalaya were supported by a deep crustal "root," concluded that the Earth's crust was in a state of hydrostatic equilibrium and floated on its fluid substrate, much like an iceberg in seawater.

Today, we tend to think of isostasy as a restraining force rather than a plate tectonic driving force like ridge push or trench pull. Isostasy is a process that acts on the topography created at the plate boundaries by continental rifting, breakup, and mountain building so as to reshape it and redistribute the stresses that have built up. The existence of regions of large topographic relief, unusually thick and thin crust, and large-amplitude gravity anomalies are all pointers that isostatic equilibrium does not prevail everywhere. The main challenge in isostatic studies has been to quantify the degree to which the hydrostatic equilibrium of Earth's crust is achieved and the time and spatial scales over which it operates.

A number of isostatic models have been proposed, three of which are illustrated in Figure 1. Airy and Pratt are local models in which changes in the Earth's topography are supported by *either* changes in the thickness of a uniform density crust or by lateral changes in density of the crust and mantle. Both models imply that the crust and mantle responds to loads (e.g., volcanoes) or unloads (e.g., erosion) locally such that neighboring regions are not involved in their support or removal. They imply an intrinsically weak crust that responds to loads and unloads locally, intensely, and in some cases by faulting. Vening Meinesz, on the other hand, is a regional model in which loads and unloads are supported by a gentle bending or flexure over a wide area rather than by faulting.

It is important to remind ourselves that the models in Figure 1 are highly idealized and only represent the state that the Earth's outermost layers would tend toward in the absence of disturbing forces. These disturbing forces include many of the great cycles of geology such as the waxing and waning of ice sheets, the growth and decay of volcanoes, and sedimentation and erosion. Each of these processes is potentially a disturbing force that could act to delay or even prevent isostatic equilibrium from being achieved.

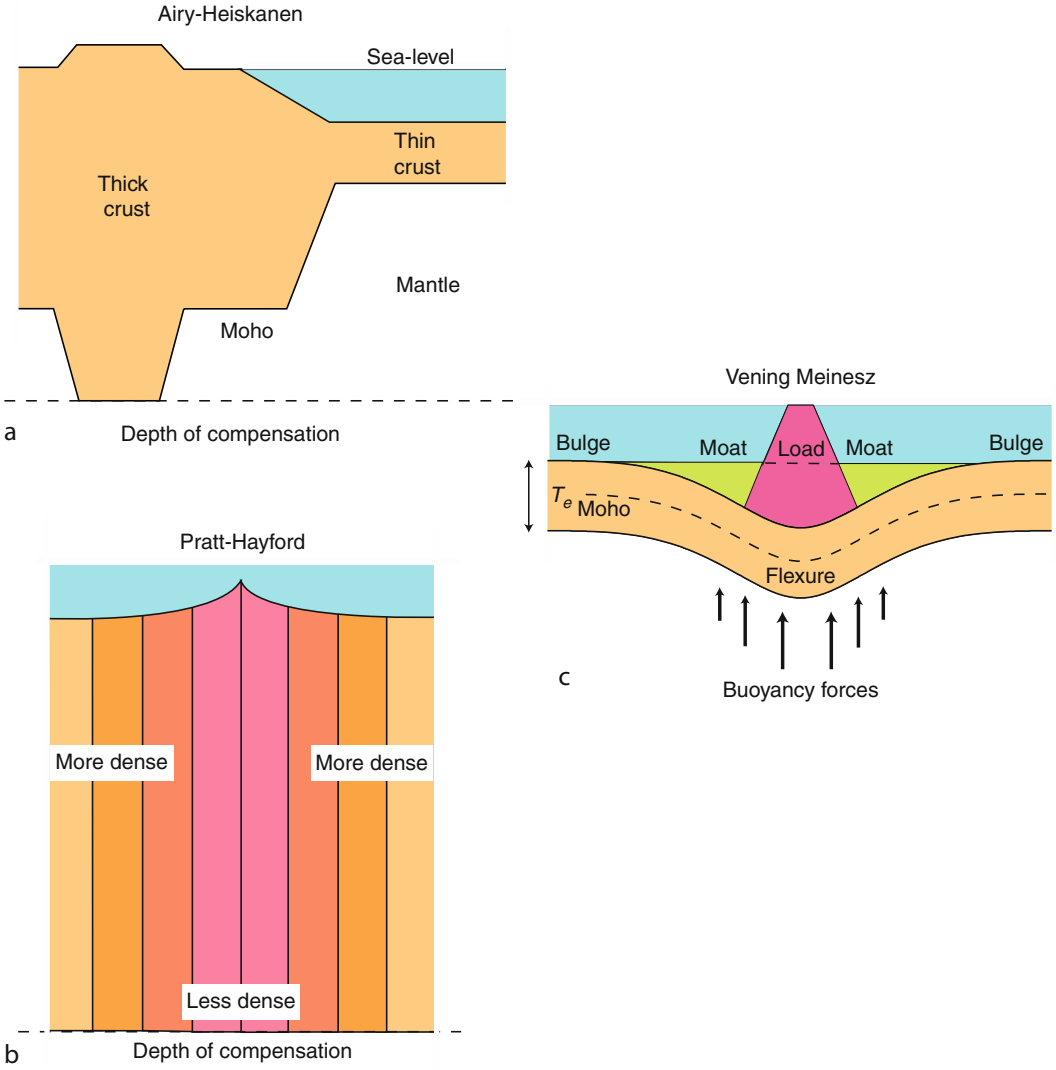
There is now a considerable body of work that indicates the Earth is a dynamic planet that is continuously being deformed. Isostasy, on the other hand, is a static concept that refers to a crust and mantle that has already moved and is now in equilibrium. It therefore appears at odds with current geodynamic views. But isostatic studies of the deformation that follows large earthquakes, late-glacial rebound, and volcano emplacement provide "snapshots" of the way that the crust and mantle responds to loads and unloads over certain temporal and spatial scales (Watts, 2007). Together, these "snapshots" provide information on lithosphere behavior and the clues as to how the crust and mantle would actually deform in response to continually applied loads and unloads.

A goal of isostatic studies is to determine the vertical motion history of the Earth's crust and mantle on short, through intermediate to long timescales. Therefore, isostasy is central to studies of environmental change, especially those that are aimed at determining the relative role of tectonics and sea-level change in controlling the position of the shoreline today and in the geological past. Most significantly, isostatic studies provide information on how the lithosphere responds to the load and unload shifts that occur during climate-driven glacial and interglacial cycles. We are presently in an interglacial, for example, yet the isostatic effects of the Last Glacial Maxima (LGM) are still clearly visible today in tide gauge and satellite Global Positioning System (GPS) data. Isostasy is a key to understanding the lithosphere on which we live and how it interacts with the asthenosphere below and the cryosphere, hydrosphere, and atmosphere above.

Concepts and applications

The isostatic models illustrated in Figure 1 imply different gravity anomalies and crustal and mantle structures. Airy and Pratt, for example, predict that free-air gravity anomalies will generally be small over elevated regions (e.g., volcanoes) because they are underlain by either a thick, uniform density, crustal "root" or a constant thickness low-density crust, the negative gravitational effect of which will compete with the positive effect of the volcano. Vening Meinesz, on the other hand, predicts a shallower depth to Moho than Airy or Pratt and, hence, the free-air gravity anomaly will be generally large over a volcano because its gravity effect will now dominate over that of the compensation.

Figure 2 shows the topography, free-air and isostatic gravity anomalies in the region of northern India where the early geodesists carried out their classical work. The isostatic anomaly has been calculated by subtracting

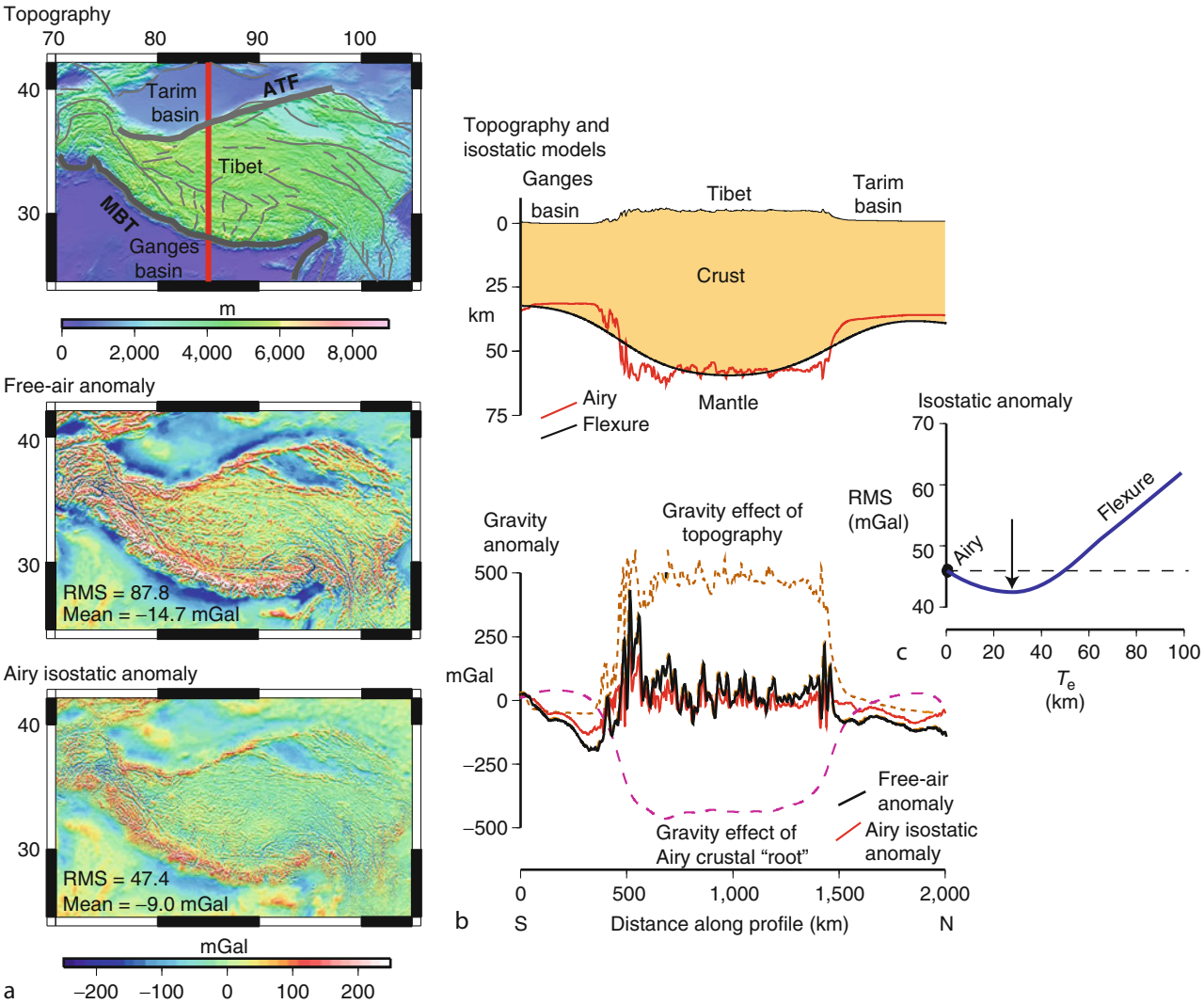


Isostasy, Figure 1 The main types of isostatic models. Each model implies a state of hydrostatic equilibrium such that the Earth's outermost layers are in a state of flotation on their more fluid substrate. The names (e.g., Airy-Heiskanen) refer to the workers who first conceived and quantified the models. Vertical (lithostatic) stresses are constant on the depth of compensation in the Airy-Heiskanen (Airy, 1855; Heiskanen, 1931) and Pratt-Hayford (Pratt, 1855; Hayford, 1909) models and vary laterally in the Vening Meinesz model (Vening Meinesz, 1931). (**a**) Airy model in which isostatic equilibrium is achieved by variations in the thickness of a uniform low-density crust that overlies a high-density mantle. (**b**) Pratt model in which isostatic equilibrium is achieved by lateral changes in the density. (**c**) Vening Meinesz model in which the topography is considered as a load and isostatic equilibrium is achieved by flexural downwarping and upwarping of the crust and mantle over a broad area. In this model, a load is supported by the strength of the lithosphere (which is determined by its effective elastic thickness, T_e) and the buoyancy of the underlying asthenosphere.

(and thereby removing) the positive gravity effect of the topography and the negative gravity effect of the Airy compensation from the observed free-air gravity anomaly. The Root Mean Square (RMS) of the free-air gravity anomaly variation about its mean is 87.8 mGal, while the RMS of the Airy isostatic gravity anomaly is 45.9 mGal. The RMS values show that the free-air gravity anomaly is substantially reduced after a correction for isostasy, and so we may conclude that equilibrium prevails to a significant degree in the northern Indian region.

Despite this success, Figure 2a shows that significant departures still remain in the Airy isostatic anomaly map. The most striking are the positive–negative anomaly "couples" that correlate with the edge of the Tibetan Plateau, particularly its boundary with the Ganges and Tarim foreland basins. The negative anomalies are usually on the basin side while the positive anomalies are on the plateau side, which implies that the Moho is deeper than expected for Airy on the basin side and shallower on the plateau side. Such a Moho geometry can be explained if we

assume that the load of the Himalaya and Kunlun mountains is supported by the strength of the underlying crust and mantle, which flexes over a broad region that extends beyond the mountains themselves. The Vening Meinesz model takes into account these ideas of loading and bending, and Figure 2c shows that when an elastic plate flexure model is used to compute the compensation rather than an Airy model, the RMS of the isostatic anomalies is reduced even further. A useful parameter in the Vening Meinesz model is the effective elastic thickness, T_e, since it is a proxy for the long-term strength of the lithosphere. Figure 2c shows that there is a well-developed minimum in the RMS of the isostatic anomaly for $T_e \sim 30$ km, suggesting this as the best fit T_e estimate for the region (see also Crosby, 2007 who found a similar best fit T_e for the northern Indian region). This is, of course, only an average and so there will be areas that are stronger and areas that are weaker than this. Indeed, Braitenberg et al.



Isostasy, Figure 2 Topography, free-air and isostatic gravity anomalies in the region of the Tibetan plateau. (**a**) *Upper panel*: Topography based on GEBCO (GEBCO, 2003). The map shows the plateau and the flanking Tarim and Ganges foreland basins, the major faults (ATF = Altyn Tagh Fault, MBT = Main Boundary Thrust) and the north–south profile (*thick red line*) plotted in (**b**). *Middle panel*: Free-air gravity anomaly based on the EGM2008 2.5 × 2.5 min grid (Pavlis et al., 2008). *Lower Panel*: Airy isostatic anomaly calculated by subtracting the gravity effect of the topography and its compensating "root" from the observed free-air gravity anomaly. Calculations assume a density of water, crust, and mantle of 1,030, 2,800, and 3,330 kg m^{-3}, respectively, and a zero elevation crustal thickness of 31.2 km. (**b**) Predicted crustal structure based on the Airy and flexure models and free-air and isostatic gravity anomaly profiles of the Tibetan plateau at longitude 85°E. (**c**) Root Mean Square (RMS) of the isostatic anomaly as a function of elastic thickness, T_e, which is a proxy for the long-term strength of the lithosphere. The Moho depth for $T_e = 0$ km corresponds to the predictions of an Airy model that implies a plate with no strength. The minimum RMS and hence the isostatic model that best explains free-air gravity anomaly data in the vicinity of the Tibetan plateau is for a flexure model with $T_e = 30$ km.

(2003) and Jordan and Watts (2005) have already shown, using inverse methods, that T_e is <30 km over the central and northern Tibetan Plateau and >30 km over the Ganges and Tarim foreland basins.

Another way to test isostasy on a regional scale is to compare compilations of the seismically constrained crustal structure directly to the predictions of isostatic models. Figure 3 shows, for example, an 18,000 km long transect crossing the Indian, Atlantic, and Pacific Oceans along which observed and calculated depths to Moho are compared. The observed Moho is based on CRUST2.0 (Bassin et al., 2000), a 2 × 2° grid of P-wave velocity and crustal thickness data derived from controlled source seismic data. The calculated Moho is based on Airy. The figure shows that the calculated Airy Moho tracks the observed seismic Moho well, especially at continental margins.

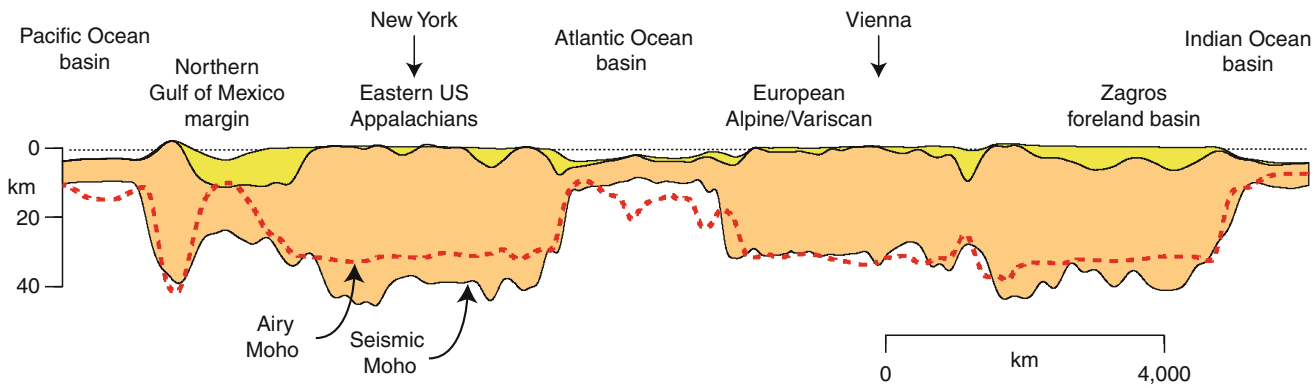
The main departures are in the Atlantic Ocean where the observed Moho depth is shallower than predicted by Airy and beneath the East Coast, U.S. margin and Persian Gulf where the observed Moho depth is deeper. The shallower depths can be explained by our assumption of a relatively high subcrustal oceanic mantle density (3,330 kg m^{-3}), which if reduced (e.g., to 3,180 kg m^{-3}) beneath a mid-ocean ridge would result in a shallower Moho depth and better agreement with the seismic Moho depths. The idea of a relatively low-density subcrustal mantle beneath the ridge suggests a Pratt rather than an Airy model of isostatic compensation. Indeed, Pratt has been assumed in thermal cooling models (Sclater and Francheteau, 1970) for the variation of seafloor depth and heat flow with age and the model may well apply to other cases of thermal isostasy, for example, at mid-plate oceanic swells (Nakiboglu and Lambeck, 1985).

The deeper depths than predicted by Airy are associated with thick sediments at the East Coast, U.S. coastal plain and the Zagros foreland basin. Here, flanking loads due to rifting and thermal contraction and thrusting and folding have flexed the Moho downward over a broad area such that Moho is deeper than would be expected for an Airy model.

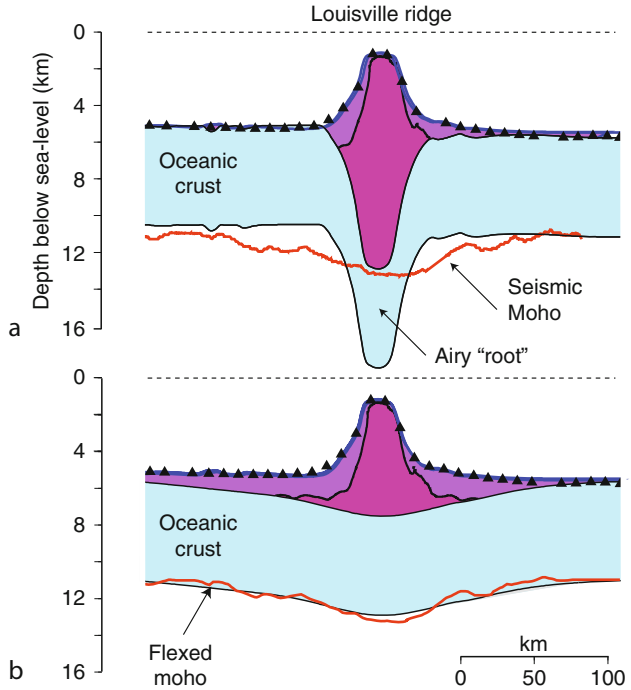
The role of flexure in contributing to crustal structure is well illustrated in Figure 4, which shows the results of a recent seismic reflection and refraction survey of the Louisville Ridge seamount chain near its intersection with the Tonga-Kermadec trench (Contreras-Reyes et al., 2010). The survey utilized a large-volume air gun source array and 39 ocean bottom seismometer receivers spaced 15 km apart were deployed. The source–receiver geometry was such that full ray coverage of the seamount and the underlying oceanic crust and uppermost ~10 km of the mantle was achieved. Figure 4a shows that the seismically constrained Moho cannot be explained by an Airy model that predicts too deep a Moho beneath the seamount and too shallow a Moho in flanking regions. The best fit between observed and calculated Moho depths (Figure 4b) is for a flexure model of isostasy with T_e =10 km. Such a model explains well the observed depth to Moho beneath *both* the seamount edifice and flanking regions.

Current investigations

One of the most important results to have emerged from isostatic studies over the past 40 years has been the link that has been made between the T_e derived from flexural modelling and the long-term strength of the lithosphere. Data from experimental rock mechanics, for example, suggest that the strength of oceanic lithosphere is limited by brittle deformation in its uppermost part and by ductile flow in its lowermost part (Goetze and Evans, 1979). This strength profile implies that loads applied to the surface (or base) of the lithosphere will be supported partly by the brittle and ductile strength of the lithosphere and partly by a central "core" that deforms elastically. It has been shown that oceanic T_e reflects the thickness of this "core," which will be small in regions of young crust, large loads, and high curvature of bending and high in regions of old



Isostasy, Figure 3 Comparison of the crustal structure based on seismic refraction data to the predicted crustal structure assuming an Airy model along an 18,000 km long great circle profile that extends from the Pacific Ocean in the west across the Atlantic Ocean to the Indian Ocean in the east. The seismic refraction data is based on CRUST2.0 http://igppweb.ucsd.edu/~gabi/rem.html and the Airy model is based on the same parameters as assumed in Figure 2.



Isostasy, Figure 4 Comparison of the observed crustal structure derived from seismic refraction data (Contreras-Reyes et al., 2010) to the predictions of an Airy and flexure model of isostasy along a NNE-SSW trending profile of the Louisville Ridge seamount chain, southwest Pacific Ocean. Black filled triangles indicate the location of the ocean bottom seismometers that were used to constrain the crustal structure. *Thick solid red lines* show the depth to the seismically constrained Moho. *Thin solid black line* separating the purple and pink filled regions show the 6.0 km/s P-wave iso-velocity contour. *Purple fill* = volcanoclastic material. *Pink fill* = volcano edifice. *Blue fill* = deformed oceanic crust. (**a**) Airy model based on the same parameters as assumed in Figure 2. (**b**) Flexure model with T_e =10 km.

crust, small loads, and low curvature (Watts and Burov, 2003). Oceanic T_e is in the range 0–40 km, and since the oceanic crust is only ~8 km thick we may conclude that not only the crust but also the mantle is involved in supporting long-term geological loads.

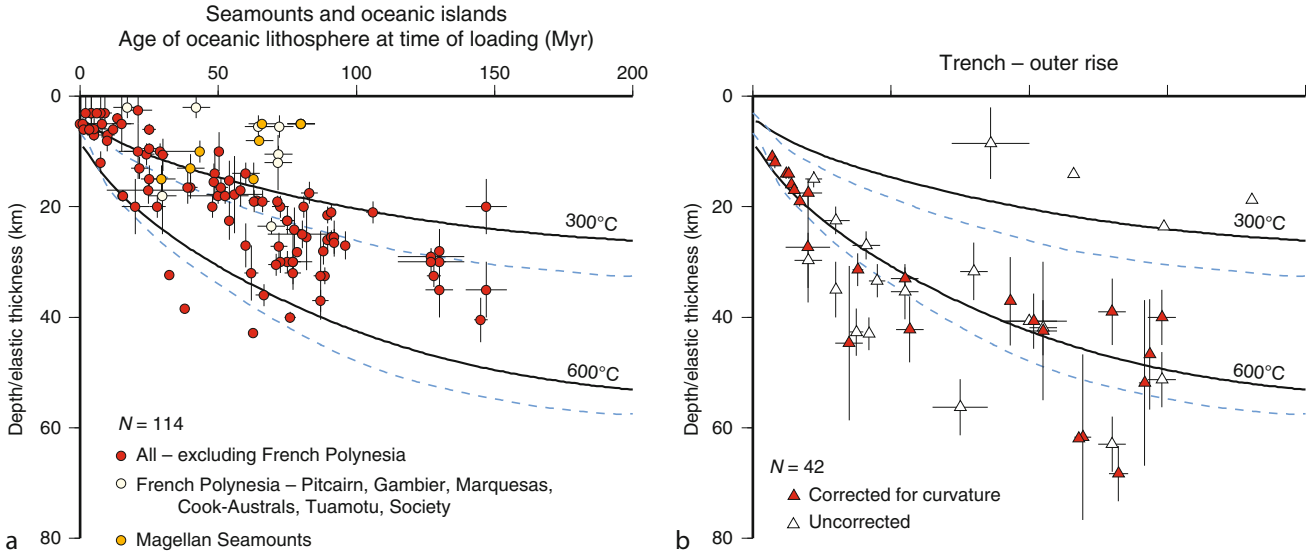
Figure 5 shows that T_e at seamounts and ocean islands and deep-sea trench and outer rise systems increases with the age of the lithosphere at the time of loading, being small at volcano and trench loads that form on young seafloor (e.g., on-ridge) and large at loads on old seafloor (e.g., off-ridge). Thus, as the oceanic lithosphere cools, subsides, and increases in age following its creation at a spreading center, it becomes more rigid in the way that it responds to loads. Interestingly, most seamount and oceanic island values fall within the 300–600°C isotherm envelope based on the cooling plate model of Parsons and Sclater (1977) while most trench and outer rise values require a higher controlling isotherm in the region 450–800°C. This difference in the range of controlling isotherm can be explained by difference in the load age, which is small (essentially present day) at deep-sea trench-outer rise and large (~1.5–100 Myr) at seamounts and oceanic islands. Indeed, a load age dependency explains much of the scatter in plots of T_e vs. the age of the lithosphere at the time of loading. Other possible contributors are uncertainties in load, infill, and mantle densities; thermal perturbations due to hot and cold spots; and yielding in regions of large loads and high curvature.

Most of the T_e estimates plotted in Figure 5 are based on a forward modelling approach in which the gravity or geoid anomaly, surfaces of flexure, or vertical motion history are calculated and compared to observations. The best fit T_e is then determined by a "trial and error" comparison of the observed and calculated. While forward modelling is a satisfactory way to estimate T_e, the number of sites where information on both load and plate age are available is limited.

Technological improvements in satellite radar altimetry and shipboard multibeam (swath) bathymetry data acquisition systems have led to a rapid increase in the number of T_e estimates. Watts et al. (2006), for example, used satellite-derived gravity data together with a reciprocal admittance technique to predict bathymetry for different values of T_e. By comparing observed and calculated bathymetry these workers obtained 9,758 individual estimates of T_e, 291 of which were at sites where both load and plate age were known. Bry and White (2007) used satellite-derived gravity and shipboard bathymetry data to determine T_e along a number of fixed length profiles of deep-sea trench-outer rise systems, while Kalnins and Watts (2009) used satellite-derived gravity and shipboard bathymetry data and an admittance technique to determine the T_e structure of the western Pacific region.

One problem with all these spectral methods is the influence of the analysis window size on the recovered T_e. Windows need to be large enough to resolve the largest T_e and small enough to avoid "contamination" from more than one type of geological feature. Watts et al. (2006) and Bry and White (2007), for example, found a greater than expected scatter in plots of T_e vs. the age of the lithosphere at the time of loading with no clear indication of a controlling isotherm. Kalnins and Watts (2009), however, used a moving window technique and found that the controlling isotherm varied regionally; this may explain some of the scatter in plots of T_e against load and plate age.

The physical meaning of T_e and its relationship to load and plate age is much less clear in the continents than it is in the oceans. Unlike oceanic lithosphere, continental lithosphere has a multilayered rheology (Burov and Diament, 1995; Meissner and Strehlau, 1982) and there may be more than one brittle and ductile layer that supports an individual load or unload. In this situation, T_e is not a particular depth in the lithosphere but a measure of its integrated strength. Continental T_e depends on both the curvature of flexure (which depends in turn on the size of the load or unload) and how well individual competent layers are coupled together, being low for high curvature



Isostasy, Figure 5 Plot of T_e vs. the age of the oceanic lithosphere at the time of loading. (**a**) Seamounts and oceanic islands and (**b**) deep-sea trench and outer rise systems. The data is based on Table 6.1 in Watts (2001) with the addition of Freedman and Parsons (1986), McQueen and Lambeck (1989), and 15 other studies carried out since 2000. *Solid and dashed lines* show the calculated depth of the 300°C and 600°C isotherms based on the cooling plate models of Parsons and Sclater (1977) and McKenzie et al. (2005), respectively.

and poorly coupled systems and high for low curvature and well coupled systems (Watts and Burov, 2003).

As in the oceans, continental T_e has been estimated using both forward and inverse modelling techniques. Most forward modelling estimates are based on fitting either the gravity anomaly or the surfaces of flexure (as inferred, e.g., from basement depth data) at rift-type, strike-slip, intra-cratonic, and foreland basins to model predictions. Sometimes, both gravity anomaly and depth to basement data are available and this yields the most reliable estimates (Haddad and Watts, 1999). Continental T_e estimates range from 5 to 100 km and no simple relationship has yet been found between T_e and load and plate age (Watts, 2001). The data is suggestive, however, of some age dependence. Archaen and Proterozoic cratons, for example, are generally associated with high T_e values ($>$70 km) while Phanerozoic and younger mountain belts generally have lower values (25–60 km) (Pérez-Gussinyé and Watts, 2005). Young rifted margins (e.g., South China Sea) appear to have lower T_e values than old margins (e.g., Amazon) (Watts et al., 2009) and this has been interpreted as the result of heating during continental breakup that weakens the lithosphere, and cooling following rifting that strengthens it (Burov and Poliakov, 2001). Plume-influenced rifts (e.g., Afar, North Atlantic) are associated with the lowest T_e values (Ebinger et al., 1989; Watts and Fairhead, 1997), but it is not known how much strength a weakened rifted lithosphere is capable of regaining.

According to the Wilson Cycle, ocean basins open and close and compressional mountain belts develop on or close to the site of extensional rifted margins. Isostatic studies of the foreland basins that flank mountain belts suggest that they may inherit the T_e structure of the underlying rifted margin (Lin and Watts, 2002). The foreland basin may be either narrow or wide, depending on whether it formed on a young or old margin. The very high ($>$70 km) T_e values that are associated with some foreland basins (e.g., central Andes) may be explained by the fact that the fold and thrust belts, the main driving load for foreland basin subsidence, will sometimes override the rifted margin altogether and "telescope" onto the more rigid cratonic interior, thereby inheriting its value.

As in the oceans, inverse modelling approaches have provided new insights into continental T_e, especially its spatial and temporal variation. The first studies (e.g., Dorman and Lewis, 1970; Banks et al., 1977; McNutt and Parker, 1978) were based on a Bouguer admittance, Fourier transform periodogram method, which yielded a number of low values. Forsyth (1985), however, pointed out that the low values might be a result of the low coherence that existed in continental regions between gravity and topography because of the presence of subsurface (i.e., buried) loads that were uncorrelated with surface topographic loads. Subsurface loads are difficult to define and so he suggested using the Bouguer coherence, rather than the Bouguer admittance, because it was less sensitive to the ratio of surface to subsurface loading. Forsyth's method was widely applied (e.g., Zuber et al., 1989) and as Bechtel et al. (1990) showed for North America, it yields low as well as high values, especially in cratonic regions. Subsequent studies have incorporated new techniques, including maximum entropy estimators (Lowry and Smith, 1995) and multitapers (McKenzie

and Fairhead, 1997). McKenzie and Fairhead (1997) recommended using the free-air admittance rather than Bouguer coherence because the surface topography, unlike buried loads, is a known load, the gravity effect of which should not really be removed from the free-air gravity anomaly. These workers argued that results based on Forsyth's Bouguer coherence technique maybe biased by the effects of erosion and his method will yield overestimates of T_e rather than true values. Pérez-Gussinyé and Watts (2005), however, used both Bouguer coherence and free-air admittance and found the highest values in Europe (>70 km) over the Archaen and Early Proterozoic (<1.5 Ga) East European craton (part of Baltica) and the lowest values (10–45 km) over the flanking Caledonian, Variscan, and Alpine orogenic belts (Figure 6). Similar results were found for North and South America (Kirby and Swain, 2009; Pérez-Gussinyé et al., 2007), Australia (Simons et al., 2000), and Africa (Pérez-Gussinyé et al., 2009), and a consensus seems to be emerging that the continental lithosphere is capable of showing both great weakness and great strength. Interestingly though, cratons are not always associated with a uniformly high T_e, and there are regions within them that have low values.

In another development, wavelets have been used instead of spectral methods. Wavelets have an advantage over spectral methods in that they are local in space and scale so that the relationship between gravity anomaly and topography can, in principal, be examined over a range of wavelengths without the need for annular averaging over discrete windows. They yield similar results as spectral methods, although the recovered T_e tends to be smoother (Audet and Mareschal, 2007; Sacek and Ussami, 2009; Stark et al., 2003; Tassara et al., 2006).

Irrespective of the methods used, caution needs to be applied when interpreting continental T_e data. This is because it is not clear whether the T_e deduced from inverse modelling refers to the same loading and unloading schemes as are assumed in forward modelling. In forward modelling, specific loads (e.g., fold and thrusts) and flexures are considered, and the T_e deduced usually refers to a particular thermo-tectonic event in the geological past (e.g., a rift or orogenic event). Inverse methods, by way of contrast, "see" *all* the loads that are acting on the lithosphere, and these include not only past loads, but sedimentary loads and erosional unloads, some of which continue to shape the Earth's surface up to the present day.

What is clear from all these studies is that Earth's lithosphere has a rich T_e structure. In the oceans, T_e varies by up to ±15 km over horizontal distances of 500–1,000 km (Watts et al., 2006). Spatial variations are even larger in the continents. The compilations of Pérez-Gussinyé and coworkers, for example, suggest ±25 km over similar distances. In many places (e.g., western Canada and USA), weak regions with T_e as low as 10 km abut strong regions with $T_e > 70$ km (Fluck et al., 2003). The consequences of such strength variations are profound, especially as they impact on the patterns of glacial isostatic rebound (Whitehouse et al., 2006), solid earth tides (Mantovani et al., 2005), and the terrane structure of continental lithosphere (Pérez-Gussinyé et al., 2007).

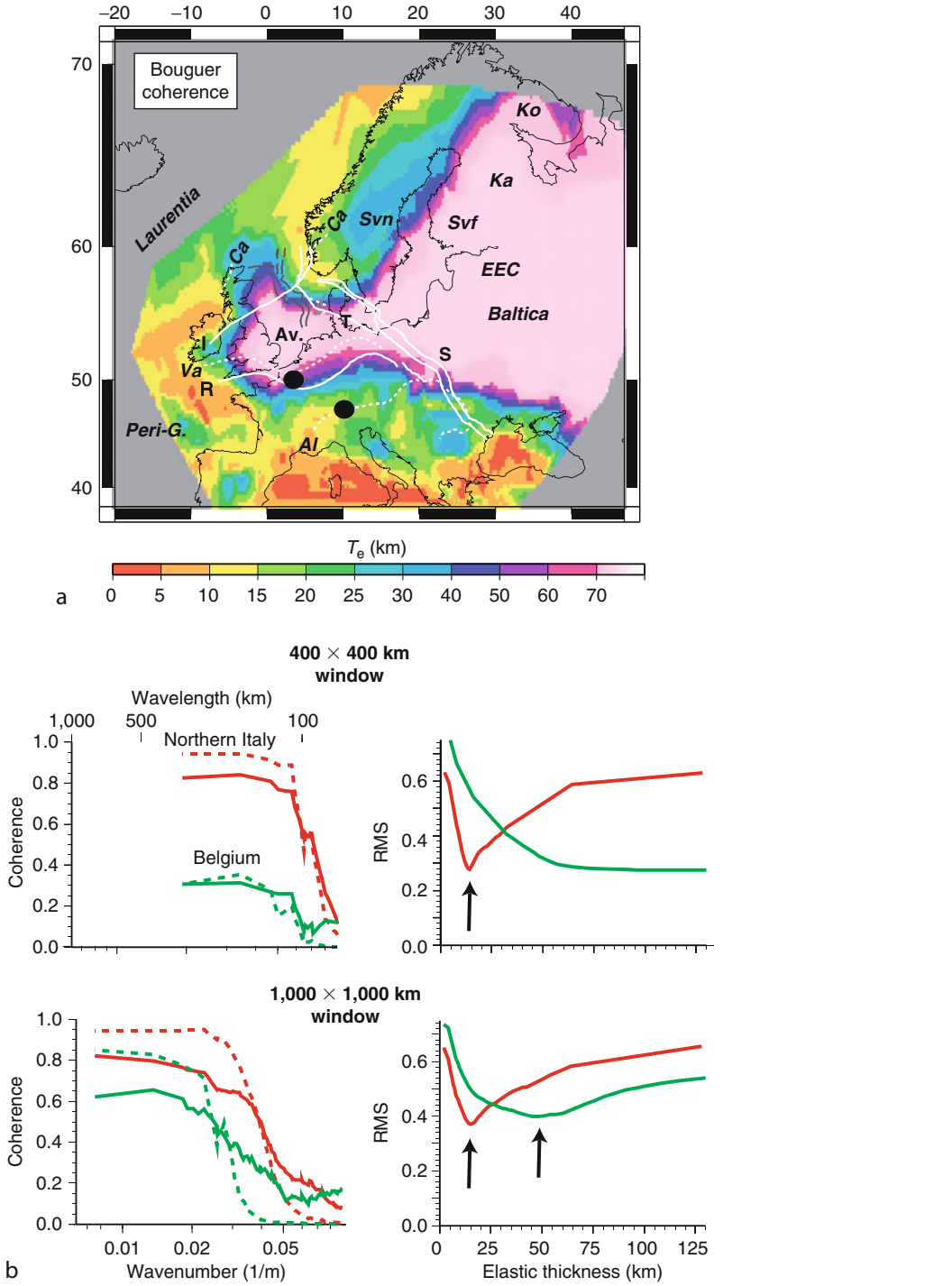
Current controversies and gaps in knowledge

Isostatic studies have never been very far from controversy as demonstrated by the vigorous debates on local vs. regional models that took place at the turn of the twentieth century. Central to these debates was the applicability of isostatic models to particular geological features, and certain "schools of thought" emerged among geologists, geophysicists, and geodesists. There were even debates either side of the Atlantic when an Airy model was preferred by Europeans and a Pratt model by North Americans. Today, there is general agreement on the relative role of local and regional models of isostasy. Debate is focused instead on the flexure model, particularly T_e, what it means and what it tells us about the long-term strength of the lithosphere, the timescales of isostatic adjustment, and the geological processes that shape the top and base of the lithosphere.

The long-term strength of the lithosphere

In the oceans, there is widespread agreement on oceanic T_e and what it indicates about the long-term strength of the lithosphere. While there is scatter in plots of oceanic T_e vs. plate age, most of it can be explained by viscoelastic stress relaxation (Watts and Zhong, 2000), curvature and yielding (McNutt and Menard, 1982), and spatial variations in the controlling isotherms that determine T_e (Kalnins and Watts, 2009).

There has not, however, been the same agreement concerning continental T_e estimates and their relationship to age. The controversy has recently come to a head in the so-called crème brûlée vs. jelly sandwich debate (Burov and Watts, 2006). Proponents of the crème brûlée model follow McKenzie and Fairhead (1997) and Jackson (2002) who believe that because T_e is small and similar in thickness to the seismogenic layer, the strength of the lithosphere resides mainly in the uppermost part of the crust. In their view, the lower crust and uppermost mantle are weak, which explains why deep earthquakes are rare. Proponents of the jelly sandwich model, however, follow Meissner (1986) who argues that the lithosphere is "jelly like" with a strong upper crust, a weak lower crust, and a strong uppermost mantle. This model is compatible with the results of seismic reflection profiling, which show a layered lower continental crust, and with the results of flexure studies, which show a wide range of T_e values. Low T_e is indicative of thin competent layers while high T_e values reflect thick competent layers. Very high T_e values (>70 km) are also permissible and occur when the upper and lower crust or the lower crust and mantle are coupled. Deep earthquakes are rare, not because the mantle is too weak but because it is too strong, and the tectonic stresses required to initiate faulting are not large enough.



Isostasy, Figure 6 The T_e structure of Europe as derived from the coherence between the observed Bouguer gravity anomaly and topography as a function of wavelength. (Modified from Pérez-Gussinyé and Watts [2005].) (**a**) T_e and its relationship to the terrane structure. Thin white lines show major sutures. *I* = Iapetus. *R* = Rheic. *A* = Alpine. *T* = Thor. *S* = Sorgenfrei-Tornquist. The tectonic provinces are Ko, Kola; Ka, Karelia; and Svf, Svecofennian. *Thin dashed lines* show the Caledonian (Ca), Variscan (Va), and Alpine (Al) deformation fronts. Peri-G. = Peri-Gondwanaland. Av = Avalonia. EEC = East European Craton. *Black filled circles* show the location of the coherence and RMS plots in (**b**). (**b**) Examples of the observed and predicted Bouguer coherence and RMS between the observed (*solid line*) and calculated (*dashed line*) coherence for a 400 and 1,000 km analysis window centered on Belgium and northern Italy. Note that northern Italy appears as a region of low T_e in both windows, but the high T_e region of Belgium only has a clear RMS minima in the larger window.

The debate has led to a reexamination of the methodologies used to estimate continental T_e and its physical meaning. Pérez-Gussinyé et al. (2004) showed that it is necessary when comparing observed and calculated free-air admittances to first window them using the same multitaper estimator, while Crosby (2007) showed that in order to recover T_e in cratonic shield regions large window sizes (up to 1,500 km) are needed; otherwise, T_e may be underestimated. Watts and Burov (2003) pointed out that while T_e may well be low and similar to the seismogenic layer thickness in some tectonic settings (e.g., Basin and Range, western USA), the two parameters are not the same. T_e reflects the integrated strength of the lithosphere while the seismogenic layer is the thickness of the uppermost part of the crust that responds to stresses by faulting and earthquakes. Finally, Handy and Brun (2004) argued that a strong uppermost mantle is required in order to explain the integrity of sinking slabs in subduction zones and the deep structure of rifted margins and orogens.

The continental T_e debate has proved a useful discussion point in isostatic studies. It highlights the importance of T_e while at the same time illustrates gaps in our knowledge about the Earth's lithosphere, especially regarding its long-term rheology. Other gaps are as follows.

Timescales of isostatic adjustment

Isostasy varies temporally, as well as spatially. Exactly how slow or fast isostasy operates is of fundamental importance to our understanding of the Earth's vertical motion history as well as to topics such as sea-level change. For example, Bloom (1967) pointed out that the height of a wave-cut notch above present day sea level depends not only on the amplitude of sea-level change, but how quickly isostatic equilibrium is achieved. If isostasy is fast then a sea-level rise will have time to load the crust and upper mantle and so the height of the notch will be low due to regional subsidence. If, on the other hand, isostasy is slow then there will be time and a notch will be cut at a higher elevation.

Several attempts have been made to estimate the time scales of isostatic adjustment associated with load shifts on the top or base of the Earth's crust. Watts (1978), for example, compared the T_e derived from oceanic flexure studies to the thermal and seismic thickness of the lithosphere and concluded that on loading there must be a rapid reduction in the thickness of the mechanical layer that supports a load as the lithosphere relaxes from its short-term seismic to its long-term elastic thickness. Beaumont (1979) used an elastic–plastic model based on the creep laws of olivine to argue that while oceanic lithosphere was essentially elastic on long-time scales, continental lithosphere continued to relax on timescales of up to several tens of million years. Willett et al. (1985) used a viscoelastic plate that overlies an inviscid substrate to model continental T_e data and showed that on loading there would be a weakening of the lithosphere that reduced the thickness of the mechanical layer by as much as a factor of 3–8, depending on the initial thermal structure of the lithosphere that was assumed. And finally, Watts and Zhong (2000) used the actual temperature structure of the oceanic lithosphere constrained from seafloor depth and heat-flow data to calculate a viscosity profile and then examined how the equivalent T_e would vary following volcano loading. They showed that because viscosity increases from low values in the lower lithosphere to high values in the upper lithosphere, the lithosphere relaxes from the bottom-up such that the reduction in the mechanical layer thickness that supports a load is initially fast and then slows with load age.

Figure 7 compares the T_e for a number of seamounts and ocean islands in the world's ocean basins to predicted curves based on a multilayered viscoelastic model. The model assumes an asthenosphere viscosity of 10^{21} Pa s and a lithosphere viscosity that increases from 10^{21} to10^{35} Pa s, depending on thermal age. The figure shows a rapid decrease in the thickness of the lithosphere that supports a load and hence a decrease in its strength. This weakening results in an increase in subsidence and a narrowing of the zone of subsidence with increase in load age. Eventually, the strength reduction slows and the thickness that supports a load approaches a steady state value.

The continents have a wider range of load and unload ages than the oceans and there are elastic thickness estimates for short-term earthquake loading (2–100 s) through intermediate (~10 ka) glacial isostatic rebound to long-term (>1 Ma) flexure. There are regions (e.g., western and eastern USA) where elastic layer thickness has been estimated for two or more timescales. These data suggest that the relaxation observed in oceanic lithosphere may also characterize the continental lithosphere, albeit over a much wider range of relaxation timescales. For example, the Hebgen Lake (Nishimura and Thatcher, 2003), Lake Bonneville (Nakibogluand Lambeck, 1983), and western Cordillera (Lowry et al., 2000) loads that reflect ages of 0–120 s, 10 ka, and ~60 Ma are associated with elastic layer thicknesses of 38, 30, and 5 km, respectively. There is therefore a decrease in T_e with increase in load age. The decrease is similar in form to the ocean loading cases modelled by Watts and Zhong (2000), but the viscosity of the asthenosphere that is required is less and $\sim 3 \times 10^{18}$ Pa s. This viscosity results in a rapid initial and a slow final weakening of the lithosphere and a transition between the two regimes at load ages of ~30 a. Interestingly, the transition appears to vary regionally such that loads on the tectonic western North America relax quicker than those emplaced on the more stable parts of eastern North America.

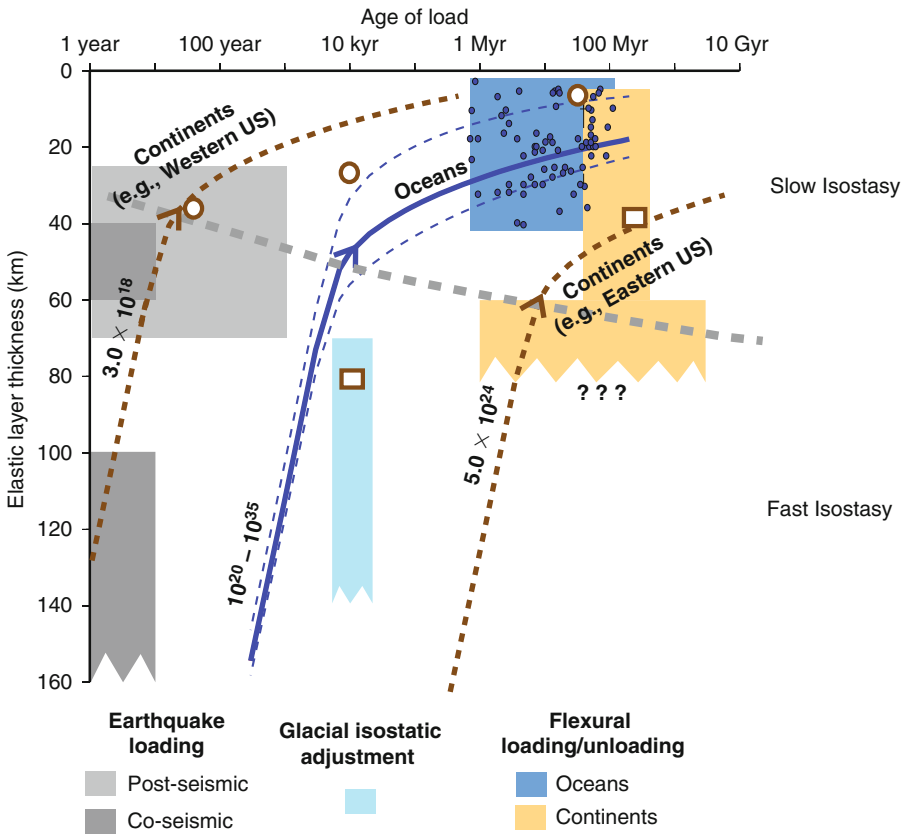
Isostasy, landscape evolution, and mantle dynamics

The behavior of the lithosphere and how it interacts with the cryosphere, hydrosphere, and atmosphere above and the asthenosphere below is a fundamental question that needs to be addressed if we are to fully understand

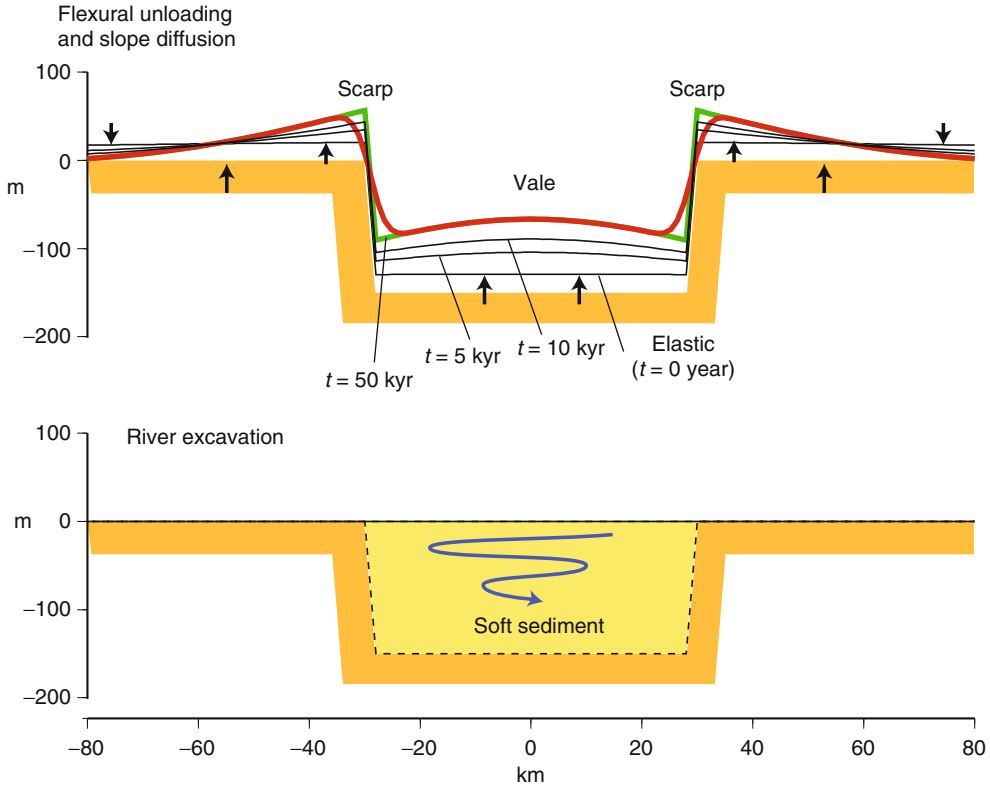
geological processes and construct realistic models. We know, for example, from flexure studies that the lithosphere behaves as a low-pass filter in the way that it responds to loads and unloads, suppressing the short-wavelength deformation and passing the long-wavelengths. The physical properties of the lithosphere therefore smooth out the effects of surface and subsurface processes such as those involved in landscape evolution and mantle convection, making it difficult to see the full effect of these processes.

Isostasy has played a major role in the development of the landscape, both in the constructive forces such as tectonics that build it and the destructive forces such as erosion that destroy it. Examples include the flexures that follow mechanical unloading on normal faults (ten Brink and Stern, 1992), fault scarp retreat (Gilchrist and Summerfield, 1990), and glacial erosion (Pelletier, 2004; Stern et al., 2005). Because flexure involves rock uplift and subsidence it is capable of modifying pre-existing drainage patterns, especially river knick points and base levels. This is particularly well seen in the rim uplifts that flank rifted and transform margins (Tucker and Slingerland, 1994) where flexure is a major factor in controlling the drainage divide that separates scarp and dip streams. Other examples have been described from the Cotswold Hills, southern England (Lane et al., 2007), where rim uplifts flank regions of fluvial incision and excavation.

Figure 8 illustrates the isostatic adjustment that would occur following fluvial incision and excavation of a 150 m deep and 60 km wide region of relatively low-density sediments. The calculations assume that the



Isostasy, Figure 7 Plot of elastic layer thickness vs. the age of load for oceanic and continental lithosphere. (Modified from Watts [2007].) The *thick solid blue line* and *thin dashed blue lines* show the predicted change in elastic layer thickness for young, intermediate, and old oceanic lithosphere according to Watts and Zhong (2000). *Filled blue circles* show the data from individual seamount and ocean island loads, which range in age from ~1 to 100 Myr. The data indicate that by ~10 kyr oceanic loads have undergone most of their relaxation. The *thick dashed brown lines* show the predicted change in elastic thickness for continental lithosphere. *Filled gray rectangles* are based on earthquake loading (post-seismic and co-seismic), *filled blue rectangles* are based glacial isostatic adjustment, and *filled beige and yellow rectangles* are based on continental flexure studies (see Watts, 2007 for references). *Open symbols* show elastic layer thickness estimates from the same tectonic region (*circles* = western USA, *squares* = eastern USA) based on the data in Nakibogluand Lambeck (1983), DiDonato et al. (2000), Lowry et al. (2000), Nishimura and Thatcher (2003), and Pazzaglia and Gardner (1994). The data suggest two modes of relaxation that correspond to "fast" and "slow" isostasy. Interestingly, the relaxation time that separates these two modes varies between oceans and continents with western USA relaxing much quicker (~10 year) than the oceans and eastern USA relaxing much slower (~70 Myr).



Isostasy, Figure 8 Simple model for flexure of the lithosphere due to river incision and excavation. The model assumes that the lithosphere responds to loads (and unloads) initially as a thick elastic plate (T_e = 80 km), then as a viscous material, and finally as a thin, essentially elastic, plate (T_e =10 km). The response shows an initial long-wavelength low amplitude uplift, a transient short-wavelength high amplitude uplift and subsidence, and a final short-wavelength high amplitude uplift and subsidence that together reflects the thinning of the lithosphere from its initial short-term thickness to its long-term elastic thickness. The *thick solid red line* shows the final topography after it has been modified by slope diffusion (assumed erosion time = 50 ka, diffusion constant = 50 $m^2\ a^{-1}$).

lithosphere behaves as a viscoelastic plate, similar to the model illustrated in Figure 7 and used by Watts and Zhong to calculate the time-dependant flexure that follows seamount loading. The main model features are an initial thick-plate elastic response, a viscous time-dependant response, and a final thin-plate, essentially elastic, response. The figure shows that lithospheric weakening due, for example, to load-induced stress relaxation has the capacity to create topography that locally either enhances or subdues the more regional deformation associated with the initial short-term response. Karner (1984) showed that topography is also created during lithospheric strengthening when, for example, the oceanic lithosphere is thermally rejuvenated (and weakened) at mid-plate hotspot swells and subsequently cools. The topography in this case subdued the original swell.

Incision and excavation, together with refilling, are common features of the fluvial response to glacial and interglacial cycles. Isostasy will therefore have a major role to play in the deformations that result. For example, incision due to meltwater-charged rivers is a feature of glacial periods when the sea level was low while valley-filling and delta construction are features of interglacials when the sea level was high. Repeated episodes of incision and valley fill will therefore resurface a landscape, creating in the process isostatic rim uplift and subsidence, as seems to have occurred in the region of the Mississippi River where it debauches in to the Gulf of Mexico (Blum et al., 2008). One consequence of this resurfacing is that the apparent sea level may vary along-strike the rifted margin of the northern Gulf of Mexico such that the Mississippi River valley in Texas subsides relative to the adjacent Alabama coast.

The Earth's lithosphere is deformed not only by surface geological loads and unloads, but also by subsurface loads, some of which originate within the crust while others have their source in the subcrustal mantle. Like surface loads, subsurface loads are associated with mass excesses and deficiencies and create flexures of the Moho and other crustal layers that are manifest in topography and gravity anomaly data. For example, subsurface loads due to intracrustal thrusting, obducted crustal blocks, and ophiolite emplacement cause flexures that contribute to both topography and gravity anomalies (e.g., Royden and Karner, 1984).

Irrespective, the tendency for isostatic loads and unloads to be compensated means that their gravity anomaly will be reduced and may approach zero at the longest wavelengths. This is most clearly seen in the free-air admittance that shows that Airy, Pratt, and flexure models have a distinct "roll-over" that approaches zero at the longest wavelengths because of isostasy (Watts, 2001). The existence in oceanic regions of admittances in the range 5–20 mGal/km at wavelengths 1,000–4,000 km (Crosby, 2007; Watts, 2007; Wieczorek, 2007) is therefore of much geodynamic interest. Watts (2007) has shown that the long-wavelength admittance remains approximately constant at ~18 mGal/km in the central Pacific Ocean as the analysis window size is increased from 1,000 to 4,000 km. He argued that the long-wavelength admittance cannot be attributed to isostasy and so must reflect dynamic processes in the subcrustal mantle.

It has been known for some time that the Earth is associated with both long-wavelength topography and gravity anomalies. The long-wavelength topography that cannot be attributed to isostatic processes was dubbed "dynamic topography" by Hager et al. (1985). It has been shown that dynamic topography is a significant contributor to the long-wavelength gravity and geoid anomaly field as well to the subsidence and uplift history of sedimentary basins, especially foreland-type (Liu and Nummedal, 2004), rift-type (Muller et al., 2000), and some intra-cratonic basins (Coakley and Gurnis, 1995) where slab-induced motions in the deep mantle may have perturbed the overlying plates. However, we are still far from fully understanding the spatial extent, magnitude, and origin of dynamic topography. More isostatic studies are therefore needed because they have the potential to separate the effects of lithospheric plate flexure from topography and gravity and geoid anomaly data and isolate the planform of motions in the Earth's mantle. The development of improved gravity anomaly and topography field data, new spectral analysis techniques, and better geodynamical models offer the most promise of addressing this problem in the future.

Conclusions

Isostasy is a principle that is central to the Earth Sciences. It was first formulated in the 1880s to describe the tendency of the crust to float, in hydrostatic equilibrium, on its denser substrate such that light areas stand at a greater elevation than dense areas. Today, we consider isostasy as a highly idealized state that the crust and upper mantle approaches, but rarely achieves. Nevertheless, the comparison of observed gravity anomalies to the predictions of local and regional isostatic models has led to a new understanding about the thermal and mechanical properties of the lithosphere. We now know, for example, that the lithosphere responds to long-term (i.e., $>10^5$ years) loading and unloading in a similar manner as would a strong elastic plate that overlies a weak fluid substrate. In the oceans, the thickness of the elastic layer that supports a volcano load increases with the age of the lithosphere at the time of loading. In the continents, the relation between elastic thickness and age is not as clear. However, there is evidence that older Archaen and Early Proterozoic cratons have high values of the elastic thickness (i.e., >70 km) while younger Middle/late Proterozoic and Phanerozoic orogenic belts and rifts have lower values. Data from experimental rock mechanics suggest that the elastic thickness derived from isostatic studies is a proxy for the long-term strength of the lithosphere. The elastic thickness differs from the seismogenic layer thickness, which is the thickness of the uppermost part of the lithosphere that responds to stresses by faulting and earthquakes and the seismic "lid" thickness, which is the mechanical thickness that supports short-term loads and unloads. Isostatic studies suggest that the timescales of isostatic adjustment varies regionally such that tectonically active areas respond to loads and unloads much quicker than cratonic areas. The behavior of the lithosphere on seismic to geologic timescales is important to take into account when modelling geological processes, especially those associated with the climate-induced glacial loading and unloading cycles.

Bibliography

Airy, G. B., 1855. On the computation of the effect of the attraction of mountain-masses, as disturbing the apparent astronomical latitude of stations of geodetic surveys. *Philosophical Transactions of the Royal Society*, **145**, 101–104.

Audet, P., and Mareschal, J.-C., 2007. Wavelet analysis of the coherence between Bouguer gravity and topography: application to the leastic thickness anisotropy in the Canadian Shield. *Geophysical Journal International*, **168**, 287–298.

Banks, R. J., Parker, R. L., and Huestis, S. P., 1977. Isostatic compensation on a continental scale: local versus regional mechanisms. *Geophysical Journal of the Royal Astronomical Society*, **51**, 431–452.

Bassin, C., Laske, G., and Masters, G., 2000. The current limits of resolution for surface wave tomography in North America. *Transactions of the American Geophysical Union*, **81**, F897.

Beaumont, C., 1979. On rheological zonation of the lithosphere during flexure. *Tectonophysics*, **59**, 347–365.

Bechtel, T. D., Forsyth, D. W., Sharpton, V. L., and Grieve, R. A. F., 1990. Variations in effective elastic thickness of the North American lithosphere. *Nature*, **343**, 636–638.

Bloom, A., 1967. Pleistocene shorelines: a new test of isostasy. *Geological Society of America Bulletin*, **78**, 1477–1494.

Blum, M. D., Tomkin, J. H., Purcell, A., and Lancaster, R. R., 2008. Ups and downs of the Mississippi delta. *Geology*, **36**, 675–678.

Braitenberg, C., Wang, Y., Fang, J., and Hsu, H. T., 2003. Spatial variations of flexure parameters over the Tibet-Quinghai plateau. *Earth and Planetary Science Letters*, **205**, 211–224.

Bry, M., and White, N., 2007. Reappraising elastic thickness variation at oceanic trenches. *Journal of Geophysical Research*, **112**, doi: 10.1029/2005JB004190.

Burov, E. B., and Diament, M., 1995. The effective elastic thickness (Te) of continental lithosphere: what does it really mean? *Journal of Geophysical Research*, **100**, 3895–3904.

Burov, E. B., and Poliakov, A., 2001. Erosion and rheology controls on synrift and postrift evolution: verifying old and new ideas using a fully coupled numerical model. *Journal of Geophysical Research*, **106**, 16461–16481.

Burov, E. B., and Watts, A. B., 2006. The long-term strength of continental lithosphere: jelly sandwich or crème brûlée? *GSA Today*, **16**, 4–10.
Coakley, B., and Gurnis, M., 1995. Far-field tilting of Laurentia during the Ordovician and constraints on the evolution of a slab under an ancient continent. *Journal of Geophysical Research*, **100**, 613–6327.
Contreras-Reyes, E., Grevemeyer, I., Watts, A. B., Planert, L., Flueh, E. R., and Peirce, C., 2010. Crustal intrusion beneath the Louisville hotspot track. *Earth and Planetary Science Letters*, **289**, 323–333.
Crosby, A. G., 2007. An assessment of the accuracy of admittance and coherence estimates using synthetic data. *Geophysical Journal International*, **171**, 25–54.
Delaney, J. P., 1940. Leonardo da Vinci on isostasy. *Science*, **91**, 546–547.
Di Donato, G., Mitrovica, J. X., Sabadini, R., and Vermeersen, L. L. A., 2000. The influence of a ductile crustal zone on glacial isostatic adjustment: geodetic observables along the U.S. East Coast. *Geophysical Research Letters*, **27**, 3017–3020.
Dorman, L. M., and Lewis, B. T. R., 1970. Experimental isostasy1: theory of determination of the Earth's response to a concentrated load. *Journal of Geophysical Research*, **75**, 3357–3365.
Dutton, C. E., 1882. Physics of the Earth's crust; by the Rev. Osmond Fisher. *American Journal of Science*, **23**(136), 283–290.
Dutton, C. E., 1889. On some of the greater problems of physical geology. *Bulletin of the Philosophical Society of Washington*, **2**, 51–64.
Ebinger, C. J., Bechtel, T. D., Forsyth, D. W., and Bowin, C. O., 1989. Effective elastic plate thickness beneath the East African and Afar plateaus and dynamic compensation of the uplifts. *Journal of Geophysical Research*, **94**, 2883–2901.
Fisher, O., 1881. *Physics of the Earth's Crust*. London: Macmillan, p. 299.
Fluck, P., Hyndman, R. D., and Lowe, C., 2003. Effective elastic thickness Te of the lithosphere in western Canada. *Journal of Geophysical Research*, **108**, 2430, doi:2410.1029/2002JB002201.
Forsyth, D. W., 1985. Subsurface loading and estimates of the flexural rigidity of continental lithosphere. *Journal of Geophysical Research*, **90**, 12623–12632.
Freedman, A. P., and Parsons, B., 1986. Seasat-derived gravity over the Musician seamounts. *Journal of Geophysical Research*, **91**, 8325–8340.
GEBCO, 2003. The GEBCO digital atlas, centenary edition (CD-ROM).
Gilchrist, A. R., and Summerfield, M. A., 1990. Differential denudation and flexural isostasy in formation of rifted-margin upwarps. *Nature*, **346**, 739–742.
Goetze, C., and Evans, B., 1979. Stress and temperature in the bending lithosphere as constrained by experimental rock mechanics. *Geophysical Journal of the Royal Astronomical Society*, **59**, 463–478.
Haddad, D., and Watts, A. B., 1999. Subsidence history, gravity anomalies, and flexure of the northeast Australian margin in Papua New Guinea. *Tectonics*, **18**, 827–842.
Hager, B. H., Clayton, R. W., Richards, M. A., Cromer, R. P., and Dziewonski, A. M., 1985. Lower mantle heterogeneity, dynamic topography and the geoid. *Nature*, **313**, 541–545.
Handy, M. R., and Brun, J.-P., 2004. Seismicity, structure and strength of the continental lithosphere. *Earth and Planetary Science Letters*, **223**, 427–441.
Hartley, R., Watts, A. B., and Fairhead, J. D., 1996. Isostasy of Africa. *Earth and Planetary Science Letters*, **137**, 1–18.
Hayford, J. F., 1909. *The Figure of the Earth and Isostasy from Measurements in the United States*. Washington, DC: Government Printing Office, p. 178.
Heiskanen, W., 1931. Isostatic tables for the reduction of gravimetric observations calculated on the basis of Airy's hypothesis. *Bulletin Géodésique*, **30**, 110–129.
Jackson, J., 2002. Strength of the continental lithosphere: time to abandon the jelly sandwich? *GSA Today*, **12**, 4–10.
Jordan, T. A., and Watts, A. B., 2005. Gravity anomalies, flexure and the elastic thickness structure of the India-Eurasia collisional system. *Earth and Planetary Science Letters*, **236**, 732–750.
Kalnins, L. M., and Watts, A. B., 2009. Spatial variations in effective elastic thickness in the western Pacific Ocean and their implications for Mesozoic volcanism. *Earth and Planetary Science Letters*, **286**, 89–100.
Karner, G. D., 1984. Thermally induced residual topography of oceanic lithosphere. *Nature*, **318**, 327–331.
Kirby, J. F., and Swain, C. J., 2009. A reassessment of spectral Te estimation in continental interiors: the case of North America. *Journal of Geophysical Research*, **114**, doi:10.1029/2009JB006356.
Lane, N., Watts, A. B., and Farrant, A., 2007. An analysis of the Cotswolds topography: insights into the landscape response to denudational isostasy. *Journal of the Geological Society of London*, **165**, 85–103, doi:110.1144/0016-76492006-76492179.
Lin, A. T., and Watts, A. B., 2002. Origin of the West Taiwan basin by orogenic loading and flexure of a rifted continental margin. *Journal of Geophysical Research*, **107**, 10.1029/2001JB000669.
Liu, S.-F., and Nummedal, D., 2004. Late Cretaceous subsidence in Wyoming: quantifying the dynamic component. *Geology*, **32**, 397–400.
Lowry, A. R., and Smith, R. B., 1995. Strength and rheology of the western U.S. Cordillera. *Journal of Geophysical Research*, **100**, 17947–17963.
Lowry, A. R., Ribe, N. M., and Smith, R. B., 2000. Dynamic elevation of the Cordillera, western United States. *Journal of Geophysical Research*, **105**, 23371–23390.
Mantovani, M. S. M., Shukowsky, W., de Freitas, S. R. C., and Neves, B. B. B., 2005. Lithosphere mechanical behaviour inferred from tidal gravity anomalies: a comparison of Africa and South America. *Earth and Planetary Science Letters*, **230**, 397–412.
McKenzie, D. P., and Fairhead, D., 1997. Estimates of the effective elastic thickness of the continental lithosphere from Bouguer and free-air gravity anomalies. *Journal of Geophysical Research*, **102**, 27523–27552.
McKenzie, D., Jackson, J., and Priestley, K., 2005. Thermal structure of oceanic and continental lithosphere. *Earth and Planetary Science Letters*, **233**, 337–349.
McNutt, M. K., and Menard, H. W., 1982. Constraints on yield strength in the oceanic lithosphere derived from observations of flexure. *Geophysical Journal of the Royal Astronomical Society*, **71**, 363–394.
McNutt, M. K., and Parker, R. L., 1978. Isostasy in Australia and the evolution of the compensation mechanism. *Science*, **199**, 773–775.
McQueen, H. W. S., and Lambeck, K., 1989. The accuracy of some lithospheric bending parameters. *Geophysical Journal*, **96**, 401–413.
Meissner, R., 1986. Twenty years of deep seismic reflection profiling in Germany – a contribution to our knowledge of the nature of the lower Variscan crust. *Journal of Geological Society of London.* Special publication number, 24, pp. 1–10.
Meissner, R., and Strehlau, J., 1982. Limits of stresses in continental crusts and their relation to the depth-frequency distribution of shallow earthquakes. *Tectonics*, **1**(1), 73–89.
Muller, R. D., Lim, V. S. L., and Isern, A. R., 2000. Late Tertiary tectonic subsidence on the northeast Australian passive margin: response to dynamic topography? *Marine Geology*, **162**, 337–352.
Nakiboglu, S. M., and Lambeck, K., 1983. A reevaluation of the isostatic rebound of Lake Bonneville. *Journal of Geophysical Research*, **88**, 439–447.

Nakiboglu, S. M., and Lambeck, K., 1985. Comments on thermal isostasy. *Journal of Geodynamics*, **2**, 51–65.
Nishimura, T., and Thatcher, W., 2003. Rheology of the lithosphere inferred from postseismicuplift following the1959 Hebgen Lake earthquake. *Journal of Geophysical Research,* **108**, doi:10.1029/2002JB002191.
Parsons, B. E., and Sclater, J. G., 1977. An analysis of the variation of ocean floor bathymetry and heat flow with age. *Journal of Geophysical Research*, **82**, 803–827.
Pavlis, N. K., Holmes, S. A., Kenyon, S. C., and Factor, J. K., 2008. An Earth gravitational model to degree 2160: EGM2008. EGU General Assembly, Vienna.
Pazzaglia, F. J., and Gardner, T. W., 1994. Late Cenozoic flexural deformation of the middle U.S. Atlantic passive margin. *Journal of Geophysical Research*, **99**, 12143–12157.
Pelletier, J. D., 2004. Estimate of three-dimensional flexural-isostatic response to unloading: Rock uplift due to late Cenozoic glacial erosion in the western United States. *Geology*, **32**, 161–164.
Pérez-Gussinyé, M., and Watts, A. B., 2005. The long-term strength of Europe and its implications for plate-forming processes. *Nature*, **436** (21 July), doi:10.1038/nature03854.
Pérez-Gussinyé, M., Lowry, A. R., Watts, A. B., and Velicogna, I., 2004. On the recovery of effective elastic thickness using spectral methods: examples from synthetic data and from the Fennoscandia Shield. *Journal of Geophysical Research*, **109**, doi: 10.1029/2003JB002788.
Pérez-Gussinyé, M., Lowry, A. R., and Watts, A. B., 2007. Effective elastic thickness of South America and its implications for intracontinental deformation. *Geochemistry, Geophysics, Geosystems*, **8**, doi:10.1029/2006GC001511.
Pérez-Gussinyé, M., Metois, M., Fernandez, M., Verges, J., Fullea, J., and Lowry, A. R., 2009. Effective elastic thickness of Africa and its relationship to other proxies for lithospheric structure and surface tectonics. *Earth and Planetary Science Letters*, **287**, 152–167.
Pratt, J. H., 1855. On the attraction of the Himalaya mountains, and of the elevated regions beyond them, upon the plumb line in India. *Philosophical Transactions of the Royal Society*, **145**, 53–100.
Pratt, J. H., 1864. Speculations on the constitution of the Earth's crust. *Proceedings of the Royal Society of London*, **13**, 253–276.
Royden, L., and Karner, G. D., 1984. Flexure of the lithosphere beneath Apennine and Carpathian foredeep basins: evidence for an insufficient topographic load. *American Association of Petroleum Geologists*, **68**, 704–712.
Sacek, V., and Ussami, N., 2009. Reappraisal of the effective elastic thickness for the sub-Andes using 3-D finite element flexural modelling, gravity and geological constraints. *Geophysical Journal International*, **179**, 778–786.
Sclater, J. G., and Francheteau, J., 1970. The implications of terrestrial heat flow observations on current tectonic and geochemical models of the crust and upper mantle of the Earth. *Geophysical Journal of the Royal Astronomical Society*, **20**, 509–542.
Simons, F. J., Zuber, M. T., and Korenaga, J., 2000. Isostatic response of the Australian lithosphere: estimation of effective elastic thickness and anisotropy using multitaper spectral analysis. *Journal of Geophysical Research*, **105**, 19163–19184.
Stark, C. P., Stewart, J., and Ebinger, C. J., 2003. Wavelet transform mapping of the effective elastic thickness and plate loading: validation using synthetic data and application to the study of the South African tectonics. *Journal of Geophysical Research*, **108**, doi:10.1029/2001JB000609.
Stern, T. A., Baxter, A. K., and Baxter, P. J., 2005. Isostatic rebound due to glacial erosion within the Transantartic Mountains. *Geology*, **33**, 221–224.
Tassara, A., Swain, C., Hackney, R., and Kirby, J., 2006. Elastic thickness structure of South America estimated using wavelets and satellite-derived gravity data. *Earth and Planetary Science Letters*, **253**, 17–36.
ten Brink, U., and Stern, T., 1992. Rift flank uplifts and hinterland basins: comparison of the Transantarctic Mountains with the Great Escarpment of Southern Africa. *Journal of Geophysical Research*, **97**, 569–585.
Tucker, G. E., and Slingerland, R. L., 1994. Erosional dynamics, flexural isostasy, and long-lived escarpments: a numerical modeling study. *Journal of Geophysical Research*, **99**, 12229–12243.
Vai, G. B., 2006. Isostasy in Luigi Ferdinando Marsili's manuscripts. In Vai, G. B., and Caldwell, W. G. E. (eds.), *The Origins of Geology in Italy.* Geological Society of America Special Paper 411.
Vening Meinesz, F. A., 1931. Une nouvelle methode pour la réduction isostatique régionale de l'intensité de la pesanteur. *Bulletin Géodésique*, **29**, 33–51.
Watts, A. B., 1978. An analysis of isostasy in the world's oceans: 1. Hawaiian-Emperor Seamount Chain. *Journal of Geophysical Research*, **83**, 5989–6004.
Watts, A. B., 2001. *Isostasy and Flexure of the Lithosphere*. Cambridge: Cambridge University Press, p. 458.
Watts, A. B., 2007. An overview. In Watts, A. B. (ed.), *Treatise of Geophysics. Vol. 6. Crust and Lithosphere Dynamics*. Amsterdam: Elsevier, pp. 1–48.
Watts, A. B., and Burov, E. B., 2003. Lithospheric strength and its relationship to the elastic and seismogenic thickness. *Earth and Planetary Science Letters*, **213**, 113–131.
Watts, A. B., and Fairhead, J. D., 1997. Gravity anomalies and magmatism at the British Isles continental margin. *Journal of the Geological Society of London*, **154**, 523–529.
Watts, A. B., and Zhong, S., 2000. Observations of flexure and the rheology of oceanic lithosphere. *Geophysical Journal International*, **142**, 855–875.
Watts, A. B., Sandwell, D. T., Smith, W. H. F., and Wessel, P., 2006. Global gravity, bathymetry, and the distribution of submarine volcanism through space and time. *Journal Geophysicl Research*, **111**, doi:10.1029/2005JB004083.
Watts, A. B., Rodger, M., Peirce, C., Greenroyd, C. J., and Hobbs, R. W., 2009. Seismic structure, gravity anomalies, and flexure of the Amazon continental margin, NE Brazil. *Journal of Geophysical Research*, **114**, doi: 10.1029/2008JB006259.
Whitehouse, P., Latychev, K., Milne, G. A., Mitrovica, J. X., and Kendall, R., 2006. Impact of 3-D Earth structure on Fennoscandian glacial isostatic adjustment: Implications for space-geodetic estimates of present-day crustal deformations. *Geophysical Research Letters*, **33**, doi:10.1029/2006GL026568.
Wieczorek, M. A., (2007). Gravity and topography of the terrestrial planets, in Treatise of Geophysics. In Sphon, T. (ed.), Planets and Moons. Elseiver, Vol. 10, pp. 165–206.
Willett, S. D., Chapman, D. S., and Neugebauer, H. J., 1985. A thermo-mechanical model of continental lithosphere. *Nature*, **314**, 520–523.
Zuber, M. T., Bechtel, T. D., and Forsyth, D. W., 1989. Effective elastic thickness of the lithosphere and mechanisms of isostatic compensation in Australia. *Journal of Geophysical Research*, **94**, 13919–13930.

Cross-references

Earth's Structure, Continental Crust
Earthquakes and Crustal Deformation
Geodynamics
Gravity Field of the Earth
Heat Flow, Seafloor: Methods and Observations
Isostasy, Thermal
Lithosphere, Continental

Lithosphere, Continental: Thermal Structure
Lithosphere, Mechanical Properties
Lithosphere, Oceanic
Lithosphere, Oceanic: Thermal Structure
Mantle Viscosity
Sedimentary Basins

ISOSTASY, THERMAL

David S. Chapman, Derrick Hasterok
Department of Geology and Geophysics, The University of Utah, Salt Lake City, UT, USA

Definition

Isostasy: (Greek: isos "equal," stasis "standstill") term used in geology to refer to the condition of gravitational equilibrium such that the rigid outer part of the earth's crust and uppermost mantle "float" at an elevation which depends on its density. This concept explains how different topographic heights can exist at the earth's surface.
Compositional isostasy: the part of isostasy traceable to density variations arising from compositional (rock type and mineralogy) differences.
Thermal isostasy: the part of isostasy traceable to density variations arising from temperature differences and thermal expansion within the rock column.
Isostatic equilibrium: the condition whereby a certain area of crust and upper mantle reaches the state of isostasy such that a hydrostatic equilibrium exists at depth below a level of compensation. The region is said to be in *isostatic equilibrium*.

Introduction

The condition in the earth whereby variations in crustal thickness and density determine the elevation of the earth's solid surface is known as isostasy and has been appreciated for a century and a half (Pratt, 1855; Airy, 1855; Watts, 2001, for general review). Local isostasy requires a hydrostatic equilibrium condition at a depth below a compensation level, implying that the integral of density over a rock column is constant. When density differences arise from differences in rock composition (rock type, mineralogy), the isostatic condition is referred to as compositional isostasy. It is compositional isostasy, and in particular the density contrast across the crust–mantle boundary, that dominates isostasy discussions in textbooks. If, on the other hand, density differences arise from different thermal states of the lithosphere and thermal expansion of rock, the condition is properly called thermal isostasy. This article is devoted to thermal isostasy, computation of thermal isostasy effects, and examples of thermal isostasy in oceanic regions and on the continents.

Thermal isostasy

Rock density is influenced by temperature through thermal expansion. If rock composition is identical in two vertical columns through the lithosphere, the warmer column has a lower density than the cooler lithosphere. Applying the isostatic condition of hydrostatic equilibrium at some depth of compensation leads to a prediction of elevation changes for the two regions.

Consider two regions of different lithospheric thermal states. Temperature differences between the two regions represented by a regional geotherm $T(z)$ and a reference geotherm $T_{\mathrm{ref}}(z)$, respectively, multiplied by a thermal expansion coefficient α_{V} and integrated over depth predict an elevation change, $\Delta\varepsilon_{\mathrm{T}}$, given by

$$\Delta\varepsilon_{\mathrm{T}} = \alpha_{\mathrm{V}} \int_0^{z_{\max}} [T(z) - T_{\mathrm{ref}}(z)]dz. \quad (1)$$

The maximum depth of integration, $z_{\max}$, is the depth at which the colder geotherm converges to a mantle adiabat assumed to be identical for both lithospheric regions.

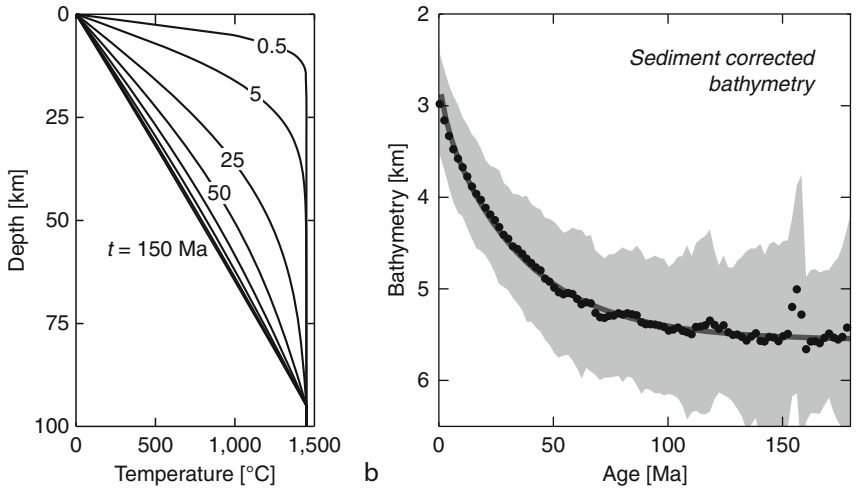
Most studies use a coefficient of thermal expansion of $3.0 \times 10^{-5}\,\mathrm{K}^{-1}$ within the crust, consistent with the values for major crustal forming rocks including granite and gabbro. Within the mantle, a value of thermal expansion of $3.2 \times 10^{-5}\,\mathrm{K}^{-1}$is used.

Oceanic thermal isostasy

The clearest example of thermal isostasy occurs in oceanic regions as hot, near-molten lithosphere at a spreading ridge cools over tens of millions of years leading to thermal contraction, increasing rock density, and subsidence of the seafloor by about 3,000 m. It is important to note that lateral compositional variations are small in oceanic regions. Oceanic crust about 7 km thick is formed at a spreading center and its thickness and composition are modified only in minor ways over its maximum 165 My existence. The thermal state of oceanic lithosphere, on the other hand, varies from a near molten state at the ridge to extremely cold lithosphere after up to 165 My of cooling. These contrasting thermal states have consequences for the rigidity of the lithosphere but in particular for its density and therefore elevation.

Almost since the discovery of plate tectonics and seafloor spreading, thermal isostasy has been invoked to explain the bathymetry of the oceans as a function of age. Figure 1a illustrates the time evolution of an oceanic geotherm for a cooling plate model. The initial condition is 1,450°C (Stein and Stein, 1993); other cooler values lead to similar subsidence curves if thermal expansion values are adjusted accordingly. The ocean lithosphere initially cools rapidly as shown by the large difference between the 0.5 My geotherm and the 5 and 25 My geotherms (Figure 1a). At longer times, the cooling slows and the geotherms reach near steady state shown by the 150 My geotherm. The average temperature change in the lithosphere between zero age at the ridge and about 150 My seafloor is 725 K.

Bathymetry is proportional to the integrated cooling of the oceanic lithosphere (Equation 1) incorporating as well the changing load of water as the seafloor subsides. This



Isostasy, Thermal, Figure 1 Thermal evolution of oceanic lithosphere after formation at a spreading oceanic ridge system. (**a**) Temperature-depth profiles as a function of lithosphere age in millions of years. (**b**) Subsidence of the seafloor as a function of seafloor age. Solid dots are data in 2 My age bins. Gray shading represents ±1 standard deviation about the mean bathymetry.

subsidence (Figure 1b) is explained very well by a simple one-dimensional (1-D) model of lithospheric cooling. A midocean ridge, or zero age lithosphere, has on average a depth of 2.7 km, a consequence of both its thermal state, crustal thickness and composition. But as the lithosphere ages and cools, temperature changes but crustal thickness and composition do not. Cooling over the first 9 My causes a subsidence of 1 km to a depth of 3.7 km. At a lithospheric age of 36 My, the solid surface has subsided a further kilometer to a water depth of 4.7 km. As the lithosphere reaches its thermal equilibrium, the bathymetry also approaches its equilibrium value of 5.5 km.

Continental thermal isostasy

On continents, thermal isostasy has been used successfully to examine the evolution of regions that mimic oceanic spreading, such as continental rifts and elements of provinces with extensive volcanism or back-arc regions (McKenzie, 1978; Jarvis and McKenzie, 1980; Brott et al., 1981; Lachenbruch and Morgan, 1990; Hyndman et al., 2005). However, direct thermal effects on continental elevation are difficult to discern because of the potential masking effect of compositional variations in lithospheric density and thickness.

The elevation effect due to compositional variation can be estimated by a simple isostatic calculation using continental extremes. A mountainous region with a crustal thickness of 50 km and a density of 2,800 kg m^{-3} (granodiorite) would have an elevation 5 km higher than a rift province with a crustal thickness of 25 km and a density of 2,900 kg m^{-3} (gabbro). Both columns assume a similar mantle density of 3,340 kg m^{-3}. The potentially large compositional effect can easily mask the effect of thermal isostasy. Therefore, in order to isolate the effect of thermal isostasy in the continents, it is necessary first to remove compositional isostatic effects.

Normalizing compositional elevation

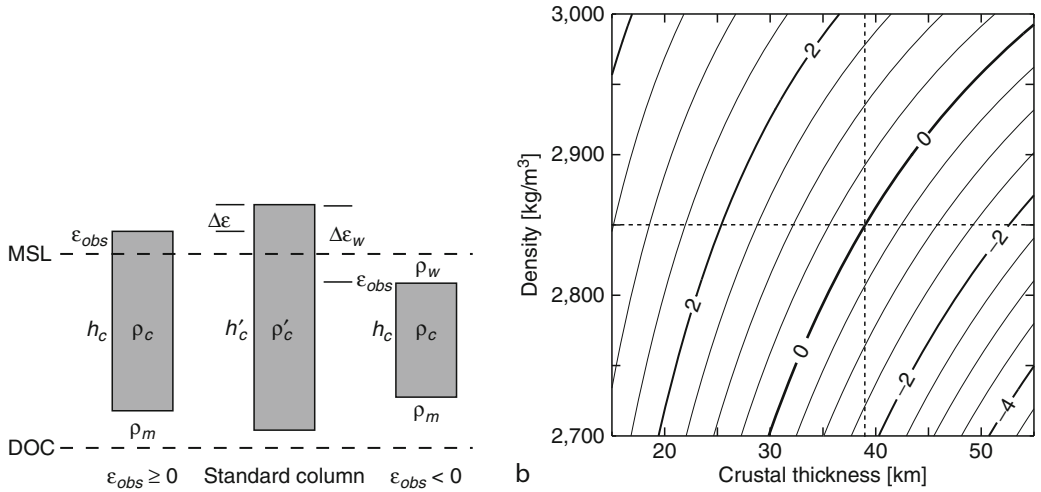
Compositional variations involving both crustal thickness and density are removed with a simple isostatic correction to the observed elevation (Han and Chapman, 1995). The adjustment normalizes any crustal column to an arbitrary crustal standard. In refining this method, Hasterok and Chapman (2007a) used a standard crustal thickness h'_c of 39 km and a standard crustal density ρ'_c of 2,850 kg m^{-3}. These values represent the average crustal thickness and density of North American provinces. A lithospheric mantle density of 3,340 kg m^{-3} is used based on the xenolith derived estimate by Griffin et al. (1999) for Proterozoic lithosphere.

Consider a crustal column with a crustal thickness, h_c, and crustal density, ρ_c, at an elevation of ε_{obs} (Figure 2a). If the density and thickness of the observed crustal column are adjusted to match the standard density and thickness, the elevation would change by an amount $\Delta\varepsilon_c$ (Han and Chapman, 1995) given by

$$\Delta\varepsilon_C = h'_c\left(1 - \frac{\rho'_c}{\rho_m}\right) - h_c\left(1 - \frac{\rho_c}{\rho_m}\right) \quad (2)$$

where ρ_m is the density of the mantle. If the province surface is below sea level (e.g., continental shelf), an additional term involving the water depth (where ε_{obs} is the bathymetry) and seawater density, ρ_w, is required and the elevation adjustment becomes

$$\Delta\varepsilon_C = h'_c\left(1 - \frac{\rho'_c}{\rho_m}\right) - h_c\left(1 - \frac{\rho_c}{\rho_m}\right) - \frac{\varepsilon_{obs}\rho_w}{\rho_m}. \quad (3)$$



Isostasy, Thermal, Figure 2 Compositional isostasy adjustments to continental elevation. (**a**) Cartoon illustrating parameters used in compositional correction for elevation. MSL is mean sea level and DOC is the depth of compensation. The observed crustal columns (*left* for regions above sea level and *right* for continental margins below sea level) are adjusted to an arbitrary standard crust (*center*) of thickness 39 km and density 2,850 kg m^{-3}. (**b**) Nomogram contouring the compositional elevation adjustment for observed crustal thickness and average crustal density. Contour interval is 0.5 km. Dashed lines for crustal thickness of 39 km and density of 2,850 kg m^{-3} represent the standard crustal column. Zero line is locus of values for which the elevation adjustment is zero.

By adding the elevation adjustment to the observed elevation one arrives at the final adjusted elevation, ε_{adj}, given by

$$\varepsilon_{adj} = \varepsilon_{obs} + \Delta\varepsilon_C. \quad (4)$$

Three important physical parameters must be estimated in order to make the compositional adjustment: ε_{obs}, h_c, and ρ_c. The observed elevation is obtained from GTOPO30 for elevations above sea level, and from bathymetric maps for elevations below sea level. Crustal thickness is obtained from 1-D whole crustal V_P models. Crustal density is estimated by using empirical velocity–density (V_P-ρ) relationships. The lateral dimension of each province analyzed varies from ~70 km in some provinces in the Western US to >500 km in shield provinces. See Hasterok and Chapman (2007a,b) for details.

The magnitudes of compositional elevation adjustments made using Equations 2 and 3 are shown in Figure 2b. The zero contour is the locus of crustal thickness and density combinations for which no crustal adjustment is made and therefore passes through the standard crust of thickness 39 km and density 2,850 kg m^{-3}. Over the extreme range of crustal thicknesses (20–60 km) and densities (2,700–3,000 kg m^{-3}), compositional elevation adjustments can be as much as 4 km. Regions that are thinner and/or more dense than the standard crust receive positive elevation adjustments, whereas regions that are thicker and/or less dense than the standard crust receive negative elevation adjustments. Most adjustments in Figure 2b fall in the range +/−1 km. From this nomogram we can also determine the uncertainty in elevation adjustment introduced by variations in crustal thickness and density. For example, to achieve an uncertainty of ± 250 m requires an uncertainty in crustal thickness less than 3 km and in density less than 43 kg m^{-3}.

Continental thermal state

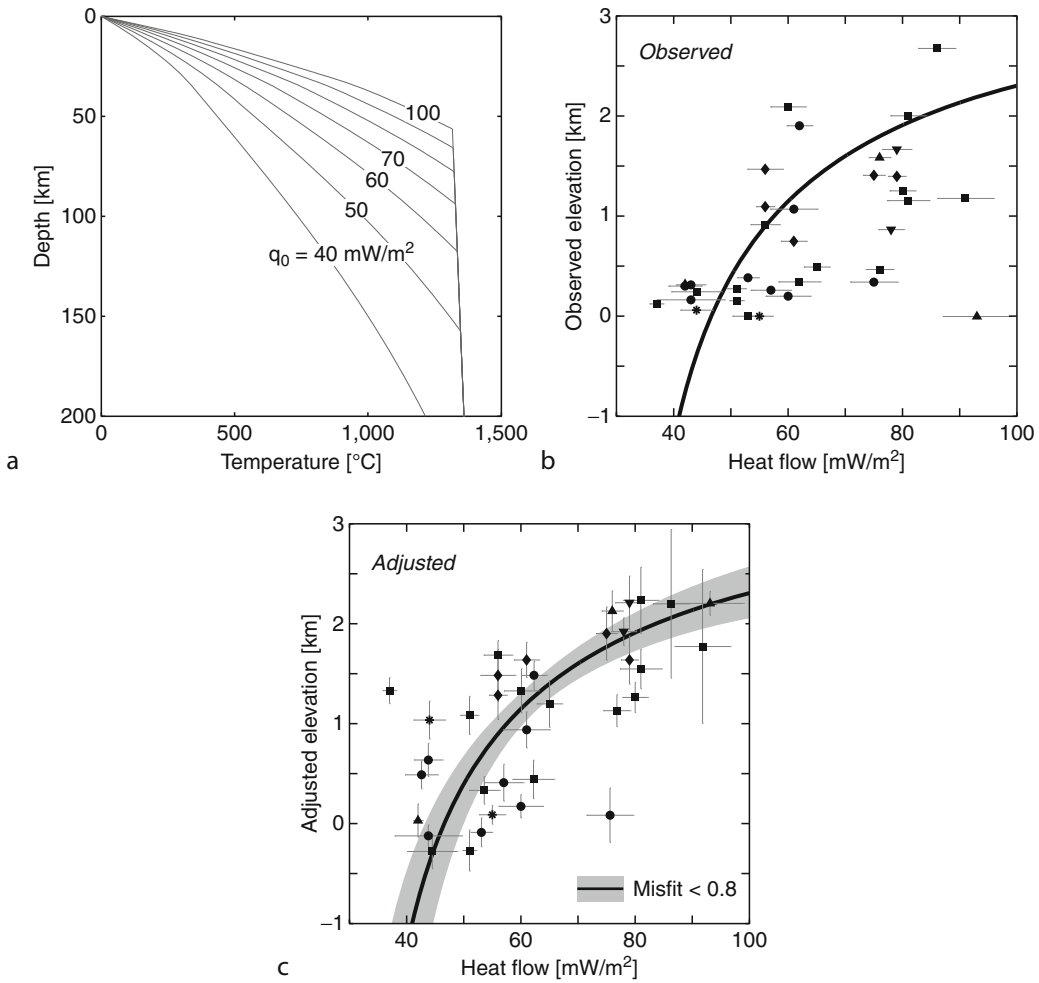
The thermal state of continental lithosphere is described by a set of geotherms. Whereas oceanic geotherms may be identified in terms of age, continental geotherms do not correlate well with rock age. Instead, the most important controlling parameter for continental geotherms is surface heat flow (Pollack and Chapman, 1977b; Chapman and Pollack, 1977; Blackwell, 1971; Lachenbruch and Sass, 1977; Rao et al., 1982).

There exist a number of geotherm models of the continental crust (Blackwell, 1971; Lachenbruch and Sass, 1977; Chapman, 1986; Chapman and Furlong, 1992). Many other models are based on locally derived estimates of thermal conductivity and/or heat production. Although locally and xenolith-derived models may be best in the region for which they are developed, they may be poor estimates when extended to other regions. Therefore, we feel it is best to use a consistent geothermal model based on the fewest parameters necessary to demonstrate the usefulness of this method. We chose the method of Chapman (1986) and Chapman and Furlong (1992) to compute our geotherms because they yield a set of geotherms that are well documented, commonly used, and yield temperatures which fall between other warm and cool models.

The properties of the geotherm model include a lithospheric structure divided into an upper and lower crust and mantle lithosphere. Thermal conductivity within the crust is determined using a pressure-/temperature-dependent relationship of Chapman (1986). Thermal

conductivities within the upper and lower crust are initially set as 3.0 and 2.65 W m^{-1} K^{-1}, respectively, corresponding generally to felsic and mafic crystalline rocks at STP conditions. Mantle lithospheric thermal conductivities include a radiation term at temperatures above 500 K. Surface heat production is determined assuming a 40% radiogenic contribution to surface heat flow with a characteristic depth of 8 km (Pollack and Chapman, 1977a). This characteristic depth is also used as the decay length of an exponentially decreasing function used to compute heat generation within the upper crust. Heat production of the lower crust is set as 0.4 μW m^{-3} consistent with many studies of exposed granulite terranes. Mantle heat production is assumed to be 0.02 μW m^{-3}, a lower value justified by Chapman (1986) and Chapman and Furlong (1992). Temperatures within the asthenosphere are computed using an adiabat with a potential temperature of 1,300 C and increase with depth of 0.3 C km^{-1}. Steady state, 1-D, conductive continental geotherms computed for surface heat flow values between 40 and 100 mW m^{-2} are shown in Figure 3a. Details of the geotherm calculation and boundary conditions used are given by Chapman (1986) and further discussed in Hasterok and Chapman, 2007a).

The thermal structure in some regions differs from the geotherm models used in this study as a result of nonsteady state processes and differences in thermophysical properties. However, we have chosen to



Isostasy, Thermal, Figure 3 Thermal isostasy on continents. (**a**) Generalized family of continental geotherms parametric in surface heat flow values from 40 to100 mW m^{-2}. Geotherms are truncated by the 1,300°C mantle adiabat. (**b**) Observed elevation versus heat flow for individual tectonic provinces of North America. Symbols denote dominant style of tectonic province: circle (shield), star (basin), square (collision), diamond (volcanic), triangle (extension). Error bars are one standard deviation of the mean for heat flow. The geotherms in (**a**) together with Equation 1 are used to calculate a theoretical thermal isostasy relationship shown as a bold line. (**c**) Elevation adjusted for compositional effects versus heat flow for the same tectonic provinces in (**b**). Elevation error bars are predicted from a Monte Carlo analysis of uncertainties in the seismic velocity to density conversion. Data trend confirms expected 3,000 m of thermal isostasy effect.

keep our thermal models as simple as possible, interpreting the more difficult and uncertain parameters, that is, heat production and thermal conductivity, as residuals from a reference thermal isostatic model.

Revealing thermal isostasy on continents

Because the thermal state of continents is not clearly related to the age of the lithosphere as it is in the oceans, one cannot confirm continental thermal isostasy by simply plotting elevation of continental regions against age of the province. Other approaches must be explored. One such approach examined elevation as a function of heat flow for several tectonic provinces around the globe (Han and Chapman, 1995; Nagihara et al., 1996). Hasterok and Chapman (2007a; 2007b) followed the same approach in much greater detail and made an assessment of the uncertainties related to removing compositional isostatic effects.

A thermal isostatic curve for the continents, derived from the geotherm family in Figure 3a and Equation 1 is shown in Figure 3b. We use a reference geotherm corresponding to a surface heat flow of 47 mW m^{-2} and assign a lithosphere having this thermal state an elevation of 0 km. Although the zero elevation reference heat flow at 47 mW m^{-2} is assumed here, the actual zero elevation could easily be set at another value. This initial assumption for the zero-elevation reference heat flow is refined further by Hasterok and Chapman (2007b). Within the range of surface heat flow for continental tectonic provinces (40–90 mW m^{-2}), the predicted elevation range resulting from thermal isostasy is approximately 3 km. This prediction is similar to the oceanic bathymetry difference between hot ridges and cold abyssal plains.

Figure 3b also demonstrates how much continental thermal isostasy is obscured by compositional effects. There is little correlation between observed elevation of 36 North American tectonic provinces and the continental thermal isostatic curve. Much of this scatter is the result of variations in crustal density and thickness between the different tectonic provinces.

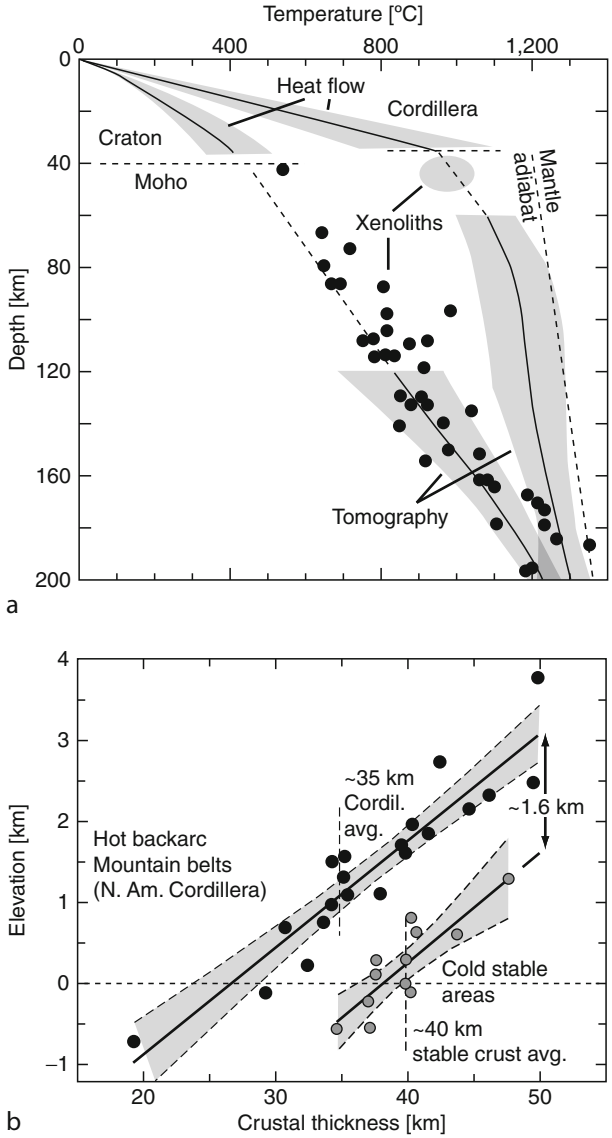
Compositional elevation adjustments for the 36 tectonic provinces of North America range from 763 ± 221 m in the Middle Rocky Mountains to 2,207 ± 119 m in the Gulf of California (Hasterok and Chapman, 2007b). Nearly all of the provinces have compositional elevation adjustments between −1 and +1 km. The results suggest that the compositional contribution to the elevation of North America accounts for ~3 km of observed elevation variation. Patterns are evident in variations of elevation adjustment as a function of crustal thickness and density (Figure 3c). The magnitude and range of elevation adjustments correlate with the type of tectonic regime that characterizes each North American province. Precambrian cratons have compositional adjustments between −0.5 and +0.5 km. Cenozoic rifts have elevation adjustments >0.5 km, and adjustments for volcanic provinces range from just under 0–1.0 km. Collisional orogens have a very large range of elevation adjustments relative to the other tectonic environments, ranging from 0.75 to 1.0 km.

The elevation adjustment is added to the observed elevation to determine the compositionally adjusted elevation (Figure 3c). Although the data shown in Figure 3c exhibit much scatter, two trends are clear. First, the adjusted elevation difference between hot and cold provinces is ~3 km, a magnitude roughly equal to the elevation range observed in the oceans between hot spreading centers and cold abyssal plains. Second, whereas observed elevations do not define a clear trend with heat flow, the adjusted province elevations are less scattered and define a pattern that more closely resembles our predicted thermal isostatic curve (black line). The considerable scatter in Figure 3c may be caused by imperfect compositional adjustment (seismic velocity characterization, velocity to density transform), by regions not being in isostatic equilibrium, or by surface heat flow not being perfectly representative of the thermal state of the lithosphere as described by the geotherm model used.

The North American Cordillera

The elevation of the North American Cordillera provides an excellent case study for thermal isostasy. Conventional wisdom suggests that many continental mountain belts, the result of continent–continent collision, are elevated because the collision process has thickened the continental crust. Airy isostasy considerations predict terrain elevations of 1,000–2,000 m as a result of thickened crust. But Hyndman (2010) point out that modern seismic data reveal that many of these collision provinces, in particular the North American Cordillera, do not have crustal roots. The Canadian Cordillera, for example, has a crustal thickness of only 35 km in comparison to the Canadian Shield crustal thickness of 40 km. But the elevation of the thinner crust Cordillera is more than 1,000 m higher than the Canadian Shield, exactly opposite to the Airy compositional isostasy prediction.

The solution to this apparent conundrum is found in thermal isostasy (Hyndman, 2010). Hyndman et al. (2005) and Currie and Hyndman (2006) have shown that high upper mantle temperatures characterize most subduction back-arc regions (see also Lewis et al., 1992; Lewis et al., 2003). And that many mountain belts also occur in current or recent back-arc locations. The temperature-depth field for the Cordillera, therefore, is much hotter than for a cratonic region (Figure 4a). The Cordillera geotherm approaches a mantle adiabat at a depth of about 60 km, defining the lithospheric thickness for the Cordillera, while the Craton geotherm does not approach the mantle adiabat until a depth of more than 200 km. If Equation 1 is applied to these two regions, the estimated difference in elevation, once compositional effects have been removed should be about 1,600 m. Figure 4b (from Hyndman, 2010) shows exactly that effect. Elevations for both back-arc mountain belts and cold stable areas



Isostasy, Thermal, Figure 4 Thermal isostasy case study for North America Cordillera (after Hyndman, 2010). (**a**) Temperature-depth plot for the northern Cordillera in comparison to the adjacent craton. Shaded regions in the mantle are general results from tomography; solid dots are temperature-depth loci from xenolith studies (see Hyndman, 2010 for details). (**b**) Elevation adjusted for composition (density) versus crustal thickness (Table 3 of Hasterok and Chapman, 2007b) for the hot Cordillera and cold stable areas of North America. The ~1.6 km elevation difference can be explained by thermal isostasy using Equation 1.

corrected for compositional density effects are plotted against local crustal thickness. Each group of data exhibit the expected trend of increasing elevation with increasing crustal thickness, but the Cordillera elevations are consistently 1,600 m higher than stable area elevations for the same crustal thickness.

Summary

Thermal isostasy, that part of isostasy traceable to density differences being caused by temperature differences and thermal expansion of rock, is an important but underappreciated condition in crustal geophysics. The condition/process of thermal isostasy received much attention in oceanic regions with the discovery of plate tectonics and seafloor spreading. It was quickly recognized that systematic cooling of the lithosphere formed at a spreading ridge could explain the 3,000 m of seafloor subsidence away from a ridge. On continents, thermal isostasy effects are obscured by compositional isostasy, principally crustal thickness, which produces over 3,000 m of elevation effects. Recent analyses, however, have developed methods to adjust elevations for compositional effects. When those adjustments are made, there is a clear trend of increasing elevation of a continental province with surface heat flow – a confirmation of thermal isostasy. The magnitude of elevation differences between hot and cold continental regions is also about 3,000 m, comparable to thermal isostasy effects in marine areas. An excellent example of thermal isostasy effects is drawn from the North American Cordillera where Airy isostasy is unable to explain the elevation of the Cordillera terrains relative to stable shield areas. Instead, the hot upper mantle associated with current or recent back-arc activity, conveniently explains the 1,600 m of elevation difference that is observed.

Bibliography

Airy, G., 1855. On the computation of the effect of the attraction of mountain-masses, as disturbing the apparent astronomical latitude of stations in geodetic surveys. *Philosophical Transactions of the Royal Society of London*, **145**, 101–103.

Blackwell, D., 1971. The thermal structure of the continental crust. In Heacock, J. (ed.), *The Structure and Physical Properties of the Earth's Crust. Geophysical Monograph Series*. Washington: AGU, Vol. 14, pp. 169–184.

Brott, C., Blackwell, D., and Ziagos, J., 1981. Thermal and tectonic implications of heat flow of the western Snake River Plain, Idaho. *Journal of Geophysical Research*, **86**, 11709–11734.

Chapman, D., 1986. Thermal gradients in the continental crust. In Dawson, J., Carswell, D., Hall, J., and Wedepohl, K. (eds.), *The Nature of the Lower Continental Crust*. Geological Society Special Publication 24, pp. 63–70.

Chapman, D., and Furlong, K., 1992. Thermal state of the continental lower crust. In Fountain, D., Arculus, R., and Kay, R. (eds.), *Continental Lower Crust. Geotectonics*. New York: Elsevier, Vol. 23, pp. 179–199.

Chapman, D., and Pollack, H., 1977. Regional geotherms and lithospheric thickness. *Geology*, **5**, 265–268.

Currie, R., and Hyndman, R., 2006. The thermal structure of subduction backarcs. *Journal of Geophysical Research*, **111**, B08404, doi:10.1029/2005JB004024.

Griffin, W., O'Reilly, S., and Ryan, C., 1999. The composition and origin of sub-continental lithospheric mantle. In Fei, Y., Bertka, C., and Mysen, B. (eds.), *Mantle Petrology: Field Observations and High Pressure Experimentation*: A Tribute to Francis R. (Joe) Boyd, Special Publication of the Geochemical Society, 6, pp. 13–45.

Han, U., and Chapman, D., 1995. Thermal isostasy: Elevation changes of geologic provinces. *Journal of the Geological Society of Korea*, **31**, 106–115.
Hasterok, D., and Chapman, D., 2007a. Continental thermal isostasy: 1. Methods and sensitivity. *Journal of Geophysical Research*, **112**, B06414, doi:10.1029/2006JB004663.
Hasterok, D., and Chapman, D., 2007b. Continental thermal isostasy: 2. Application to North America. *Journal of Geophysical Research*, **112**, B06415, doi:10.1029/2006JB004664.
Hyndman, R., Currie, C., and Mazzotti, S., 2005. Subduction zone backarcs, mobile belts, and orogenic heat. *Geological Society of America Today*, **15**, 4–10.
Hyndman, R., 2010. The consequences of Canadian Cordillera thermal regime in recent tectonics and elevation: a review. *Canadian Journal of Earth Sciences*, **47**, 621–632.
Jarvis, G., and McKenzie, D., 1980. Sedimentary basin formation with finite extension rates. *Earth and Planetary Science Letters*, **48**, 42–52.
Lachenbruch, A., and Morgan, P., 1990. Continental extension, magmatism, and elevation; formal relations and rules of thumb. *Tectonophysics*, **174**, 39–62.
Lachenbruch, A., and Sass, J., 1977. Heat flow in the United States and the thermal regime of the crust. In Heacock, J. (ed.), *The Earth's Crust: Its Nature and Physical Properties, Geophysical Monograph Series*. Washington: AUG, Vol. 20, pp. 625–675.
Lewis, T., Bentkowski, W., and Hyndman, R., 1992. Crustal temperatures near the Lithoprobe Southern Cordillera Canada transect. *Canadian Journal of Earth Science*, **29**, 1197–1214.
Lewis, T., Hyndman, R., and Flueck, P., 2003. Heat flow, heat generation, and crustal temperatures in the northern Canadian Cordillera: thermal control of tectonics. *Journal of Geophysical Research*, **108**, 2316–2334, doi:10.1029/2002JB002090.
McKenzie, D., 1978. Some remarks on the development of sedimentary basins. *Earth and Planetary Science Letters*, **40**, 25–32.
Nagihara, S., Lister, C., and Sclater, J., 1996. Reheating of old oceanic lithosphere: deductions from observations. *Earth and Planetary Science Letters*, **139**, 91–104.
Pollack, H., and Chapman, D., 1977a. Mantle heat flow. *Earth and Planetary Science Letters*, **34**, 174–184.
Pollack, H., and Chapman, D., 1977b. On the regional variation of heat flow, geotherms, and lithospheric thickness. *Tectonophysics*, **38**, 279–296.
Pratt, J., 1855. On the attraction of the Himalaya mountains, and the elevated regions beyond them, upon the plumb-line in India. *Philosophical Transactions of the Royal Society of London*, **145**, 53–100.
Rao, R., Rao, G., and Reddy, G., 1982. Age dependence of continental heat flow – fantasy and facts. *Earth and Planetary Science Letters*, **59**, 288–302.
Stein, C., and Stein, S., 1993. Constraints on Pacific midplate swells from global depth–age and heat flow– age models. In Pringle, M. (ed.), *The Mesozoic Pacific: Geology, Tectonics, and Volcanism, Geophysical Monograph Series*. Washington: AGU, Vol. 77, pp. 53–76.
Watts, A., 2001. *Isostasy and the Flexure of the Lithosphere*. Cambridge: Cambridge University Press.

Cross-references

Heat Flow, Continental
Lithosphere, Continental: Thermal Structure
Lithosphere, Oceanic: Thermal Structure

L

LEGAL CONTINENTAL SHELF

Ray Wood[1], Stuart A. Henrys[1], Vaughan Stagpoole[1], Bryan Davy[1], Ian Wright[2]
[1]GNS Science, Lower Hutt, New Zealand
[2]National Oceanography Centre, Southampton, UK

Synonyms

Continental shelf; Extended continental margin; Extended continental shelf

Definition

The legal continental shelf comprises the submerged prolongation of the land mass of a coastal State beyond 200 nautical miles from the territorial sea baselines, and consists of the seabed and subsoil of the shelf, the slope, and the rise. Formulae to determine the outer limits of the legal continental shelf are in article 76 of the United Nations Convention on the Law of the Sea. These formulae are based on seafloor morphology and sediment thickness. The outer limits of the legal continental shelf cannot exceed 350 nautical miles from the territorial sea baselines or 100 nautical miles beyond the 2,500 m isobath.

Introduction

The United Nations Convention on the Law of the Sea (UNCLOS) was adopted in 1982 after 9 years of discussion, bargaining, and compromise (Division for Ocean Affairs and the Law of the Sea DOALOS, 2007). Significant features of UNCLOS relevant to the management of natural resources include establishment of the 200 nautical mile (M) Exclusive Economic Zone (EEZ) and of the legal continental shelf.

Within the 200 M EEZ a State has rights to resources in the water column, such as deep-sea fisheries, and resources on and beneath the seafloor, such as oil, gas, and other minerals. Within the legal continental shelf, a State has rights to resources on and beneath the seafloor, but not to resources in the water column.

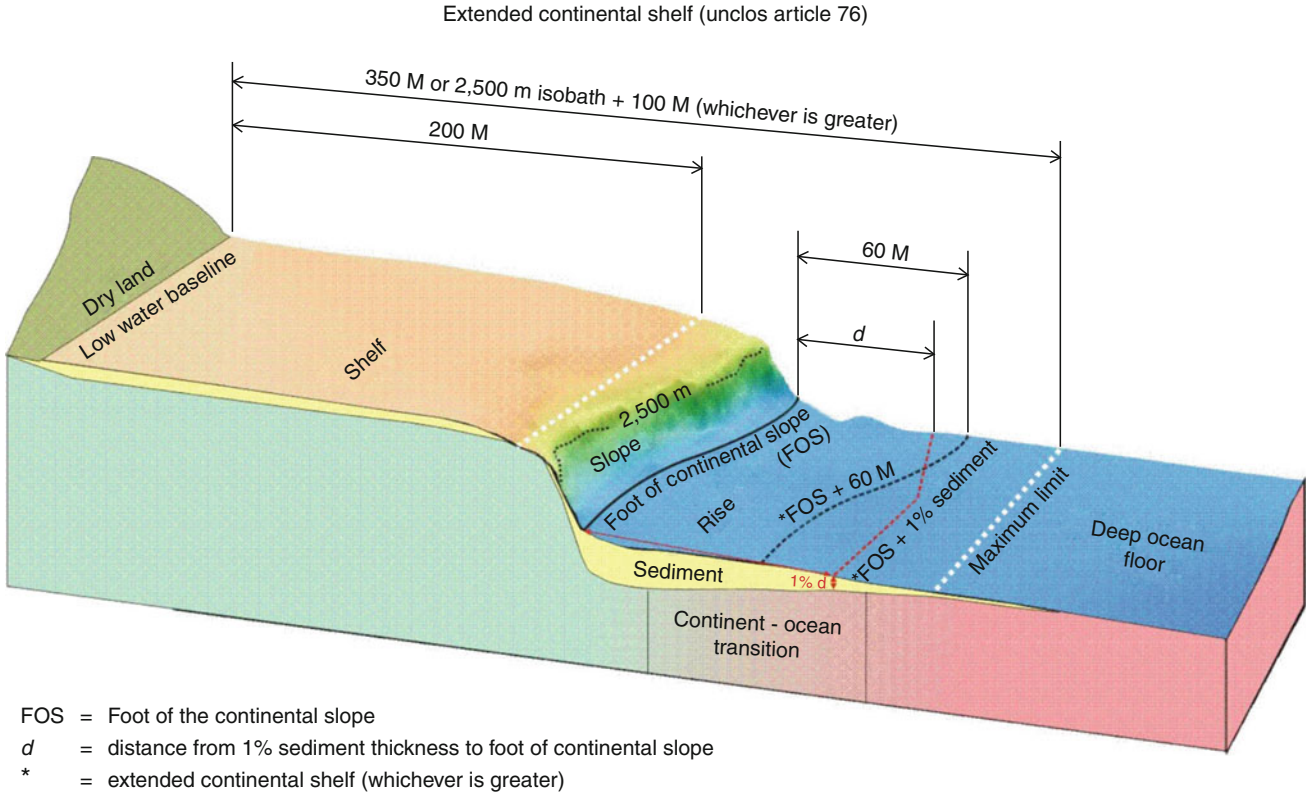
Areas of legal continental shelf in submissions to the United Nations cover large parts of the oceans and include seafloor with a wide variety of morphologic and geologic characteristics. The attribute these areas have in common is an interpreted link to the land mass of a coastal State. Techniques to establish the outer limits of the legal continental shelf are described in article 76 of UNCLOS and Annex II of the Final Act of the Third United Nations Conference on the Law of the Sea (the Statement of Understanding).

Establishing the outer limits of the legal continental shelf

Delimitation of the 200 M EEZ is a geodetic calculation using basepoints and/or baselines established along the coastline of the State, subject to negotiations with other States where EEZs overlap. Article 76 of UNCLOS states that the continental shelf comprises the submarine prolongation of the State's land mass and consists of the shelf, slope, and rise but it does not include the deep ocean floor (Figure 1). The article sets out a process whereby States can define the outer limits of a legal continental shelf beyond 200 M. The Statement of Understanding extends the principles of article 76 to coastal States in the southern part of the Bay of Bengal that have an extraordinarily extensive rise.

Application of the concept of prolongation to the complex seafloor and geological relationships of continental margins can be challenging. Establishing whether prolongation occurs and the nature of that prolongation are of primary importance because they determine (a) whether a legal continental shelf exists and (b) how far it extends beyond 200 M. Article 76 refers to a variety of submarine features, including plateaux, rises, caps, banks, spurs,

Harsh K. Gupta (ed.), *Encyclopedia of Solid Earth Geophysics*, DOI 10.1007/978-90-481-8702-7,



Legal Continental Shelf, Figure 1 Cartoon showing a hypothetical continental margin, summarizing the formulae and constraints described in article 76 that determine the outer limits of the continental shelf (Modified from Kapoor and Kerr, 1986).

submarine ridges, and oceanic ridges. Submarine ridges, for example, are subject to the 350 M constraint (Symonds and Brekke, 2004), whereas other features are subject to the 2,500 m + 100 M constraint. On margins with expansive, relatively shallow slopes the 2,500 m + 100 M constraint could lie far seaward of the 350 M line.

Prolongation of land mass

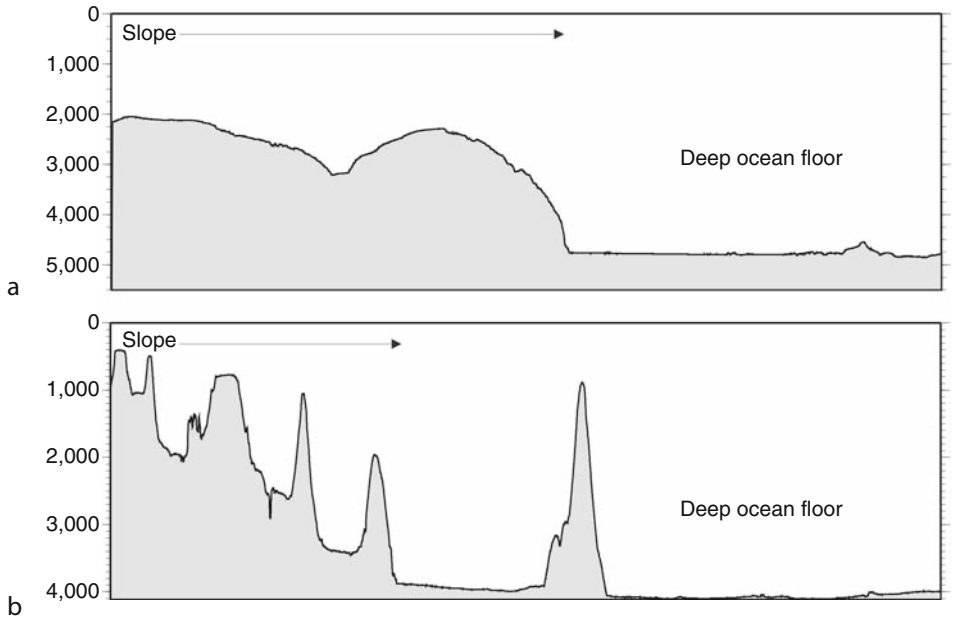
There are three aspects of prolongation of a land mass:

- Morphologic – the seafloor shape is a continuation of the land mass morphology
- Geologic – the rocks beneath the seafloor are the same as, or related to, those of the land mass
- Tectonic – the rocks beneath the seafloor share their history with those of the land mass

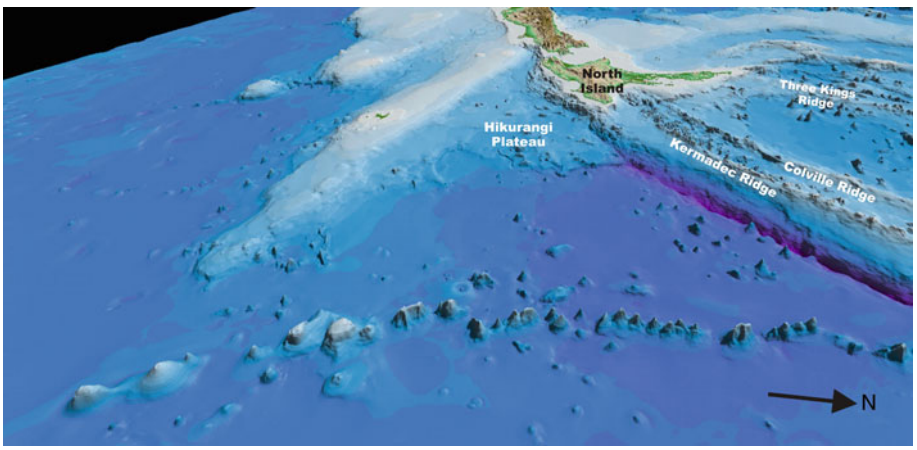
The morphological transition from land mass to deep ocean floor is the result of the composition and density of the rocks beneath the seafloor, the geologic processes that form and shape them, and the tectonic forces that act on them. This transition is relatively simple on many margins, with a shelf dipping gently to a depth of several hundred meters, a steeper slope deepening to a depth of several thousand meters, and on some margins a gently dipping rise lying between the slope and the almost flat deep ocean floor (Figure 2a). Other margins have a complex morphology with ridges, seamounts, and canyons and the transition with the deep ocean floor can be harder to recognize (Figure 2b).

Prolongation of a land mass cannot be established solely on the basis of rock type. Land masses are usually the result of igneous, metamorphic, sedimentary, and tectonic processes that form geologically complex crust that is generally distinctly different from the crust of the deep ocean floor. The deep ocean floor is the result of sea floor spreading and is characterized by basaltic rocks. Rocks such as granite, marble, and sandstone only form as part of continent building processes and can be used to infer that an area is continental in origin, but land masses are not always composed of continental-type rocks. Iceland and other volcanic islands are examples of land masses that are composed of basaltic rocks, and oceanic rocks can be thrust onto or accreted to a land mass or continent by tectonic movements.

Lines of volcanoes are common above active and fossil subduction zones, particularly around the Pacific. Volcanic arcs often extend into large continental blocks and the rocks associated with them form the core of many continents. The Three Kings Ridge, Colville Ridge, and Kermadec Ridge are three volcanic arcs that extend thousands of kilometers north of the North Island of New Zealand (Figure 3). These ridges are part of the



Legal Continental Shelf, Figure 2 Bathymetry profiles across a continental margin showing a relatively simple (**a**) and a more complex morphology of the continental slope (**b**).



Legal Continental Shelf, Figure 3 Oblique view of the seafloor showing the continuity of volcanic arcs (Three Kings Ridge, Colville Ridge, and Kermadec Ridge) and the Hikurangi Plateau with the land mass of New Zealand.

morphological, geological, and tectonic prolongations of the land mass of New Zealand (e.g., Mortimer et al., 2007).

Tectonic prolongation implies that the rocks beneath the seafloor have a shared history with the rocks of the land mass. Shared history refers to the processes of continent building: sediment deposition, igneous activity, and suturing or accretion of rocks along plate margins. Disruption of this history results from fragmentation of the land mass by the tectonic processes of rifting and seafloor spreading. Continents and land masses are part of the plate tectonic conveyor system, and their extent and composition changes with time as the result of interactions along the plate margins.

The geologic, morphologic, and tectonic boundaries are seldom, if ever, abrupt, and implicit in the determination of prolongation is an assessment of the degree of the continuity or uniformity of the geology, morphology, and tectonic history.

The Hikurangi Plateau is an example of a basaltic tectonic block that is a part of the prolongation of the New Zealand land mass (Figure 3). The plateau was accreted to the New Zealand continent when it collided with a subduction zone along the Gondwana margin more than

100 million years ago (Davy, 1993). The plateau stands more than 1,000 m above the deep ocean floor and is morphologically, geologically, and tectonically connected to the New Zealand land mass.

Application of UNCLOS article 76

The process for defining the outer limits of a legal continental shelf beyond 200 M has four steps (Figure 4); (1) the area of prolongation of the land mass is determined using two formulae, (2) constraint lines are determined using two methods, (3) the appropriate constraints are applied to the area of prolongation, and (4) the resulting line composed of formulae and/or constraint lines is smoothed with 60 M bridging lines.

Both formulae use the location of the foot of the continental slope as a reference point. The Commission on the Limits of the Continental Shelf (CLSC) (1999) declares a preference for identifying the foot of the continental slope as the point of maximum change in seafloor gradient as the general rule, and use of "evidence to the contrary" as exceptions to the rule. The guidelines state that these exceptions can be applied when the foot of the continental slope is not located reliably by the point of maximum change in sea floor gradient, and that they provide an opportunity for States to use geological and geophysical evidence to locate the foot of the continental slope.

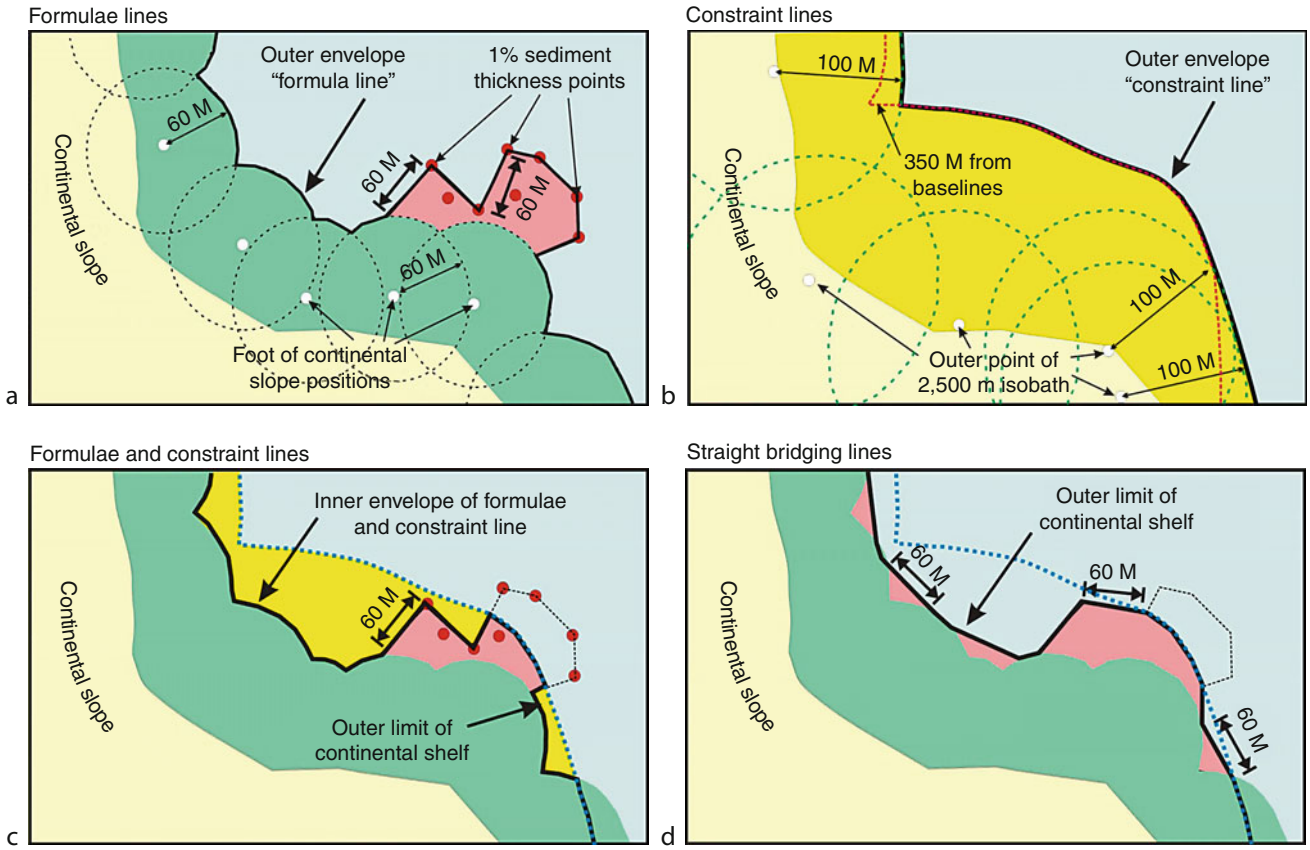
Locating the foot of the continental slope

There are two steps to find the foot of the continental slope:

1. Identify the region of the base of the continental slope.
2. Determine the location of the foot of the continental slope within that region.

Identification of the region of the base of the continental slope may be on the basis of morphological evidence and/or geological and geophysical evidence.

In areas where the morphology of the continental margin can be clearly subdivided into shelf, slope, rise, and deep ocean floor (Figure 1), the region of the base of the continental slope is where the lower continental slope meets the rise, or where it meets the deep ocean floor in



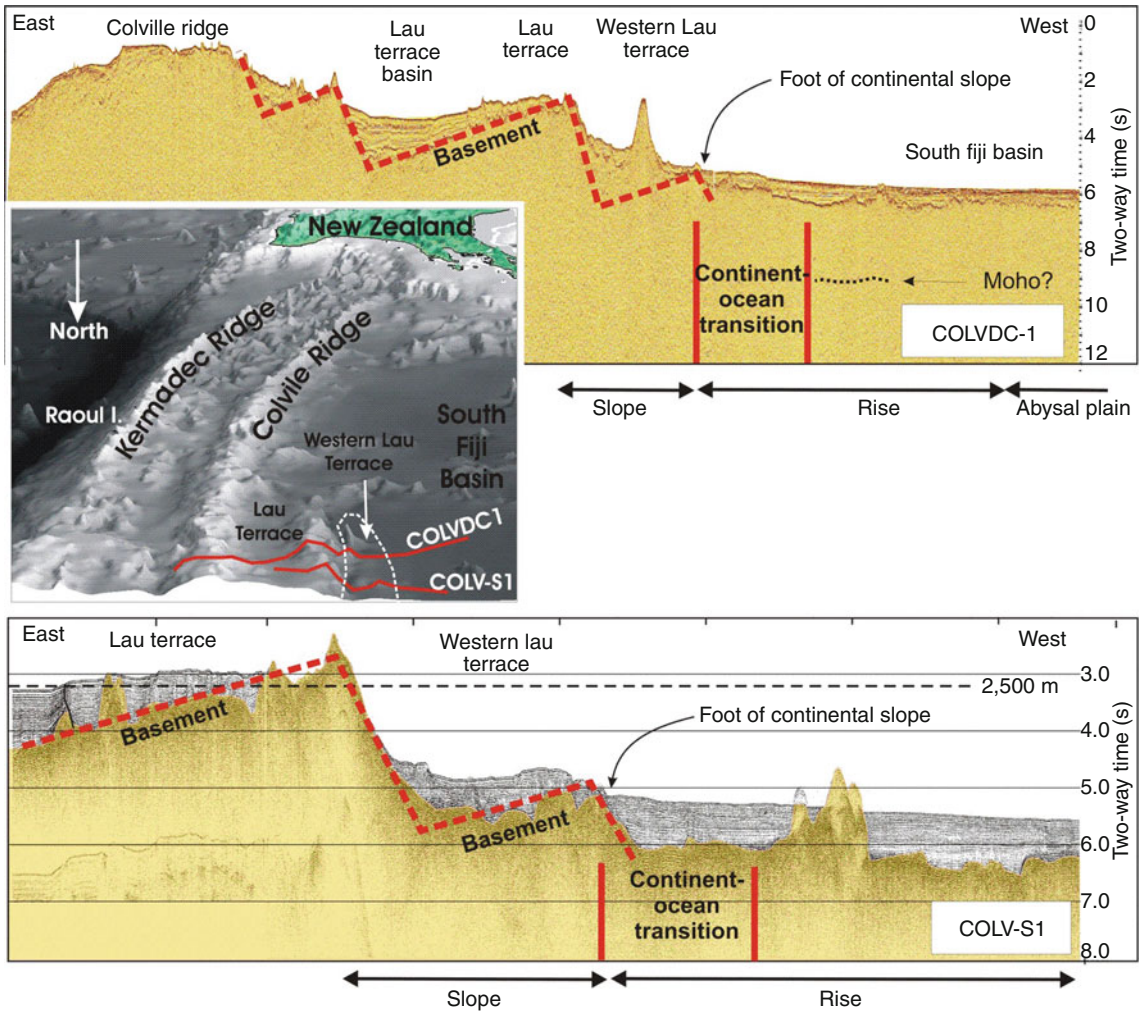
Legal Continental Shelf, Figure 4 The four steps for establishing the outer limits of the legal continental shelf: (**a**) Determine foot of slope positions and apply formulae in paragraph 4 of article 76; (**b**) Establish constraint line by application of paragraphs 5 and 6 of article 76; (**c**) Intersect formula and constraint lines to establish fixed points along outer limits of the continental shelf; (**d**) Delineate the outer limits of the continental shelf with straight bridging lines no more than 60 m long between fixed points, as specified in paragraph 7 of article 76.

cases where a rise is absent. Regional gradients can be used to narrow the search for the region of the base of the continental slope. Regional gradients less than 1° are generally considered representative of the rise and deep ocean floor, and regional gradients greater than 2° are generally considered representative of the slope. There are many exceptions to these values, however, and other evidence, such as the erosion effects of deep-sea currents, underwater slides, margin collapse, local volcanic activity, and deep-sea canyons must be considered before finally establishing the region of the base of the continental slope. The width of the region of the base of the continental slope is typically 4–10 km, but can vary according to the complexity of the margin.

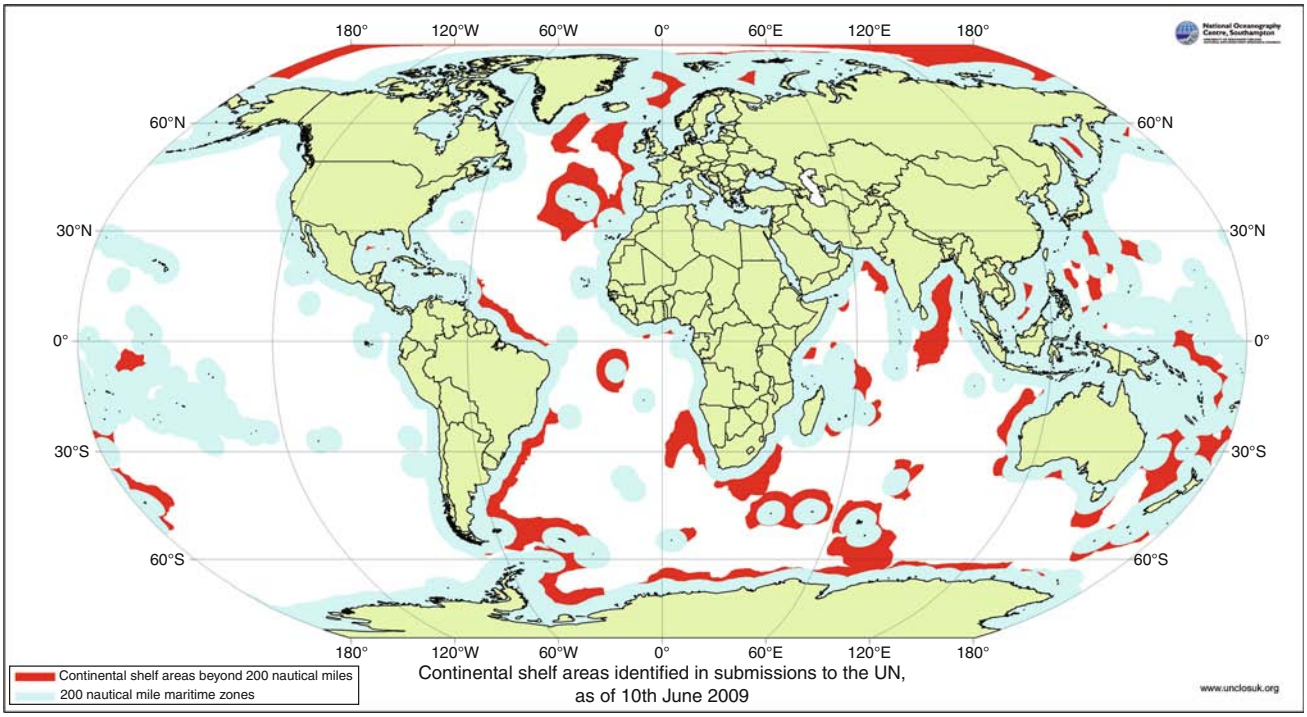
The point of maximum change in gradient is determined by computing the second derivative of the bathymetry within this region. The foot of the continental slope is the point with the maximum value of the second derivative of the bathymetry.

Situations in which the maximum change in seafloor gradient rule might not reliably locate the foot of the continental slope include continental margins that have a very smooth transition from slope to rise and no single point represents a significant maximum change in gradient, and continental margins with irregular seafloor morphology and a point of maximum change in gradient that is not consistent with other data that indicate the location of the foot of the continental slope.

The region of the Western Lau Terrace, north of New Zealand (Figure 5) provides an example of the use of geological and geophysical evidence to identify the region of the base of the continental slope. This part of the New Zealand continental margin is characterized by terraces and sediment-filled basins formed by tilted basement blocks that step down to the deep ocean floor of the South Fiji Basin. The continent ocean transition lies at the western edge of the Western Lau Terrace and the foot of the continental slope is identified as the point with



Legal Continental Shelf, Figure 5 Seismic sections and a location map (looking south) of the Lau Terrace and Western Lau Terrace. Basement fault blocks are shown by *red dashed lines*.



Legal Continental Shelf, Figure 6 Red areas indicate areas of legal continental shelf submitted to the United Nations. Please also refer to www.unclosuk.org.

the maximum change in seafloor gradient within the transition zone.

Once the foot of the continental shelf points are established, then the outer limits of the continental margin are constructed either 60 M from these points, or at fixed points where the sediment thickness is at least 1% of the distance to these points, whichever is greater (Figure 4b).

Establishing constraint lines

The constraint line is constructed from the line 350 M from the basepoints from which the breadth of the territorial sea is measured and the line 100 M from the 2,500 m water depth contour, whichever is greater. Construction of the constraint line requires consideration of whether the features are submarine ridges or are natural components of the continental margin. The continental shelf cannot extend beyond the constraint line, so the intersection of the constraint line with the area inside the formula lines defines fixed points along the outer edge of the continental shelf (Figure 4c).

Straight bridging lines

The final step in the construction of the outer limits of the continental shelf is to join the fixed points with straight bridging lines no more than 60 M long (Figure 4d). The outer limits of the continental shelf therefore can be comprised of points that lie on arcs 60 M from foot of the slope positions, points where the sediment thickness is at least 1% of the distance to foot of the slope positions, points 350 M from the territorial sea basepoints, points 100 M from the 2,500 m isobath, and points on bridging lines between fixed points.

Summary

Ratification of UNCLOS by coastal States has led to a renewed interest in the morphology and geology of continental margins. These States have applied the terms of article 76 and the Statement of Understanding to their continental margins and identified large areas of legal continental shelf (Figure 6). At the time of writing most of these areas are awaiting recommendations by the United Nations Commission on the Limits of the Continental Shelf.

Bibliography

Commission on the Limits of the Continental Shelf (CLSC) 1999. Scientific and technical guidelines of the Commission on the Limits of the Continental Shelf. CLCS/11 (http://www.un.org/Depts/los/clcs_new/clcs_home.htm).

Davy, B., 1993. The influence of subducting plate buoyancy on subduction of the Hikurangi-Chatham Plateau beneath the North Island, New Zealand. In Watkins J., Zhigiang, F., and McMillen K. (eds.), *Advances in the Geology and Geophysics of the Continental Margin*. AAPG Memoir 53, 75–91.

Division for Ocean Affairs and the Law of the Sea (DOALOS) 2007. The United Nations Convention on the Law of the Sea (a historical perspective). http://www.un.org/Depts/los/convention_agreements/convention_overview_convention.htm

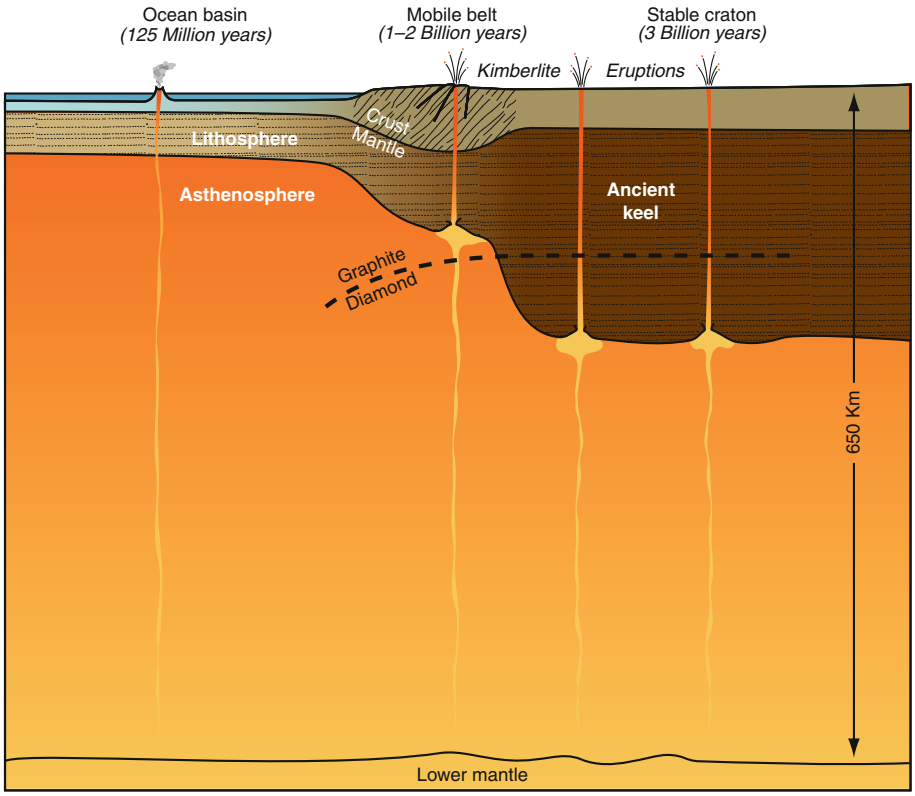
Kapoor, D. C., and Kerr, A. J., 1986. *A Guide to Maritime Boundary Delimitation*. Toronto: Carswell. 123 p.
Mortimer, N., Herzer, R. H., Gans, P. B., Laporte-Magoni, C., Calvert, A. T., and Bosch, D., 2007. Oligocene-Miocene tectonic evolution of the South Fiji Basin and Northland Plateau, SW Pacific Ocean: evidence from petrology and dating of dredged rocks. *Marine Geology*, **237**(1/2), 1–24.
Symonds, P. A., and Brekke, H., 2004. The ridge provisions of article 76 of UN convention on the law of the sea. In Nordquist, M. H., Moore, J. N., and Heidar, T. H. (eds.), *Legal and Scientific Aspects of Continental Shelf Limits*. Laiden: Martinus Nijhoff, pp. 169–199.

LITHOSPHERE, CONTINENTAL

David E. James
Department of Terrestrial Magnetism, Carnegie Institution of Washington, Washington, DC, USA

Definition and introduction

The continental lithosphere consists of the continental crust and, typically, some nonconvecting part of the underlying upper mantle (Figure 1). In plate tectonics terms, the continental lithosphere is part of the rigid outer rind of the Earth, which is segmented into several major plates. The cold lithosphere lies atop a hotter, more mobile (low strength) asthenosphere. In this context, oceanic lithosphere for the most part obeys comparatively simple thermal models, with the boundary between lithosphere and asthenosphere represented by an isotherm marking the transition in mantle peridotite between elastic and ductile behavior (but see *Lithosphere, Oceanic* for the growing list of complications). By contrast, the continental lithosphere is heterogeneous and its structure highly variable. That contrast reflects the fact that oceanic lithosphere is formed in much the same way worldwide and that it is continually recycled into the interior of the Earth via the plate tectonic "conveyer belt." Continental lithosphere, on the other hand, is generally too buoyant to be subducted, although an increasing number of instances have been reported in which parts of continental lithospheric mantle show evidence of gravitational instability, forming "drips" where lithosphere delaminates and sinks into the deeper mantle (Zandt et al., 2004). The detachment of continental mantle from the crust in such circumstances may be an important factor in intraplate



Lithosphere, Continental, Figure 1 Schematic cross section showing transition from oceanic lithosphere to younger continental lithosphere and into the thick, cratonic lithosphere (tectosphere) beneath the ancient continental nucleii. Melts rising from depth ascent through relatively thin oceanic and young continental lithosphere to produce volcanoes at the surface, whereas thick tectosphere acts as a barrier to melts percolating from depth, requiring evolved, volatile-rich kimberlitic magmas to bore a conduit through the mantle to the surface.

deformation and volcanism (Carlson et al., 2005). Thus, while no known oceanic lithosphere is older than about 200 my, the continental lithosphere has evolved over billions of years, one consequence of which is that it is seen today as the resulting collage of many superimposed tectonic events, including rifting, subduction, continental collision, accretion, and hot spot magmatism, that recur repeatedly over geologic time.

The study of continental lithosphere has been revolutionized over the past two decades with the advent of portable broadband seismology. Passive array experiments (see *Seismological Networks*) have made it possible to obtain high-resolution images of deep lithospheric structures beneath every continent (James, 2007). The greatest concentration of high-density broadband deployments has been in western North America, Europe, and Asia, with large-scale experiments also in Africa, Central and South America, and Australia. Within the USA, the Transportable Array of the national EarthScope project has already resulted (at the time of this writing) in uniformly high-resolution seismic imaging of the mantle beneath the western half of the nation. The lithosphere so revealed consists of a collage of overlapping structures produced by ongoing subduction, trench migration, ridge override, regional extension, mantle upwellings and volcanism, and lithospheric delamination. In some tectonically active regions, a thin crust is the only lithosphere remaining intact (Carlson et al., 2005). This complex and disrupted lithosphere is in marked contrast to that of stable cratonic regions, including the continental nuclei of north-central North America, western Australia, and southern Africa. The cratons of southern Africa are perhaps the best studied of the Archean cratons of the world, where results from large-scale passive array experiments show that lithospheric roots as deep as 250 km or more underlie the undisturbed parts of the cratons (James and Fouch, 2002; Ritsema and van Heijst, 2000).

Despite remarkable progress in imaging the continental lithosphere over the past 2 decades, a precise definition of continental lithosphere remains elusive. No single definition of continental lithosphere can be made to fit all circumstances. Particularly controversial is the definition of what constitutes the base of the lithosphere. Many definitions of continental lithosphere appear in the literature – seismological, mechanical, rheological, thermal, and compositional. Depending on one's view, the continental lithosphere may be a *mechanical boundary layer,* a *thermal boundary layer*, or a *chemical boundary layer.* In the remainder of this entry, we shall explore these various concepts in some detail. Our goal here is to describe the continental lithosphere in terms of global Earth structure, plate tectonics, and the long-term evolution of the continents. By this view, the continental lithosphere is envisaged to be a long-lived plate tectonic unit, consisting both of continental crust and, typically, a very significant mantle "keel" that is attached to and translates with the continent. In this sense, it is seen as a stable component of the continent and does not participate in the convective processes of the deeper dynamic mantle.

Mechanical models

A simple expression of the lithosphere is *elastic thickness*. While rarely used to describe continental lithospheric in a geologic or structural sense, elastic thickness is a useful concept in that it models the mechanical response of the lithosphere as an elastic plate overlying a weak (asthenospheric) substrate. As discussed in detail elsewhere in this volume (see *Lithosphere, Mechanical Properties*), the elastic thickness of the lithosphere may be determined through topographic response to surface loads (such as volcanoes), the relationship between topography and Bouguer gravity anomalies, or the shape of depositional surfaces in basins. The thickness of elastic lithosphere so determined is rarely as much as 100 km and in areas of extension or hotspot activity may be less even than crustal thickness. The elastic thickness of the continents is chiefly controlled by the thermal state of the lithosphere (see *Lithosphere, Mechanical Properties*). As Forsyth (Forsyth, 1989) points out, however, the apparent elastic thickness is also affected by the state of stress and the rheological stratification in the crust and mantle, so mechanical thickness may not bear a simple relationship to a particular isotherm.

One important aspect of the mechanical structure of the continental lithosphere as it relates to lithospheric evolution is the large decrease in strength of crustal rocks in the deep crust. The steep geothermal gradient in the crust produces increasing ductility of crustal rocks with depth. The brittle zone in tectonic areas typically extends only to about 15 km, below which the rocks deform plastically. The topmost part of the underlying mantle, however, is relatively much stronger as temperatures there are still far from the melting point of mantle rocks. Thus, it is widely posited that there is a zone of weak lower crust sandwiched between strong upper crust and strong upper mantle, creating a possible zone of crust–mantle decoupling (see *Lithosphere, Mechanical Properties*).

While the flexural lithosphere is a useful *mechanical* concept, it fails as a plate tectonic description of lithosphere. In many tectonic regions, the flexural lithosphere is not even as thick as the continental crust. As low-density sialic crust is not easily subducted into the mantle, it must persist as the primary long-lived component of continental lithosphere. From the perspective of plate tectonics, it may be argued that the full definition of the lithosphere should account both for the long-term structural coherence of continental plates and the impact of asthenospheric processes that affect either the formation or the removal of lithospheric mantle beneath the crust.

Thermal models

A convenient definition for *thermal lithosphere*, or the thermal boundary layer, is the outer layer of the Earth in

which heat transfer is dominated by conduction (see *Heat Flow, Continental*; *Lithosphere, Continental: Thermal Structure*). Thus, the stronger lithosphere acts as a barrier to thermally induced buoyancy forces that drive convective heat transfer in the underlying asthenosphere (Morgan, 1984). In the case of the oceanic lithosphere, this conductively cooled layer of the Earth is thought to have finite thickness (i.e., the lithosphere ceases to undergo thermal contraction as it ages beyond about 80 Ma, implying that heat is supplied to the lithosphere from the underlying asthenosphere) (see *Lithosphere, Oceanic*; *Lithosphere, Oceanic: Thermal Structure*). For the continents, the situation is far more complex. The continental lithosphere does not undergo simple monotonic cooling, but may be subjected to repeated episodes of thermal or tectonic disturbances. Thus, superimposed upon the geological age of the lithosphere is the thermal or tectonic age of the lithosphere, factors that have led to significant compositional heterogeneity in both crust and lithospheric mantle.

We consider first the thickness of the thermal boundary layer beneath the stable cratonic cores of the continents. Here, the lithosphere is both at its thickest and its strongest. If the effective viscosity of the asthenosphere beneath the cratonic lithosphere is taken to be about 10^{21} Pa s, based upon glacial unloading and gravity/topography correlations (see *Mantle Viscosity*), it implies a temperature of transition from conductive to convective heat transfer of about 1300–1400°C for typical mantle compositions (Morgan, 1984). If crustal heat production is taken into account, the thickness of the thermal lithosphere of the stable continental cratons as calculated from surface heat flow may range from about 90 to 220 km (Rudnick et al., 1998).

The lithospheric thickness obtained by thermal modeling depends upon other factors, the most important of which is the composition, including volatile content, of the subcrustal mantle. If the continental lithospheric mantle consists of low-density peridotite depleted of its low-melting basaltic fraction and devolatilized over time due to metamorphic and magmatic events (see discussion below), the continental mantle will be less dense and substantially more refractory than the rest of the mantle. The chemical boundary layer represented by this depleted subcontinental mantle peridotite may stabilize the continental mantle root against thermal disruption to depths below the thermal boundary layer (Jordan, 1978, 1981).

Tectono-thermal events, such as those that are widespread across the western USA, will have the effect of thinning the existing lithosphere. Regions where prior thermal disturbances have occurred, therefore, may be more prone to reactivation, whereas stable cratons, particularly those buffered by bordering Proterozoic mobile belts as in southern Africa, should be comparatively less vulnerable to thermal perturbations from the underlying asthenosphere (Pollack, 1986). Tectonic reactivation seems to be particularly true of continental rifting, where repeated episodes apparently recur along the same long-lived zones of weakness. The greatly expanded scope of tomographic studies in recent years has revealed a number of localities in which various forms of lithospheric delaminations, or "drips" appear to be occurring, including a number in the western USA (e.g., [West et al., 2009; Zandt et al., 2004]). Such processes suggest that lithospheric mantle beneath tectonically active areas of the continent may be thinned or removed and later regenerated. Indeed, a wide range of epeirogenic (vertical) movements within continental interiors have been interpreted to be due to thermal disturbances that produce lithosphere heating and thinning, followed by conductive relaxation and asymptotic thickening of the lithospheric thermal boundary layer.

Seismological/compositional models – tectosphere

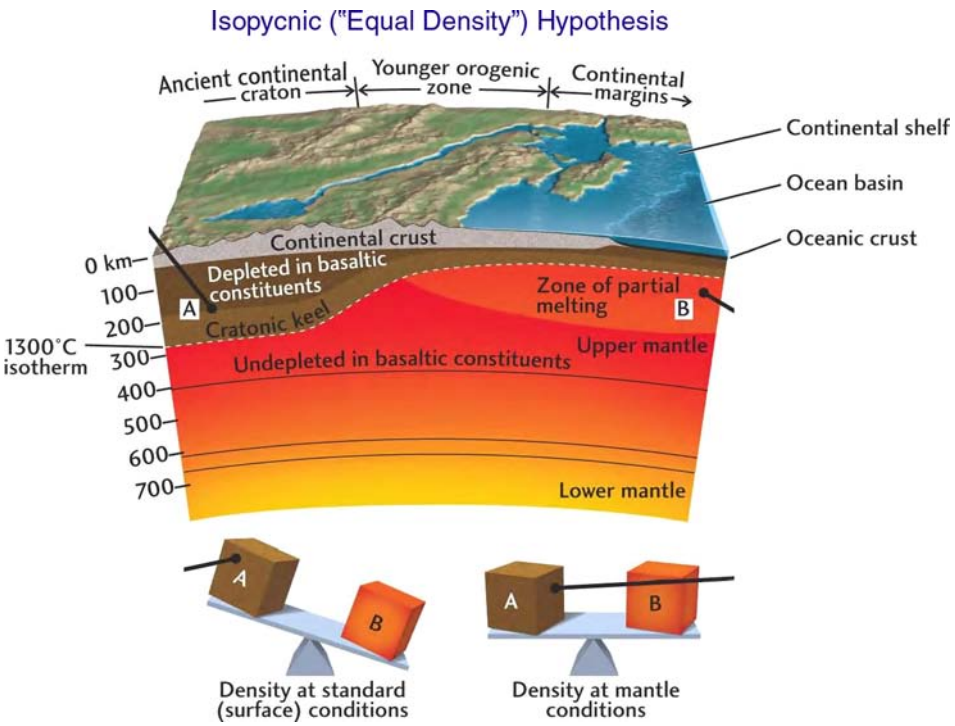
It is well known that average seismic velocities, especially shear velocity, are much higher under cratons than under oceanic or tectonically active areas. This was first shown by surface wave dispersion measurements and later by measurements of vertically travelling ScS phases. Sipkin and Jordan (Sipkin and Jordan, 1980) showed that ScS one-way travel-time anomalies associated with the upper mantle beneath stable continental interiors could be as large as 3.5 s (fast) relative to the upper mantle beneath the western Pacific. Moreover, the high-velocity continental paths are also associated with high Q_s, suggesting that the anomalies are due to thermal and/or compositional variations in the upper mantle (Jordan, 1981). Since the early studies of Sipkin and Jordan, numerous other studies based on three-dimensional inversion of travel-time anomalies for velocity structure have been carried out regionally across the globe (see summaries in [Carlson et al., 2005; James, 2007]). The results nearly all show that regional velocity contrasts, even between provinces entirely within the continents, extend to at least 200–250 km (James, 2007).

The notion of a long-lived thermochemical boundary layer in the upper mantle beneath vast regions of Precambrian crust raised many difficult questions as to how they formed and how they survived over geologic time and to a significant extent these questions remain only partially answered. There is general agreement, however, that the longevity and stability of ancient cratonic roots must be due, at least in part, to a combination of factors, primarily compositional buoyancy and strength and secondarily buffering of cratons by marginal Proterozoic mobile belts. The concept of a stabilizing chemical boundary layer is at the heart of the *temperature-depletion compensation* hypothesis laid out by Jordan in an important series of papers on the continental *tectosphere* (see [Jordan, 1981] for references) and discussed in some detail below. The term tectosphere was introduced by Jordan in an attempt to circumvent some of the difficulties associated with the conflicting definitions of lithosphere alluded to above. It is a particularly convenient concept for understanding those seismological and petrological/geochemical aspects

of deep continental structure that appear to indicate continent-ocean heterogeneity to depths of at least 200 km, and perhaps considerably more.

There is a large body of knowledge about the composition of deep cratonic mantle from the study of xenoliths that have been carried to the surface in explosive volcanic pipes. The development of accurate mineralogical geothermometers and geobarometers, which allow the equilibrium depth and temperature of xenolith source regions to be determined, made it possible to "map" xenolith compositions as functions of position in the mantle. A comprehensive summary of results from the geochemical study of mantle xenoliths can be found in (Carlson et al., 2005). Most of the ultramafic nodules associated with older cratonic areas are highly depleted in their basaltic melt fraction to a depth of at least 200 km and are approximately the same age as the overlying crust (Carlson et al., 2005; James et al., 2004). Depleted mantle xenoliths, which are by far the most common, are characterized by significantly lower density and higher seismic velocities than the more typical fertile suboceanic mantle (Boyd and McCallister, 1976). This fact led Jordan (Jordan, 1981) to postulate a thick *chemical boundary layer* for the continental tectosphere.

The effect of removing a basaltic partial melt from a fertile mantle peridotite (such as the hypothetical pyrolite of suboceanic mantle) is to reduce the density of the residue and to increase seismic velocities. Jordan compared densities of depleted peridotite nodules with densities of a hypothetical pyrolite composition and shows that average depleted peridotite of subcontinental mantle has a normative density about 1.3% less than that of pyrolite. This reduction in density will approximately compensate for the density contrast in a homogeneous mantle that would result from the temperature contrast of about −400° between a typical shield geotherm and the average oceanic adiabat. The model then, is based on the supposition that the lower density of the subcontinental mantle due to basalt depletion is gravitationally balanced by the density decrease in the oceanic mantle caused by a much higher geotherm, hence the expression *temperature-depletion compensation* (Figure 2). The thickness of the chemical boundary layer is generally taken to be about 150–200 km.



Lithosphere, Continental, Figure 2 Schematic diagram (adapted from Jordan, e.g., [Jordan 1988] with reference to previous work) showing the principal components of the isopycnic ("equal density") hypothesis. The cold, strong lithosphere (A) extends to at least 250 km beneath the ancient cratons and migrates with the craton during plate motions. At comparable depths beneath the oceans (B), the mantle is hot, weak, and mobile. The depleted mantle peridotite that makes in the cratonic keel is depleted in heavier constituents (e.g., Fe) and is therefore intrinsically less dense than more fertile oceanic mantle at the same temperature and pressure. At the same time, the sub-cratonic mantle is colder than the suboceanic mantle so the densities are approximately equal (the "temperature-depletion compensation" model). The chemical buoyancy of the cratonic root, along with its anhydrous strength, helps stabilize it against convective disruption over geologic time.

The tectosphere model remains controversial. The mechanisms by which a chemical boundary layer can be built up to form a deep mantle keel beneath ancient continental crust are poorly understood, although there are many modes by which magma can be extracted from continental mantle over time (Carlson et al., 2005). The most difficult aspect of deep mantle roots to model, however, has been their remarkable dynamical stability over billions of years of a chemical boundary layer whose thickness varies laterally by 150 km or more (Shapiro et al., 1999). Convective instability should act to disperse and thin the zone of depleted mantle, and the fact that this does not occur appears to be due largely to the anhydrous, highly viscous, nature of the depleted mantle keel.

One aspect of lithospheric formation that may be important in some cases involves underplating of continental lithosphere by oceanic lithosphere. There is evidence, for example, that among the eclogites brought up in kimberlite pipes in Africa some may be relicts of basaltic oceanic crust emplaced beneath the continental lithosphere in Precambrian time (MacGregor and Manton, 1986). The data on which this conclusion is based come primarily from measurements of oxygen isotopic ratios combined with trace element and radiogenic isotopic ratios. The oxygen isotopic ratios (low ∂O^{18}), in particular, are difficult to explain as originating anywhere other than in oceanic crust. Similar evidence for underplating of cratonic mantle by oceanic lithosphere has also been found in sulfide inclusions in eclogitic diamonds from southern Africa, where episodic diamond formation has been linked to the accretion of oceanic lithosphere onto early cratonic nuclei, with subsequent stabilization of the tectospheric mantle (Shirey et al., 2001).

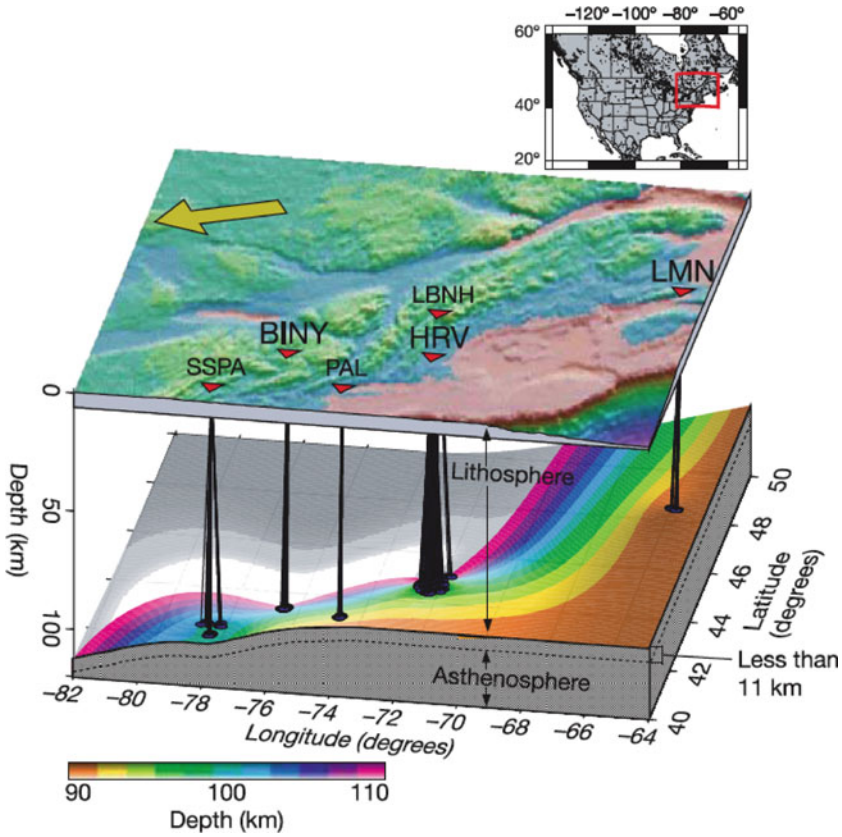
Lithosphere–asthenosphere boundary

A fundamental concept of plate tectonic theory is that the cold, rigid lithospheric plates of the Earth's outer rind are decoupled from the hot, ductile underlying mantle of the convecting Earth at the lithosphere–asthenosphere boundary (LAB). The detailed nature of this global boundary, despite many years of intense study, is still very poorly known. Yet understanding the seismic and rheological character of the decoupling zone between lithosphere and asthenosphere is a critical element in understanding the dynamical interaction between the lithospheric plates moving across the surface of the globe and the convecting asthenospheric mantle beneath. While seismic studies of the LAB beneath the ocean basins show that in many cases – although by no means all – the LAB can plausibly be related to an isotherm or rheological interface, the LAB beneath continents is far more problematic, particularly beneath the stable continental interior, where the LAB has been remarkably difficult even to detect, and its specific nature remains largely unknown.

Much of the information gleaned over the past 2 decades on the velocity structure and thickness of the continental lithosphere comes from seismic tomography, which images 3-D variations in velocity distribution in the upper mantle. Where continental lithospheric mantle is present, it is seen as a higher velocity layer overlying a lower velocity asthenosphere. On the other hand, tomographic imaging, while an invaluable tool for 3-D mapping of velocity structure, provides little or no information as to the sharpness of the lithosphere–asthenosphere boundary or the velocity contrast across it. This is a significant limitation, as knowledge of the fine structure of the LAB is crucial to sorting out the competing effects of varying geothermal gradients and the presence or absence of hydrous/carbonate fluids or partial melts for decoupling the continental lithosphere from the mobile mantle beneath.

Direct approaches to determining velocity structure at the base of the lithosphere generally involve investigating the small seismic signals produced either by the *conversion* or the *reflection* of teleseismic body waves incident on the underside of the lithosphere–asthenosphere discontinuity (e.g., [Rychert et al., 2005; Sacks and Snoke, 1977]). A converted phase is generated when some fraction of the energy of a compressional/shear wave that is obliquely incident on a velocity discontinuity in the Earth is converted to a shear/compressional wave. One widely used method, *receiver function analysis* (see *Seismic, Receiver Function Technique*), can effectively isolate the small signals on seismic records that represent conversions from the LAB. In principle, the waveform and amplitude of converted phases can provide detailed information about the structure (sharpness) and the velocity contrast – including the sign of the velocity contrast – across a discontinuity. Similarly, underside reflections from the LAB, which are seen on seismic records as small precursors to the large amplitude SS or PP surface reflections from distant earthquakes can be used to determine the sharpness and velocity contrast across the LAB. Unlike conversions, however, underside reflections sample substantial areas of the LAB (typically ~1,000 km in the case of SS phases) and, therefore, measure average regional, rather than local, structure.

Unlike most discontinuities in the Earth, the LAB represents a reversal in velocity with depth, the effect of which is to produce a converted or reflected signal with opposite polarity to that produced by a velocity increase. Results of receiver function studies of converted phases from the LAB have been decidedly mixed. An important study published in 2005 (Rychert et al., 2005) revealed a sharp LAB at a depth of about 100 km beneath northeastern USA (Figure 3). The authors suggested on the basis of the magnitude of the negative shear-wave velocity contrast across the LAB (~5–10%) and the sharpness of the velocity transition (~10 km or less) that the LAB cannot be produced by thermal or compositional contrast alone, but must represent a transition from solid (melt-free) lithosphere to a porous asthenosphere containing at least a few percent of partial melt or fluid phase. Thus, the LAB could correspond to the solidus, or point of incipient



Lithosphere, Continental, Figure 3 A schematic 3-D rendering of LAB conversion points beneath six permanent seismic stations (*red* triangles) in northeastern USA (From [Rychert et al., 2005]. With permission). Inset map at top shows location of 3-D rendering. Topography is shown as shaded relief on the upper 3-D panel. The lower surface represents the base of the lithosphere, which ranges in depth from 90 km (*orange*) to 110 km (*pink*). *Blue* circles on the LAB discontinuity surface indicate the conversion points of the Ps phases. Black lines connect piercing points to the station at which the conversion is observed (Rychert et al., 2005).

melting, in the mantle beneath the continent, or it could mark the depth at which an impermeable barrier (the lithosphere) blocks the upward migration of fluids or partial melts from deeper in the mantle. In either case, the LAB in the Rychert et al. study represents a remarkably sharp rheological boundary, implying an abrupt transition from rigid lithosphere to ductile asthenosphere. Subsequent studies employing similar methodologies to those of Rychert et al., however, have proven difficult to interpret (e.g., [Rychert and Shearer, 2009]), in that LAB depths so measured, particularly those beneath the stable continental interior, are incompatible either with tomographic images or with expected lithospheric thickness based on xenolith data or heat flow analysis (Pollack, 1986; Romanowicz, 2009). Thus, while the lithosphere–asthenosphere boundary beneath continents may be sharp and readily identified in some areas, notably tectonic regions or those of younger age, it would appear that beneath large parts of the stable continents the LAB may simply be too diffuse and/or heterogeneous to be identified as a single coherent discontinuity.

Summary and conclusions

Lithosphere, as it is understood in plate tectonic phraseology, is a term that was developed almost entirely in the context of oceanic plates. There, the discontinuity separating oceanic lithosphere from asthenosphere is widely recognized to be a rheological boundary between a strong melt-free plate and a weak, probably partially molten asthenosphere. In many important respects, we now know that this easily visualized oceanic model – universally displayed in textbooks both elementary and advanced – is a cartoonish oversimplification when applied to the continental lithosphere. Although there do appear to be regions where the continental lithosphere is almost as well defined as it is beneath ocean basins, the distinction between lithosphere and asthenosphere at depth beneath the continents, specifically the boundary between the two, is still shrouded in uncertainty. The failure of our ability seismologically to determine where lithosphere ends and asthenosphere begins has major implications for plate tectonic models in a dynamical earth. If we do not understand the coupling (or lack thereof) between lithosphere

and asthenosphere, and the depth at which that occurs, then the task of sorting out the geodynamical drivers for plate tectonics becomes an exceedingly difficult exercise. While seismology ultimately has the ability to ferret out the truth of how continental plates are coupled or decoupled from the mobile mantle beneath, we are not quite at that point in the science. Nonetheless, the study of the continental lithosphere, particularly the lithospheric mantle, is entering a period of unprecedented expansion. In the USA, the EarthScope program's Transportable Array of 400 broadband seismic systems has migrated to the country's midsection, with nearly 1,000 stations occupied at the time of this writing. Similar large-scale programs are well underway in Japan, China and Europe, giving hope that within the next several years the continental lithosphere and the lithosphere–asthenosphere boundary will have been mapped at high resolution over much of the globe's surface.

Bibliography

Boyd, F. R., and McCallister, R. H., 1976. Densities of fertile and sterile garnet peridotites. *Geophysical Research Letters*, **3**(9), 509–512.

Carlson, R. W., Pearson D. G., and James D. E., 2005. Physical, chemical and chronological characteristics of continental mantle. *Reviews of Geophysics*, **43**, RG1001, 1–24.

Forsyth, D. W., 1989. Lithosphere: mechanical properties. In James, D. (ed.), *The Encyclopedia of Solid Earth Geophysics*. New York: Van Nostrand Rhinehold, pp. 655–660.

James, D. E., 2007. Crust and lithospheric structure – Natural source portable array studies of the continental lithosphere. In Dziewonski, A. M., and Romanowicz, B. (eds.), *Treatise on Geophysics*. New York: Elsevier, pp. 479–531.

James, D. E., and Fouch, M. J., 2002. Formation and evolution of Archaean Cratons: insights from Southern Africa. In Ebinger, C., Fowler, C. M. R., and Hawkesworth, C. J. (eds.), *The Early Earth: Physical, Chemical and Biological Development*. London: Geological Society, pp. 1–26.

James, D. E., Boyd, F. R., Schutt, D., Bell, D. R., and Carlson, R. W., 2004. Xenolith constraints on seismic velocities in the upper mantle beneath southern Africa. *G-cubed*, **5**, 1–32, doi:10.1029/2003GC000551(Q01002).

Jordan, T. H., 1978. Composition and structure of the continental tectosphere. *Nature*, **274**, 544–548.

Jordan, T. H., 1981. Continents as a chemical boundary layer. In *The origin and evolution of the Earth's continental crust*, Transactions Royal Society London, Series A, Vol. 301, number 1461, pp. 359–373.

MacGregor, I. D., and Manton, W. I., 1986. Roberts Victor eclogites: ancient oceanic crust. *Journal of Geophysical Research*, **91**, 14063–14079.

Morgan, P., 1984. The thermal structure and thermal evolution of the continental lithosphere. *Physics and Chemistry of the Earth*, **16**, 107–193.

Pollack, H. N., 1986. Cratonization and thermal evolution of the mantle. *Earth and Planetary Science Letters*, **80**, 175–182.

Ritsema, J., and van Heijst, H., 2000. New seismic model of the upper mantle beneath Africa. *Geology (Boulder)*, **28**(1), 63–66.

Romanowicz, B., 2009. The thickness of tectonic plates. *Science*, **324**, 474–476.

Rudnick, R. L., McDonough, W. F., and O'Connell, R. J., 1998. Thermal structure, thickness and composition of continental lithosphere. *Chemical Geology*, **145**, 395–411.

Rychert, C. A., and Shearer, P., 2009. A global view of the lithosphere-asthenosphere boundary. *Science*, **324**, 495–498.

Rychert, C. A., Fischer K. M., and Rondenay S., 2005. Scattered wave imaging of a sharp lithosphere-asthenosphere boundary beneath eastern North America. *Nature*, **436**, 542–545.

Sacks, I. S., and Snoke, J. A., 1977. The use of converted phases to infer the depth of the lithosphere-asthenosphere boundary beneath South America. *Journal of Geophysical Research*, **82**, 2011–2017.

Shapiro, S. S., Hager, B. H., and Jordan, T. H., 1999. Stability and dynamics of the continental tectosphere. In van der Hilst, R. D., and McDonough, W. F. (eds.), *Composition, Deep Structure and Evolution Of Continents*. Amsterdam: Elsevier, pp. 115–133.

Shirey, S. B., Carlson, R. W., Richardson, S. H., Menzies, A. H., Gurney, J. J., Pearson, D. G., Harris, J. W., and Wiechert, U., 2001. Emplacement of eclogite components into the lithospheric mantle during craton formation. *Geophysical Research Letters*, **28**(13), 2509–2512.

Sipkin, S. A., and Jordan, T. H., 1980. Multiple ScS travel times in the Western Pacific: implications for mantle heterogeneity. *Journal of Geophysical Research*, **85**, 853–861.

West, J. D., Fouch, M. J., Roth, J. B., and Elkins-Tanton, L. T., 2009. Vertical mantle flow associated with a lithospheric drip beneath the Great Basin. *Nature Geoscience*, **2**, 439–444.

Zandt, G., Gilbert, H. J., Owens, T. J., Ducea, M., Saleeby, J., and Jones, C. H., 2004. Active foundering of a continental arc root beneath the southern Sierra Nevada in California. *Nature*, **431**, 41–46.

Cross-references

Heat Flow, Continental
Lithosphere, Continental: Thermal Structure
Lithosphere, Mechanical Properties
Lithosphere, Oceanic
Lithosphere, Oceanic: Thermal Structure
Mantle Viscosity
Seismic, Receiver Function Technique
Seismological Networks

LITHOSPHERE, CONTINENTAL: THERMAL STRUCTURE

Claude Jaupart[1], Jean-Claude Mareschal[2]
[1]Institut de Physique du Globe, Paris, France
[2]Centre GEOTOP-UQAM, University of Québec at Montréal, Montréal, QC, Canada

Definition

Thermal conduction is the dominant form of heat transport across the lithosphere. The thermal structure of the continental lithosphere and its thickness are determined by the distribution of heat-producing elements in crust and lithospheric mantle, and the boundary condition at its base. Vertical seismic velocity profiles, geothermobarometry on mantle xenoliths, and sedimentary records of subsidence provide additional constraints on lithospheric thickness and temperature. The base of the lithosphere can be defined as the intersection of the conductive temperature profile with the mantle isentrope.

Introduction

The lithosphere is the superficial shell, or *boundary layer*, that lies at the top of the convecting mantle and experiences little internal deformation. The concept of lithosphere has evolved over the years, leading to various definitions and estimates of its thickness. The base of the lithosphere was initially defined by a seismic low-velocity zone associated with partial melting. In the oceans, it was also determined by the variations of sea floor depth and surface heat flux as a function of age. It was clear from the beginning that all these definitions are related to the thermal structure. For instance, partial melting requires temperatures above the solidus of mantle rocks. These definitions were introduced, and tested, in the oceans with varying degrees of success but could not be readily extended to the continents. Over large continental areas, no seismic low-velocity zone can be detected and, for a long time, the meager scatter-shot heat flux data set allowed conflicting interpretations.

In the oceans, the lithosphere is made out of homogeneous starting material that is well characterized. It is affected by few thermal perturbations and tectonic events during its short residence time at Earth's surface. Its thermal structure and thickness are essentially functions of only one variable, age, and evolve in simple ways that can be studied with several independent geophysical techniques. In comparison, the continental lithosphere is heterogeneous and thick. Not only was it formed a long time ago by processes that are still debated today, but it has been affected by a succession of perturbations and tectonic events which have left their imprints. Its thermal structure is sensitive to the large amounts of heat-producing elements in the crust and cannot be characterized by age only.

Because the continental lithosphere is thick, its thermal relaxation time is very large, which offers opportunities to study mechanisms that are no longer active today. Because of its old age, it preserves structural and chemical records of ancient geological processes. From a purely thermal standpoint, it plays an important role in regulating the Earth's heat loss and in storing large quantities of radioactive elements that are no longer available to power mantle convection.

The thermal boundary layer of mantle convection

The lithosphere can be defined from a purely thermal standpoint. There are two basic mechanisms of heat transport within the Earth, conduction and convection. Convection refers to the transport of energy by matter that is set in motion by buoyancy forces. Without mass transport, energy can be transferred by conduction in association with a temperature gradient. These mechanisms are important in different parts of the Earth. At shallow depths beneath the surface, the cold upper layer does not deform easily on geological timescales, so that conduction is the dominant heat transport mechanism. The lithosphere may be defined from this perspective, but this definition does not account for heat that is supplied by convection from below. One must add a relatively thin convective boundary layer at the base of the lithosphere, which connects the rigid and purely conductive upper region to the well-mixed mantle below. In this basal boundary layer, heat transport by conduction is not negligible, as discussed below in more detail.

Thermal structure

In the continental lithosphere, heat is in part supplied from below by convection and in part generated by the decay of radioactive elements. In steady state and assuming that heat transport occurs in the vertical direction only, the heat balance equation is:

$$0 = -\frac{dq}{dz} + H \tag{1}$$

where z is depth, q is the vertical heat flux, and H the rate of heat production. In the upper part of the thermal boundary layer, there is no convection and the heat flux is given by the Fourier law of heat conduction:

$$q = -\lambda \frac{dT}{dz} \tag{2}$$

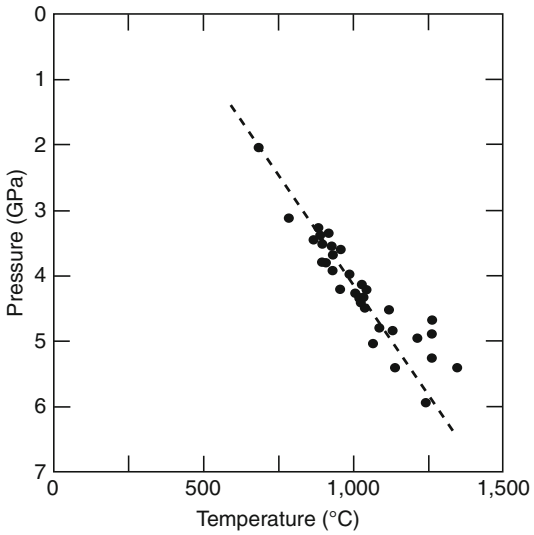
where T is temperature and λ thermal conductivity. The temperature gradient can be measured at Earth's surface in deep boreholes, and determined in the lithospheric root below the crust using the thermodynamic equilibrium conditions between mineral assemblages found in sample from the mantle, mantle xenoliths found in kimberlite pipes. Typical values in stable continents are 15 K km^{-1} and 5 K km^{-1}, respectively (Figure 1). The difference between these two values is due to radiogenic heat production in the continental crust. Below the lithosphere, the temperature gradient is dictated by the convective regime and is expected to be close to an isentrope. By definition, entropy is conserved when going up or down an isentrope, such that:

$$\frac{dT_S}{dz} = \frac{\alpha g T_S}{C_p} \tag{3}$$

where α is coefficient of thermal expansion, g is acceleration of gravity, T_S is temperature, and C_p is specific heat at constant pressure. For the Earth's mantle, the isentropic gradient is typically 0.5 K km^{-1}, which is much less than the conductive gradient in the lithospheric mantle. One identifies the isentrope by its *potential temperature*, that is, the value of its temperature at atmospheric pressure.

Three different depths may be defined in the thermal boundary layer (Figure 2). The shallowest boundary, h_1, corresponds to the base of the rigid upper part and of what we shall call the thermal lithosphere. The deepest boundary, at depth h_3, corresponds to the top of the well-mixed mantle and lies along the mantle isentrope. With no knowledge of boundary layer characteristics and heat transport mechanisms, one cannot determine h_1.

An intermediate depth, h_2, is obtained by downward extrapolation of the conductive geotherm to the mantle isentrope. h_3 is obtained from seismic velocity anomalies, as it is such that temperatures do not deviate significantly from those beneath oceans or from the average mantle velocity profile. These thickness determinations say nothing about h_1. Yet, it is h_1 that defines the mechanically coherent unit (the "plate"), which moves at Earth surface and sets the thermal relaxation time that follows tectonic and magmatic perturbations. Uncertainty on this thickness has severe consequences because the diffusive relaxation time is $\propto h^2/\kappa$, where κ is thermal diffusivity.



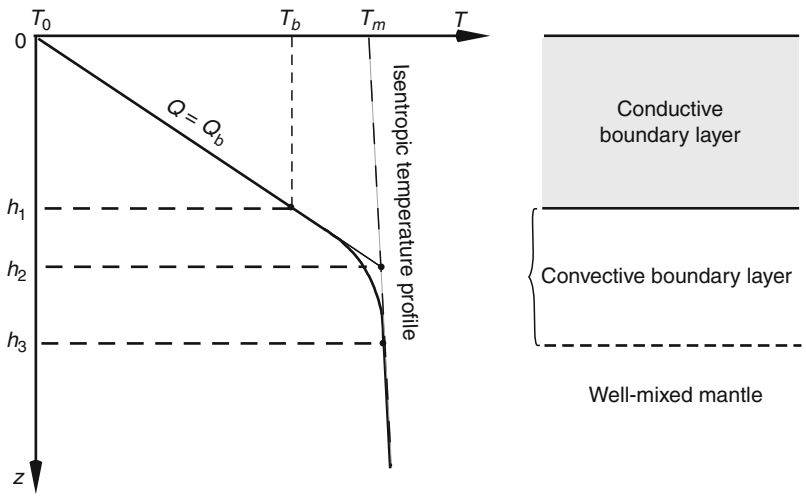
Lithosphere, Continental: Thermal Structure, Figure 1 Pressure and temperature estimates from studies of mineral assemblages in xenoliths from the lithospheric mantle beneath the Kirkland Lake kimberlite pipe, Abitibi sub-province, Canadian Shield. Data are taken from Vicker (1997). The best fit (*dashed line*) indicates a temperature gradient of 5 K km^{-1}. Note that it intersects Earth's surface at a temperature of about 400°C, showing that the temperature gradient is larger than 5 K km^{-1} in the crust.

Three contributions to the surface heat flux

In steady state, Q_0, the heat flux at Earth's surface, is obtained by integrating the heat balance Equation 1 and can be broken down into three components:

$$Q_0 = Q_{\text{crust}} + Q_{\text{lith}} + Q_b \tag{4}$$

where Q_{crust} and Q_{lith} stand for the contributions of heat sources in the crust and in the lithospheric mantle and Q_b is the heat flux at the base of the lithosphere. To describe heat transport mechanisms and the lithosphere structure, one must introduce three temperatures, T_0 at the upper boundary, which, for all practical purposes, may be taken as fixed and equal to 0°C, T_b at the base of the lithosphere, and T_m along an isentrope in the well-mixed convective interior (Figure 2). One can reasonably assume that the mantle potential temperature is the same beneath continents and oceans, and hence that it can be deduced from the composition of mid-ocean ridge basalts (Kinzler and Grove, 1992). Today, this temperature is about 1,325°C. T_b, the temperature at the base of the thermal lithosphere (the rigid lid), cannot be determined from geophysical measurements.



Lithosphere, Continental: Thermal Structure, Figure 2 *Left*: Schematic vertical temperature profile through the continental lithosphere, illustrating three different thicknesses and temperatures. Thickness h_1 corresponds to the rigid upper part of the thermal boundary layer. h_2 is determined by downward continuation of the temperature profile in the rigid upper part to the isentrope that approximates the temperature profile in the well-mixed convective mantle. h_3 denotes the base of the thermal boundary layer, such that temperature is along the well-mixed isentrope. *Right*: three different regions with different heat transport characteristics. Heat is transported by conduction only above depth h_1 and by convection only below depth h_3. Between these two depths, a convective boundary layer is such that both heat transport mechanisms are important.

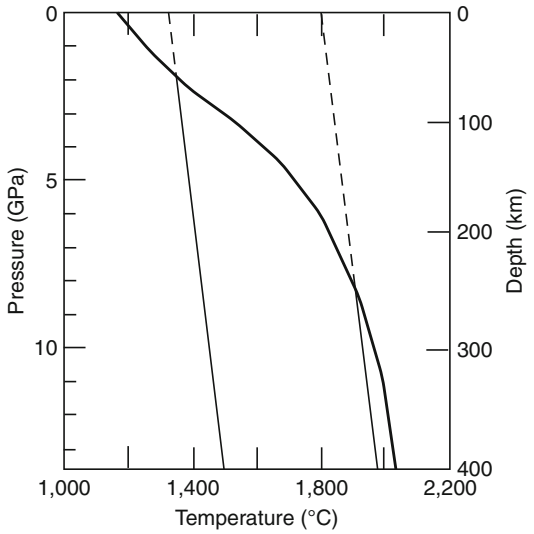
In the oceans, one may ignore Q_{crust} and Q_{lith} and, for sea floor older than 100 My, the surface heat flux is equal to the basal heat flux. In this case:

$$Q_0 = Q_b = \lambda \frac{T_b - T_0}{h_1} = \lambda \frac{T_m - T_0}{h_2} \quad (5)$$

Well-established convection theory leads to a closure equation relating Q_b to the temperature difference across the convective boundary layer, $(T_m - T_b)$ (Davaille and Jaupart, 1993; Solomatov, 1995). In this case, therefore, measurement of the surface heat flux Q_0 allows determination of the basal heat flux and of the temperature difference $(T_m - T_b)$. In turn, this leads to the values of thicknesses h_1 and h_2 as well as of basal temperature T_b. In the oceans, therefore, the values of all the variables can be deduced from surface measurements. In the continents, the problem is more complicated because the surface heat flux includes the crustal contribution Q_{crust}, which may vary by large amounts. This crustal component is typically about 30 mW m^{-2} out of a total of 42 mW m^{-2} in Archean (i.e., older than 2.5 Gy) cratons. In such conditions, surface heat flux measurements do not provide a direct measure of the heat flux at the base of the lithosphere, which must be inferred from other data with methods that will be discussed below. Q_{crust} depends on the composition of the continental crust, which itself reflects the past history of magma emplacement and tectonic deformation/accretion. Observations show that it varies by large amounts even in a single geological province. It cannot be calculated from first physical principles and must be determined on a case-by-case basis using a host of different measurements.

Controls on lithosphere thickness

The lithosphere owes its physical properties to partial melting and melt extraction processes which leave a solid residue that is both less dense and mechanically stronger than the starting mantle material. Mechanical strength is due to both the lower temperature in the lithosphere than in the convective mantle, and to dehydration that occurred during partial melting (Pollack, 1986). Melting may be achieved by two different mechanisms operating in different settings. One mechanism is decompression in an upwelling and has been documented in considerable detail in mid-ocean ridges (McKenzie and Bickle, 1988). Decompression melting begins at a depth, which depends on mantle temperature and water content (e.g., 60–80 km for the oceans today). Continents are not linear features, and were probably not generated by linear mantle upwellings such as present-day mid-ocean ridges. Their formation has been linked to large mantle plumes or subduction zones (Carlson et al., 2005). For mantle plumes, the mechanism is basically the same as that beneath mid-ocean ridges and the depth where melting starts dictates the thickness of the buoyant solid residue. In a dry mantle, high potential temperatures are required for the generation of thick lithospheric roots (Herzberg, 1983) (Figure 3). For a thickness



Lithosphere, Continental: Thermal Structure, Figure 3 Solidus temperature of dry mantle peridotite as a function of pressure and depth, from Herzberg (1983). Also shown are two isentropes that intersect the solidus at depths of 60 and 250 km.

of 250 km, for example, the mantle potential temperature must be about 1,800°C. On the other hand, if the Archean mantle was wet, melting occurred at shallower depths and at lower temperatures (Grove and Parman, 2004). In this case, a second event is required to account for the large thickness of continental lithosphere, involving large-scale compression and/or thrusting. For both types of models, the lithosphere thickness provides a strong constraint on the mechanism of continent formation.

Heat flux at the base of the continental crust and at the base of the lithosphere

Constraints on the thermal structure of the continental lithosphere may be obtained with several geophysical methods. To obtain continuous temperature profiles, however, one can only use surface heat flux data and downward continue shallow temperature measurements. To this aim, one needs to specify the heat flux at the base of the lithosphere. A first step is to determine the amount of heat produced in the continental crust or, alternatively, the heat flux at the Moho discontinuity.

Crustal heat production

Uranium and thorium are the main heat-producing elements in rocks. They are located mostly in accessory minerals and grain boundaries, which depend weakly on the bulk chemical composition. Thus, their concentrations are not related to physical properties such as density and seismic velocity and cannot be retrieved from large-scale geophysical studies. In addition, they vary by large amounts at all scales, from that of a petrological thin section to that of a whole massif. Within an apparently

homogeneous pluton, they can change in both vertical and horizontal directions due to fluid migration and late-stage alteration. In the Bohus granite, Sweden, for example, thorium concentrations vary by a factor of 5 over horizontal distances as small as a few tens of meters (Landstrom et al., 1980).

In a geological province, radioelement concentrations cannot be estimated using data from other provinces and must be measured in all the major rock types present. They depend on lithology and are high in granites and metasediments, so that a geological map gives a rough idea of the spatial distribution of heat production. With the very wide ranges that exist for each rock type, however, the standard deviation of the heat production distribution is large, and often larger than the mean. As a consequence, a single thermal model for a province would gloss over the important lateral temperature variations that occur due to the heterogeneous crustal structure. For a reliable thermal calculation in a specific area, one cannot use the local heat production values of a few rock types found on the outcrop because temperatures are sensitive to heat that is generated over rather large volumes. Average heat production values must be determined on a scale that is intermediate between the dimensions of individual plutons and the size of the province.

Vertical variations of heat production can be determined in deep boreholes or from exposed crustal sections. On a vertical scale of ≈10 km, data from the deep boreholes at Kola, in the Russian part of the Baltic Shield, (Kremenentsky et al., 1989) and the KTB, Germany, (Clauser et al., 1987), show no systematic change with depth. At KTB, heat production between 8 and 9 km depth is the same as between 1 and 2 km, and higher than above 1 km. At Kola, heat generation in the Archean rocks between 8 and 12 km is higher (1.47 μW m^{-3}) than in the shallower Proterozoic section (0.4 μW m^{-3}).

Over a crustal thickness scale, heat production is lower in mid-crustal assemblages than in the upper crust. This vertical variation is not monotonous and cannot be described by a simple function valid everywhere, as shown by exposed crustal sections such as the Vredefort in South Africa (Nicolaysen et al., 1981), the Cordilleran core complexes of Arizona (Ketcham, 1996), or the Pikwitonei-Sachigo and Kapuskasing-Wawa areas of Canada (Fountain et al., 1987; Ashwal et al., 1987). One robust feature is the depletion of lower crustal assemblages in comparison to upper and middle crustal ones (Table 1).

Table 1 lists estimates derived from global geochemical models of the crust and a very large number of measurements. The crust is split into three different crustal horizons corresponding to average regional seismic models. The model crust is 40 km thick with middle crust in the 12–23 km depth range, and is such that Q_{crust} = 36 mW m^{-2}. Another method relies on heat flux data and will be explained below (Table 2). With this method, one may also evaluate heat production variations between Provinces. The two independent methods lead to comparable results and show that crustal heat generation typically accounts for more than half of the surface heat flux.

Lithosphere, Continental: Thermal Structure,
Table 1 Element concentration and heat production in the continental crust, from Rudnick and Gao (2003). Element concentrations are given in ppm

	U	Th	K	μW m^{-3}
Upper crust	2.7	10.5	23,300	1.65
Middle crust	1.3	6.5	19,200	1.00
Lower crust	0.2	1.2	5,080	0.19
Total crust	1.3	5.6	15,080	0.89

Lithosphere, Continental: Thermal Structure,
Table 2 Estimates of bulk continental crust heat production from heat flux data, from Jaupart and Mareschal (2003)

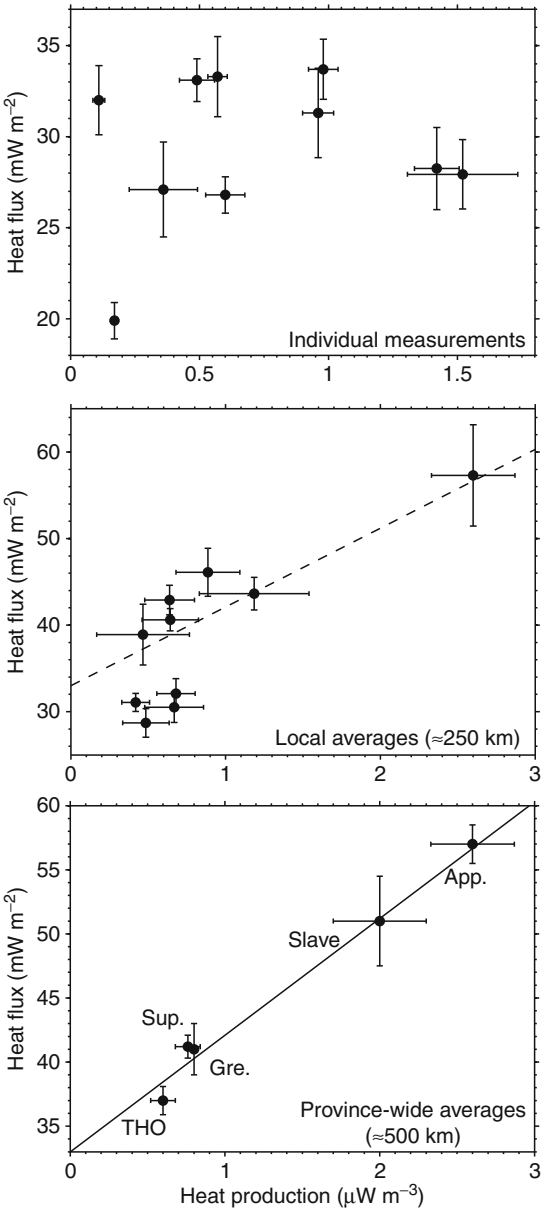
Age group	A^a	Q^b_{crust}
Archean	0.56–0.73	23–30
Proterozoic	0.73–0.90	30–37
Phanerozoic	0.95–1.21	37–47
Total continents	0.79–0.99	32–40

[a]Range of heat production in μW m^{-3}
[b]Range of the crustal heat flow component in mW m^{-2}

Heat flux and heat production

With the large variations in the concentrations of heat-producing elements that have been documented in crustal rocks, it is not surprising that, in continents, the surface heat flux changes by large amounts over small horizontal distances of a few tens of kilometers. The largest anomalies occur in association with enriched granitic plutons. These plutons are usually of small size, and the local value of the heat flux is not representative of the average crust that lies below the measurement site. In such conditions, studies of the thermal structure of continents require large data sets. The challenge is to obtain average values of heat flux and heat production that are representative of the whole crust in an area. We discuss the scales over which this can be done in a reliable manner.

There is only a weak relationship between local values of heat flux and heat production in a geological province (Figure 4). For a representative thermal model, one must determine an average heat flux value over an area that is sufficiently large for the smoothing of small-scale variations due to isolated anomalous massifs. Figure 4 also shows data for 10 ≈ 250 km × 250 km windows distributed in North America. These are located in the same area as the individual heat flux values discussed previously, and show that a relationship between heat flux and heat production begins to emerge at this scale. Even at this scale, the relationship is not defined tightly due to insufficient sampling of heat production. A single heat flux value



Lithosphere, Continental: Thermal Structure,
Figure 4 Relationship between heat flux and radiogenic heat production in North America at different scales, from Levy et al. (2010). *Top panel*: for individual measurements, there is clearly no relationship between the two. Middle panel: at an intermediate scale of about 200 km, a relationship between the average values of heat flux and heat production begins to emerge. The *dashed line* corresponds to the best-fit linear relationship defined by province-wide averages. *Bottom panel*: shows a strong correlation between the average values of heat flux and heat production for the five major geological provinces of North America. The best-fit linear relationship has a slope of 9 km and an intercept of 33 mW m^{-2}. Symbols are as follows: App, Appalachian province (400 My); Gre, Grenville province (1,100 My); THO, Trans-Hudson Orogen (1,800 My); Sup, Superior province (>2,500 My); Slave, Slave province (>2,500 My).

is sensitive to crustal heat production over large horizontal distances and records the contribution of many different geological units. In contrast, the heat production values correspond to a small number of rocks found at shallow depth in boreholes used for heat flux determinations. The limited sampling achieved at a scale of 250 km does not account for the full range of rock types and heat production rates in the crust. For representative heat production averages, one needs to work at a larger scale. Figure 4 shows data for five major geological provinces in North America with typical dimensions of ≈500 km × 500 km. These provinces cover the geological history of the stable continent and represent different types of continental crust. Furthermore, they have been extensively deformed and eroded, so that rocks from a large range of depths can be found at the surface. The metasedimentary-plutonic belts of the western Superior Province, Canadian Shield, for example, include rocks from all metamorphic grades up to granulite facies. The province-wide averages exhibit a remarkable linear relationship, with a heat flux intercept of ~33 mW m^{-2} corresponding to crust with zero surface heat production. The data for the 10 ≈ 250 km × 250 km windows lie close to this global linear relationship (Figure 4).

The Moho heat flux is equal to:

$$Q_M = Q_0 - Q_{\text{crust}} = Q_{\text{lith}} + Q_b \qquad (6)$$

which depends on heat production in the lithospheric mantle and on the heat flux at the base of the lithosphere. Both components come from depths larger than the crustal thickness, so that lateral variations, if they exist, get smoothed out by horizontal diffusion. For an average lithospheric thickness of 250 km, variations of the basal heat flux Q_b over wavelengths smaller than 500 km are not detectable in the surface heat flux (Mareschal and Jaupart, 2004). A similar statement can be made for the contribution of heat sources in the lithospheric mantle, Q_{lith}. Thus, on scales that are smaller than 500 km, one must consider that the Moho heat flux is uniform and that variations of the surface heat flux come solely from changes of crustal heat production.

The strong correlation that is observed between the province-wide average surface heat flux and heat production indicates that variations of the Moho heat flux are small throughout the whole North American continent (Figure 4). If there were large variations of the Moho heat flux, they would need to be compensated by opposite variations of the average lower crustal heat production. No physical mechanism can explain how such independent variables can be linked to one another. Variations of the Moho heat flux are therefore less than the intrinsic uncertainty on the heat flux measurements and the magnitude of departures from the heat flux–heat production relationship, which is about 2–3 mW m^{-2} (Mareschal and Jaupart, 2004). The absolute value of the Moho heat flux is not specified by this analysis, however, and this issue is discussed in the next section.

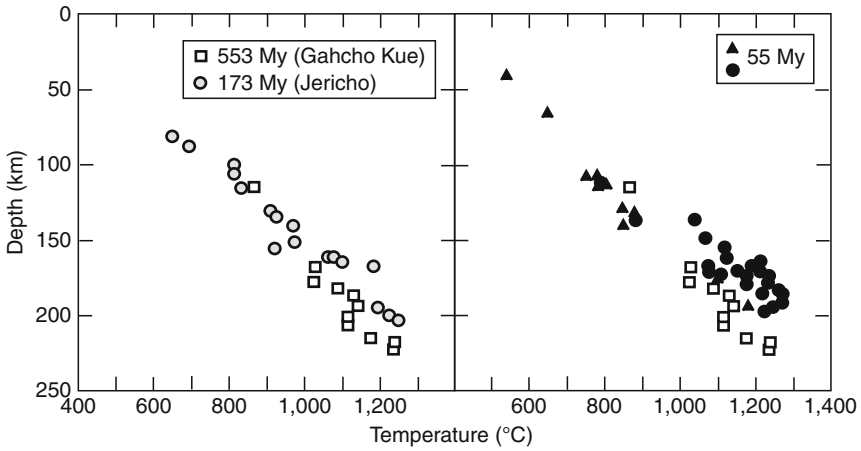
The Moho heat flux

Direct determination of the heat flux through the lithospheric mantle can be made using xenolith samples brought from lithospheric depths by powerful *kimberlitic* eruptions. Coexisting mineral assemblages and mineral compositions allow estimates of pressure and temperature and of the geothermal gradient. Together with a thermal conductivity estimate, they lead to values of the heat flux beneath the crust. For the Kirkland Lake pipe, within the Superior Province, Canada, (Figure 1), Rudnick and Nyblade (1999) obtained a best-fit Moho heat flux estimate of ≈18 mW m^{-2} within a total range of 17–25 mW m^{-2}. Russell and Kopylova (1999) derived a best-fitting value of 15 mW m^{-2}, within a range between 12–24 mW m^{-2}, for the Jericho kimberlite area of the Archean Slave Province, also in Canada. Figure 5 shows data from several kimberlite pipes in the Slave Craton (D. Canil, 2007, personal communication). There are significant temperature differences between the different data sets. Surprisingly, the youngest eruptions, which provide recent samples of the lithospheric mantle, yield the highest temperatures. A similar result is obtained in the Kaapvaal craton, South Africa (Michaut et al., 2007). If these temperature variations are indeed due to age and reflect the time evolution of lithospheric temperatures, they are not consistent with secular cooling of the Earth's mantle. The pipes do not overlap so that the data may in fact record lateral variations of temperature in the lithospheric mantle, which can only be explained by changes of heat production. Nevertheless, these data lead to similar values of the temperature gradient and heat flux.

Bounds on the Moho heat flux can be obtained by taking advantage of the variations of heat flux and crustal structure in a geological province. Combined with heat production data for the various rock types involved, one can isolate the variable crustal component and the uniform Moho heat flux. In Canada, Moho heat flux values of 10–15 mW m^{-2} have thus been derived for the Grenville Province, the Trans-Hudson Orogen (THO), for the Abitibi belt of the Superior Province (Jaupart and Mareschal, 2007). Crustal thickness and composition can also be constrained by combining seismic refraction, gravity, and heat flux. Such methods lead to values of Q_M between 7 and 15 mW m^{-2}.

Another method relies on P_n velocities from seismic refraction surveys (Perry et al., 2006). For given values of the surface heat flux, the Moho temperature depends on the amount of crustal heat production and hence on the Moho heat flux. One can therefore obtain a relationship between Moho temperature and seismic velocity with the Moho heat flux as control parameter. One can also derive independently a relationship between temperature and seismic velocity using laboratory measurements of elastic properties and seismic velocities, with the bulk mantle composition as control parameter. The two relationships are only consistent with one another over a restricted range of composition and Moho heat flux. For depleted mantle compositions appropriate to the Superior Province, a good fit between predicted and observed P_n velocity values is obtained if the Moho heat flux is within a range of 12–25 mW m^{-2} (Perry et al., 2006).

The three different and independent methods that we have described have all been applied to the Abitibi belt in the Canadian Shield and can be combined to tighten the final range to 12–15 mW m^{-2} for that sub-province. Lower and upper bounds on the Moho heat flux have been derived using other arguments. Upper bounds on the mantle heat flux are obviously provided by the lowest heat flux



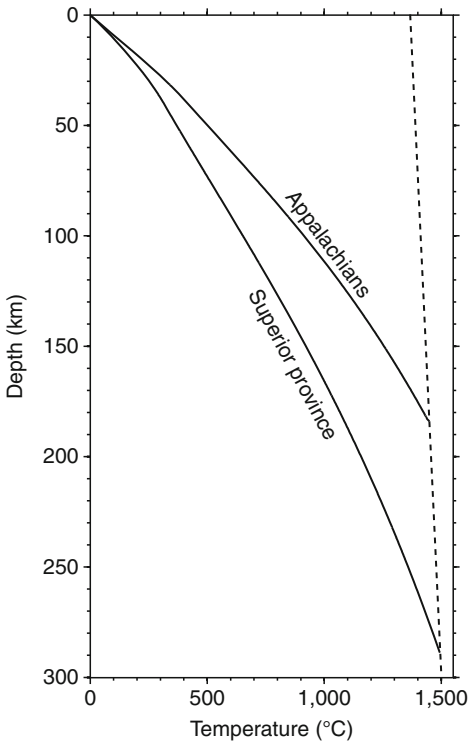
Lithosphere, Continental: Thermal Structure, Figure 5 Relationship between temperature and depth in the lithospheric mantle of the Slave Province, Canada (from D. Canil, 2007, personal communication). The Slave Province is small and has been pierced repeatedly by kimberlite pipes in the same area. *Left panel*: data from the two oldest pipes. *Right panel*: data for two young pipes compared to those for the oldest pipe in the area. On average, temperature estimates from the two young pipes are larger than those of the older pipes by about 150 K. Such differences may indicate lateral temperature variations in the lithospheric mantle.

measured. Values of 20–23 mW m^{-2} have been reported within the Canadian Shield (Mareschal et al., 2000). Regional values as low as 18 mW m^{-2} have also been reported for the Baltic or Siberian Shield. One can refine this estimate to 18 mW m^{-2} using lower bounds on crustal heat production.

The methods that we have described here rely on different data and hence are associated with different sources of uncertainty. That such completely independent methods converge to similar results allows some confidence in the final range obtained. Values lower than 12 mW m^{-2} are not consistent with the xenolith data whereas values higher than 18 mW m^{-2} can be excluded because of the heat flux data. This upper bound appears to be valid for all the shields (Jaupart and Mareschal, 2007).

Lithospheric geotherms

Using information on heat flux values at the surface and at the Moho, one may determine geotherms through the lithosphere. The calculation requires values for thermal conductivity that are discussed in the Appendix. Figure 6 shows two such geotherms for two provinces in North America, which are consistent with both seismic and heat flux data (Levy et al., 2010). These two geotherms intersect the well-mixed mantle isentrope at different depths, indicating large lateral thickness variations for the continental lithosphere. The maximum thickness is reached beneath the Archean Superior province and is estimated to be 280 ± 30 km.



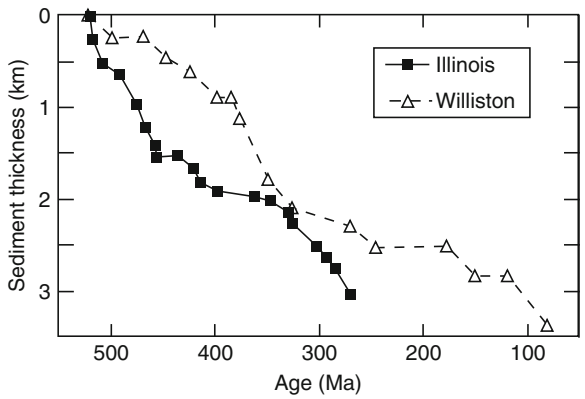
Lithosphere, Continental: Thermal Structure, Figure 6 Two steady-state lithospheric geotherms beneath the Appalachian and Superior provinces of North America, from Levy et al. (2010). These two geotherms were deduced from measurements of the surface heat flux and heat production, and were constrained by requiring consistency with seismic data. The *dashed line* is the isentrope for a potential temperature of 1,350°C.

Thermal transients

Because the continental lithosphere is thick, it is characterized by a long thermal relaxation time. For a thickness of 250 km and a thermal diffusivity κ equal to 10^{-6} m^2 s^{-1}, the characteristic diffusion time $\tau = h^2/\kappa$ is ≈2 Gy. This is much longer than that of the oceanic lithosphere, which is only about 200 My (for a thickness of 80 km). We shall now discuss two types of thermal transients in thick lithosphere.

Thermal relaxation following a thermal/tectonic perturbation

Thermal relaxation of the lithosphere following a deep thermal perturbation can lead to subsidence, which has been recorded in intracratonic sedimentary basins located away from active plate boundaries. Subsidence histories for two such basins of North America, the Michigan and Williston basins, are shown in Figure 7 and emphasize their long durations exceeding 200 My. On the basis of models for heat transport and cooling in the vertical direction only, the thickness of the North American lithosphere is found to about 115 km beneath the Michigan basin and 270 km beneath the Williston (Haxby et al., 1976; Ahern and Mrkvicka, 1984). There is no obvious explanation for such a large thickness difference in the middle of the North American continent, which suggests that the subsidence models may be oversimplified. These models rely on 1-D calculations and hence are only valid if the thermal anomaly stretches over a horizontal distance that is much larger than the lithosphere thickness. The Michigan and Williston basins have been attributed in large part to plate flexure driven by a deep load (Nunn and Sleep, 1984;



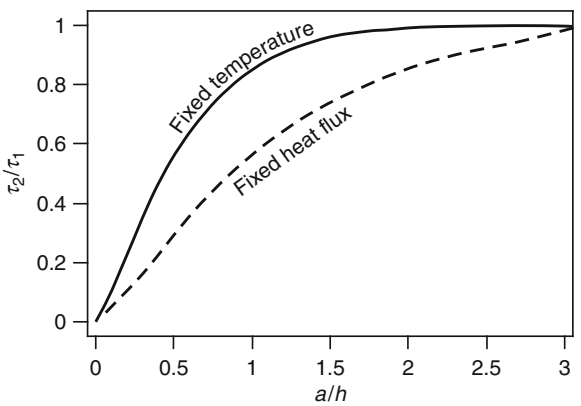
Lithosphere, Continental: Thermal Structure, Figure 7 Subsidence histories for the Michigan and Williston basins, North America.

Ahern and Mrkvicka, 1984). This load is related to the initial thermal anomaly and has a radius of about 120 km beneath the Michigan basin (Haxby et al., 1976), which is smaller than the thickness of the lithosphere beneath the North American craton. In these conditions, the assumption of purely vertical heat transfer is not tenable. Kaminski and Jaupart (2000) have reevaluated thermal models for the continents by taking into account lateral heat transport within thick lithosphere. They considered a lithosphere of thickness h with a cylindrical thermal anomaly of radius a between depths h and b. At the surface ($z = 0$), temperature is fixed at T_0. At steady state, the base of the lithosphere is at $T = T_0 + \Delta T$. Assuming, for example, that the thermally perturbed region is initially at a uniform temperature $T_0 + \Delta T$, temperature can be written as the sum of the equilibrium temperature and a dimensionless perturbation θ:

$$T(r, z, t) = T_0 + \Delta T \left[\frac{z}{h} + \theta(r, z, t)\right], \qquad (7)$$

where r is the radial distance. The temperature perturbation may be expressed in terms of a vertical component θ_Z, which corresponds to the 1-D solution, and a radial component, which can be calculated using Bessel functions. For thermal anomalies with planforms that are not circular, lateral heat transfer is dominated by the smallest horizontal dimension.

The acceleration of cooling due to lateral heat loss can be measured simply by the time to achieve 90% of the final subsidence, noted τ_2, scaled to that for an infinitely wide initial perturbation, τ_1 (corresponding to the 1-D calculation) (Figure 8). Results are sensitive to the boundary condition at the base of the lithosphere and show that the impact of lateral heat transport is significant for $a/h < 1$,



Lithosphere, Continental: Thermal Structure, Figure 8 Time to achieve 90% subsidence in thick lithosphere as a function of the radius of the deep lithospheric thermal anomaly, a (scaled to the lithosphere thickness h). The subsidence time has been scaled to the value obtained with a 1-D thermal model for purely vertical heat transport. Results have been obtained for two different boundary conditions at the base of the lithosphere. (Adapted from Kaminski and Jaupart (2000)).

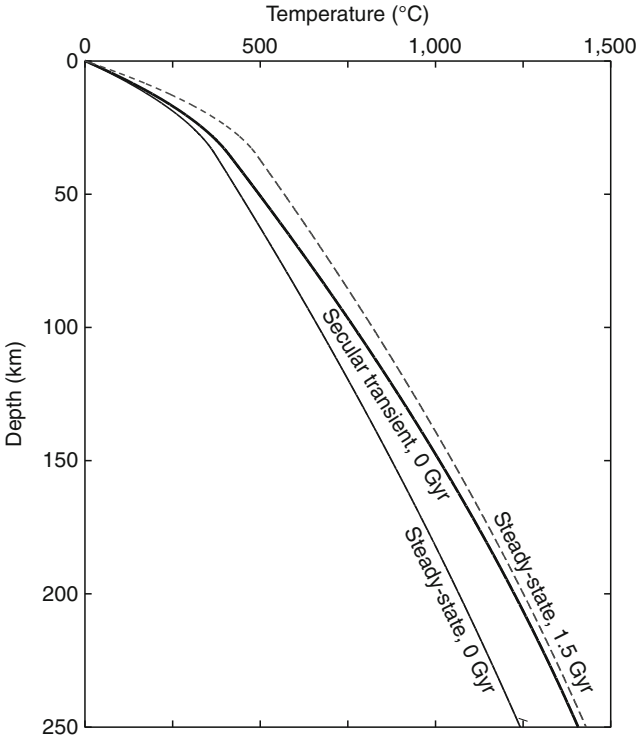
corresponding to $a <\approx 250$ km in the North American craton. This is the case of all the major intracratonic basins. In comparison to the oceans, thermal relaxation in the continents is lengthened by the larger lithosphere thickness and shortened by lateral heat transport. Different subsidence histories and durations are generated by the same basic process operating over different lateral distances.

Secular transients

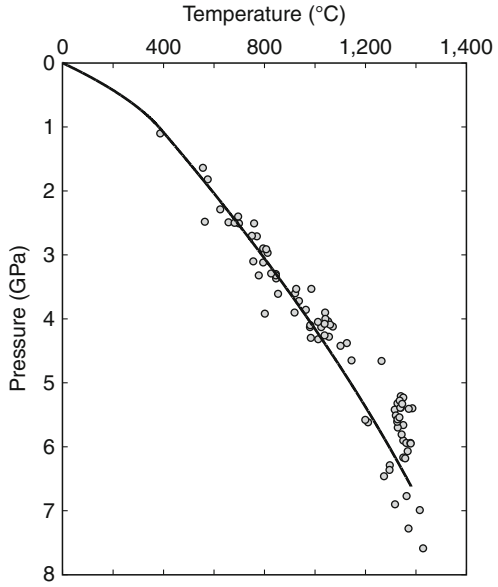
We have shown above that continental temperatures depend strongly on radiogenic heat production. Thus, continents cool down continuously due to the rundown of radioactivity with time. One should also add that conditions at the base of the lithosphere also evolve with time due to secular cooling of the whole Earth. This is ill constrained and we focus our attention to the effects of decaying heat sources. For the Appalachian geotherm of Figure 6, for example, crustal heat production contributes about 300°C. For average Th/U and K/U ratios, radiogenic heat production decreases by a factor of 2 in about 2.7 Gy. We thus deduce that continental temperatures decrease at a rate of about 50 K Gy^{-1}, which is comparable to the bulk mantle cooling rate (Abbott et al., 1994). This simple argument relies on a quasi steady-state approximation, such that ancient geotherms are calculated using the steady-state heat equation together with past values of heat production. For thick lithosphere, this approximation is not valid and may lead to significant errors.

In thick continental lithosphere, the timescale for diffusive heat transport is comparable to the half-lives of Uranium, Thorium, and Potassium. As noted above, the characteristic time for heat diffusion, equal to h^2/κ is 2 Gy for 250 km thick lithosphere, which is close to the decay time of radiogenic heat production (≈2.7 Gy). Thus, heat diffusion is not efficient enough to keep up with the decaying heat sources and lithospheric temperatures are not in equilibrium with the instantaneous rate of heat production. In these conditions, the vertical temperature profile exhibits significant curvature and may be hotter than a steady-state profile by as much as 150 K depending on the absolute value of the heat generation rate (Figure 9). Another consequence is that forcing a steady-state temperature profile through xenolith (P,T) data leads to an overestimate of the mantle heat flux.

For illustration purposes, Figure 9 shows the instantaneous vertical temperature profile with decaying sources and a steady-state geotherm derived for the same present-day values of crust and mantle heat production. The difference between the two geotherms increases with depth because of increasingly inefficient transport of heat to the Earth's surface. For comparison, we have also calculated a steady-state geotherm corresponding to heat production values 1.5 Gy ago. The transient geotherm falls between the steady-state geotherms calculated for heat production rates at 0 and 1.5 Gy, showing that it is sensitive to heat generation in the past. At the base of the



Lithosphere, Continental: Thermal Structure,
Figure 9 Present-day vertical temperature profile for transient thermal model with decaying heat sources (*thick line*) compared with two steady-state geotherms (*solid line* for present-day heat production; *dashed line*, for heat production rate values 1.5 Gy ago) for the same set of parameters (h = 250 km, crustal heat production A_c = 1.0 μW m^{-3}, lithospheric mantle heat production A_m = 0.02 μW m^{-3}, basal heat flux Q_b = 10 mW m^{-2}). Surface heat flow is of 49 mW m^{-2} for the transient solution, 46 mW m^{-2} for the present-day steady-state profile, and 62 mW m^{-2} for the 1.5 Gy steady-state profile.



Lithosphere, Continental: Thermal Structure,
Figure 10 Xenolith P-T data from the Kimberley, Jagersfontein, Premier and Frank Smith kimberlite pipes, Kaapvaal craton, South Africa (Rudnick, 2003, personal communication). *Thick line*: vertical temperature profile for a transient thermal model with decaying heat sources and for a lithosphere thickness of 225 km. Parameter values are crustal heat production A_c = 0.8 μW m^{-3}, mantle heat production A_m = 0 025 μW m^{-3}, and basal heat flux Q_b = 12 mW m^{-2}. Note that the geotherm terminates above several deep data points, due to the starting assumption on lithosphere thickness.

lithosphere, all geotherms are parallel because of the constant heat flux bottom boundary condition. Secular transient behavior is not important at shallow depths, due to the efficiency of diffusive heat transport over small distances. In the crust, differences between the transient and present-day steady-state profiles are very small and can be neglected for all practical purposes. Between 50 and 150 km depth, heat flux is about 13 mW m^{-2} for the present-day steady-state model and 16 mW m^{-2} for the transient one. Maximum departure from thermal equilibrium is achieved at the base of the lithosphere. The magnitude of transient effects essentially depends on mantle heat production as well as on lithosphere thickness. Large values of heat production in the lithosphere do not introduce large transients at shallow depths. In contrast, even small values of heat production lead to significant effects in a thick lithosphere.

The secular transient geotherm is compared to xenolith (P-T) data from Kaapvaal craton, South Africa, for h = 225 km in Figure 10. For this solution, the surface heat flux is 47 mW m^{-2}, which is close to measurements in this region, and heat production in the lithospheric mantle is 0.025 μW m^{-3}, which lies in the lower range of values proposed by Rudnick and Nyblade (1999). The basal heat flux is 12 mW m^{-2}. The Kaapvaal xenolith (P,T) data suggest a change of slope in the geotherm at a pressure of about 5 GPa (Bell et al., 2003). According to some authors, this marks the transition between lithosphere and asthenosphere, with the deepest data points belonging to the convecting mantle below the lithosphere. Rudnick and Nyblade (1999) and Bell et al. (2003) have argued that all data points belong to unperturbed lithosphere. Significant curvature of the secular transient geotherm allows a good fit to all the data and does not suggest a break in the geotherm.

Due to the intrinsic decay of radioactive heat sources, the lithospheric mantle therefore undergoes secular cooling even when thermal conditions at the base of the lithosphere remain steady. Predicted cooling rates are in the range of 50–150 K Gy^{-1}. Such transient behavior also has implications for heat flow studies. Assuming that the decay of radiogenic heat production can be represented by a single exponential, such that $A_m(t) = A_o \exp(-\alpha t)$,

the contribution of lithospheric heat production to the surface heat flux is:

$$\Delta Q_{\text{lith}}(t) = A_o h \frac{tan\sqrt{\alpha\tau}}{\sqrt{\alpha\tau}} exp(-\alpha t) \quad (8)$$

This is compared to the instantaneous amount of heat produced in the lithosphere:

$$\Delta Q_{\text{lith},i}(t) = A_o h exp(-\alpha t) \quad (9)$$

Ratio $Q_{\text{lith}}/Q_{\text{lith},i}$ increases as a function of lithosphere thickness and is 1.5 for h = 300 km, for example. This shows that surface heat flux measurements record some time-average of the deep lithospheric heat production.

Summary

The continental lithosphere is much thicker than its oceanic counterpart and its thermal structure depends strongly on radiogenic heat production in both the crust and in the mantle. Because of these two characteristics, its behavior depends on its geological history and is not a simple function of age. Beneath Archean Shields, it can be as thick as 300 km.

Appendix: thermal conductivity

For an isotropic medium, Fourier's law is:

$$\mathbf{q} = -\lambda \nabla T \quad (10)$$

where λ is the thermal conductivity measured in W m^{-1} K^{-1}. More generally, if the medium is not isotropic, the thermal conductivity must be defined as a second order tensor. Thermal conductivity can be broken into lattice and radiative components, which have different behaviors as temperature and pressure change.

At low temperatures, heat transport is in large part effected by lattice vibrations. Lattice conductivity decreases with increasing temperature.

At larger temperatures, electromagnetic waves transport energy in continuous media. Radiation is rapidly attenuated but the medium reemits energy if the temperature is sufficiently high. In a temperature gradient, the net energy flux will be $\propto T^3 \nabla T$, which has the same form as the flux for lattice conduction. One can therefore define a radiative component of conductivity, λ_R, such that the radiative heat flux is:

$$\mathbf{q_R} = -\lambda_R \nabla T \quad (11)$$

The gray body law states that:

$$\lambda_R = \frac{16}{3} \frac{n^2}{\varepsilon} \sigma T^3 \quad (12)$$

where n is the refractive index, $\in$ is the opacity, and σ is the Stefan constant of black-body radiation (σ = 5.67 × 10^{-8} W m^{-2} K^{-4}).

Lattice conductivity

For individual crystals, theoretical arguments suggest that, below the Debye temperature, the temperature dependence of lattice conductivity takes the following form:

$$\lambda(T) = \lambda_{298} \sqrt{\frac{298}{T}} \quad (13)$$

For polycrystalline assemblages, this equation is no longer valid and one relies on empirical fits to the data. Reasonable agreement can be obtained with a relationship of the type:

$$\lambda(T) = A + \frac{B}{350 + T} \quad (14)$$

where T is in °C and where constants A and B depend on the rock type (Table 3).

The effect of pressure on conductivity can be treated independently of temperature, so that one can write:

$$\lambda(T, P) = \lambda_o(T)(1 + \beta P) \quad (15)$$

where β is a constant coefficient (Table 4). The effect of pressure increases conductivity by about 20% between the surface and 200 km.

Radiative conductivity

For olivine, several independent laboratory measurements suggest that:

$$\lambda_R = 0.368 \times 10^{-9} T^3 \quad (16)$$

This equation leads to $0.13 < \lambda_R < 1.8$ W m^{-1} K^{-1} for $700 < T < 1{,}700$ K. It is only valid in a single crystal if the mean free path of photons is independent of temperature.

Lithosphere, Continental: Thermal Structure,
Table 3 Constants for calculating the thermal conductivity of different rock types, from Clauser and Huenges (1995)

Rock type	T range (°C)	A	B
Metamorphic rocks	0–1,200	0.75	705
Felsic rocks	0–1,400	0.64	807
Mafic rocks	50–1,100	1.18	474
Ultramafic rocks	20–1,400	0.73	1,293

Lithosphere, Continental: Thermal Structure,
Table 4 Pressure dependence of lattice conductivity

Material	β^a (× 10^{-2} GPa^{-1})	Reference
Mafic granulites (Finland)	2.4 ± 0.8	Kukkonen et al. (1999)
Olivine	3.2	Xu et al. (2004)

[a]Coefficient in Equation 15 for thermal conductivity

For mantle rocks, one must account for scattering and for the effect of interfaces in a mineral assemblage. Such complications led Marton et al. (2005) to use a constant radiative conductivity component λ_R = 1 W m^{-1} K^{-1} for temperatures higher than 700 K.

Bibliography

Abbott, D., Burgess, L., and Longhi, J., 1994. An empirical thermal history of the Earth's upper mantle. *Journal of Geophysical Research*, **99**, 13835–13850.

Ahern, J. L., and Mrkvicka, S. R., 1984. A mechanical and thermal model for the evolution of the Williston basin. *Tectonics*, **3**, 79–102.

Ashwal, L. D., Morgan, P., Kelley, S. A., and Percival, J., 1987. Heat production in an Archean crustal profile and implications for heat flow and mobilization of heat producing elements. *Earth and Planetary Science Letters*, **85**, 439–450.

Bell, D. R., Schmitz, M. D., and Janney, P. E., 2003. Mesozoic thermal evolution of the southern African mantle lithosphere. *Lithos*, **71**, 273–287.

Carlson, R. W., Pearson, D. G., and James, D. E., 2005. Physical, chemical, and chronological characteristics of continental mantle. *Reviews of Geophysics*, **43**, 2995–3007.

Clauser, C., and Huenges, E., 1995. Thermal conductivity of rocks and minerals. In Ahrens, T. J. (ed.), *A Handbook of Physical Constants: Rock Physics and Phase Relations*. Washington, DC: AGU, pp. 105–126.

Clauser, C., Gieses, P., Huenges, E., Kohl, T., Lehmann, H., Rybach, L., Safanda, J., Wilhelm, H., Windlow, K., and Zoth, G., 1987. The thermal regime of the crystalline continental crust: implications from the KTB. *Journal of Geophysical Research*, **102**, 18417–18441.

Davaille, A., and Jaupart, C., 1993. Transient high-Rayleigh-number thermal convection with large viscosity variations. *Journal of Fluid Mechanics*, **253**, 141–166.

Fountain, D. M., Salisbury, M. H., and Furlong, K. P., 1987. Heat production and thermal conductivity of rocks from the Pikwitonei-Sachigo continental cross section, central Manitoba: implications for the thermal structure of Archean crust. *Canadian Journal of Earth Sciences*, **24**, 1583–1594.

Grove, T. L., and Parman, S. W., 2004. Thermal evolution of the earth as recorded by komatiites. *Earth and Planetary Science Letters*, **219**, 173–187.

Haxby, W. F., Turcotte, D. L., and Bird, J. M., 1976. Thermal and mechanical evolution of the Michigan basin. *Tectonophysics*, **36**, 57.

Herzberg, C. T., 1983. Solidus and liquidus temperatures and mineralogies for anhydrous garnet lherzolite to 15 GPa. *Physics of the Earth and Planetary Interiors*, **32**, 193–202.

Jaupart, C., and Mareschal, J. C., 2003. Constraints on crustal heat production from heat flow data. In Rudnick, R. L. (ed.), *Treatise of Geochemistry*. Amsterdam: Elsevier Science. The Crust, Vol. 3, pp. 65–84.

Jaupart, C., and Mareschal, J. C., 2007. Heat flow and thermal structure of the lithosphere. In Watts, A. B. (ed.), *Treatise on Geophysics*. New York: Elsevier. The Lithosphere, Vol. 6, pp. 217–251.

Kaminski, E., and Jaupart, C., 2000. Lithosphere structure beneath the phanerozoic intracratonic basins of North America. *Earth and Planetary Science Letters*, **178**, 139–149.

Ketcham, R. A., 1996. Distribution of heat-producing elements in the upper and middle crust of southern and west central Arizona: evidence from the core complexes. *Journal of Geophysical Research*, **101**, 13611–13632.

Kinzler, R. J., and Grove, T. L., 1992. Primary magmas of mid-ocean ridge basalts 2. Applications. *Journal of Geophysical Research*, **97**, 6907–6926.

Kremenentsky, A. A., Milanovsky, S. Y., and Ovchinnikov, L. N., 1989. A heat generation model for the continental crust based on deep drilling in the Baltic Shield. *Tectonophysics*, **159**, 231–246.

Kukkonen, I. T., Jokinen, J., and Seipold, U., 1999. Temperature and pressure dependencies of thermal transport properties of rocks: implications for uncertainties in thermal lithosphere models and new laboratory measurements of high-grade rocks in the central Fennoscandian shield. In *Surveys in Geophysics*. The Netherlands: Kluwer, Vol. 20, pp. 33–59.

Landstrom, O., Larson, S. A., Lind, G., and Malmqvist, D., 1980. Geothermal investigations in the Bohus granite area in southwestern Sweden. *Tectonophysics*, **64**, 131–162.

Levy, F., Jaupart, C., Mareschal, J. C., Bienfait, G., and Limare, A., 2010. Low heat flux and large variations of lithospheric thickness in the Canadian Shield. *Journal of Geophysical Research*, **115**, B06404.

Mareschal, J. C., and Jaupart, C., 2004. Variations of surface heat flow and lithospheric thermal structure beneath the North American craton. *Earth and Planetary Science Letters*, **223**, 65–77.

Mareschal, J. C., Poirier, A., Rolandone, F., Bienfait, G., Gariépy, C., Lapointe, R., and Jaupart, C., 2000. Low mantle heat flow at the edge of the North American continent, Voisey Bay, Labrador. *Geophysical Research Letters*, **27**, 823–826.

Marton, F. C., Shankland, T. J., Rubie, D. C., and Xu, Y., 2005. Effects of variable thermal conductivity on the mineralogy of subducting slabs and implications for mechanisms of deep earthquakes. *Physics of the Earth and Planetary Interiors*, **149**, 53–64.

McKenzie, D., and Bickle, M. J., 1988. The volume and composition of melt generated by extension of the lithosphere. *Journal of Petrology*, **29**, 625–679.

Michaut, C., Jaupart, C., and Bell, D. R., 2007. Transient geotherms in Archean continental lithosphere: new constraints on thickness and heat production of the subcontinental lithospheric mantle. *Journal of Geophysical Research*, **112**, B04408.

Nicolaysen, L. O., Hart, R. J., and Gale, N. H., 1981. The Vredefort Radioelement Profile extended to supracrustal strata at Carletonville, with implications for continental heat flow. *Journal of Geophysical Research*, **86**, 10653–10661.

Nunn, J. A., and Sleep, N. H., 1984. Thermal contraction and flexure of intracratonal basins: a three-dimensional study of the Michigan basin. *Geophysical Journal of the Royal Astronomical Society*, **79**, 587.

Perry, H. K. C., Jaupart, C., Mareschal, J.-C., and Shapiro, N. M., 2006. Upper mantle velocity-temperature conversion and composition determined from seismic refraction and heat flow. *Journal of Geophysical Research*, **111**, B07301.

Pollack, H. N., 1986. Cratonization and thermal evolution of the mantle. *Earth and Planetary Science Letters*, **80**, 175–182.

Rudnick, R. L., and Gao, S., 2003. Composition of the continental crust. In Holland, H. D., and Turekian, K. K. (eds.), *Treatise on Geochemistry*. Oxford: Elsevier-Pergamon, Vol. 3, pp. 1–64 (ed. R.L. Rudnick).

Rudnick, R. L., and Nyblade, A. A., 1999. The thickness and heat production of archean lithosphere: constraints from xenolith thermobarometry and surface heat flow. In Fei, Y., Bertka, C. M., and Mysen, B. O. (eds.), *Mantle Petrology: Field Observations and High Pressure Experimentation: A Tribute to Francis R. (Joe) Boyd*. Houston: Geochemical Society, pp. 3–12.

Russell, J. K., and Kopylova, M. G., 1999. A steady-state conductive geotherm for the north central Slave, Canada: inversion of

petrological data from the Jericho kimberlite pipe. *Journal of Geophysical Research*, **104**, 7089–7101.
Solomatov, V. S., 1995. Scaling of temperature- and stress-dependent viscosity convection. *Physics of Fluids*, **7**, 266–274.
Vicker, P. A., 1997. *Garnet peridotite xenoliths from kimberlite near Kirkland Lake, Canada*. Unpublished MS thesis, University of Toronto, Toronto, p. 127.
Xu, Y., Shankland, T. J., Linhardt, S., Rubie, D. C., Langenhorst, F., and Klasinski, K., 2004. Thermal diffusivity and conductivity of olivine, wadsleyite and ringwoodite to 20 GPa and 1373 K. *Physics of the Earth and Planetary Interiors*, **143**, 321–336.

Cross-references

Deep Seismic Reflection and Refraction Profiling
Energy Budget of the Earth
Heat Flow Measurements, Continental
Heat Flow, Continental
Heat Flow, Seafloor: Methods and Observations
Lithosphere, Continental
Lithosphere, Mechanical Properties
Lithosphere, Oceanic: Thermal Structure
Mantle Convection
Radiogenic Heat Production of Rocks
Sedimentary Basins
Seismic Velocity-Temperature Relationships

LITHOSPHERE, MECHANICAL PROPERTIES

Evgueni Burov
ISTEP – Institute of Earth Sciences of Paris (UMR 7193 UPMC-CNRS), University of Pierre et Marie Curie, Paris, France

Synonyms

Geosphere

Definition

Lithosphere (*mechanical*). This is the rigid ("litho" = stone) outer layer of the Earth that remains mechanically strong over geological time spans. *Mechanical lithosphere* includes rigid layers of crust and outermost mantle capable to maintain high differential tectonic stresses, from 10 MPa to 1 GPa. *Mechanical lithosphere* is 1.5–2 times thinner than "seismic," "thermal," or "chemical" lithosphere.

Mechanical properties of the lithosphere. This term refers to the *integrated strength* of lithospheric plates, their *rheological structure* and parameters, and mechanical behavior in response to various tectonic loads.

Introduction

Mechanical properties of the lithosphere are of primary importance for local and global geodynamics. In particular, compared to the convective mantle, high long-term mechanical strength makes the lithosphere a unique stress/strain guiding and accumulating envelope with lasting mechanical memory. High strength prohibits internal *heat advection*, so *thermal conduction* is main heat transfer mechanism in the lithosphere, in contrast to the convective mantle. High strength also stabilizes vertical lithological structure of lithosphere making it a *stagnant layer*. In contrast to viscous mantle, long-term *rheology* of the lithosphere is strongly influenced not only by its *ductile* but equally *elastic* and *brittle* properties. It is probably the nonviscous properties of the lithosphere that shape it in the characteristic *plate tectonics* patterns.

The term *lithosphere* has been introduced in the second half of the nineteenth century, while the notion of *mechanical lithosphere* appeared in early twentieth century, in conjunction with that of *seismic lithosphere* (see *Earth's Structure, Global*), after formulation of the *continental drift theory* by Wegener and first interpretations of *regional isostasy* by J. Barrel and Vening-Meinesz (Watts, 2001; see entry *Isostasy*). The fact that the lithosphere has finite measurable strength has been demonstrated from observations and models of *regional isostatic compensation* of large topographic loads. Before that, the lithosphere was considered either as a very strong solid layer (*Pratt's model*) or, in turn, a weak fractured layer (*Airy's model*). *Postglacial rebound studies* of early twentieth century have contributed to the definition of the *mechanical lithosphere* as the uppermost layer of the solid earth characterized by slow *viscoelastic relaxation*, in contrast to the underlying, relatively low-viscosity asthenosphere. The long-term mechanical base of the lithosphere, h_m, is limited by the depth to isotherm 500–600°C in oceans and 700–800°C in continents, compared to almost twice as deep 1,330°C isotherm delimiting the thermal lithosphere (see entries *Lithosphere, Oceanic: Thermal Structure*; *Lithosphere, Continental: Thermal Structure*). As suggested on the basis of recent mantle–lithosphere interaction models (e.g., Schmeling et al., 2008), it is the elastic and plastic properties of the lithosphere that essentially determine the geometry of lithospheric plates and the mechanisms of formation of constructive, destructive, and transform plate boundaries at global scale. At smaller scale, *mechanical properties* of the lithosphere control formation and evolution of major geological structures such as rifts, passive margins, foreland basins, mountain ranges, plateau or strike-slip faults. They also control short-term processes such as seismicity (Watts and Burov, 2003).

Mechanical properties at different timescales

Mechanical properties of the lithosphere are timescale dependent (e.g., Watts, 2001). At seismic timescales (<0.5 h), the lithosphere behaves as an elastic or brittle-elastic layer of a thickness largely exceeding h_m. At *post-seismic timescales* (from days to years), the uppermost crust relaxes a part of seismically induced stress evoking some rather poorly understood short-timescale viscoelastic properties. At *postglacial rebound* timescales (~several 1,000 year), lithosphere remains largely elastic, with thickness of the elastic domain almost as important as its *seismic thickness*. At geodynamic timescales

(>1 Myr), the thickness of the *elastic core* is reduced to values smaller than h_m, and the stresses are slowly relaxed at timescales on the order of several Myr. The long-term properties of the lithosphere were first assessed from geodynamic-scale observations, first of all, *regional isostasy*. According to these data, the lithosphere exhibits a large spectrum of long-term behavior from quasi-elastic to brittle and viscous. Observation of crustal and lithosphere scale faults and *distributed seismicity* shows that some domains within the lithosphere deform in brittle-plastic regime over long time spans, while relaxation of deformation below, for example, oceanic volcanic islands, points to long-term viscoelastic strength. However, the long-term properties of the lithosphere are more generally studied indirectly using extrapolations from *rock mechanics data*.

Sources of information on the mechanical properties of the lithosphere

Flexural studies and *experimental rock mechanics data* are major sources of quantitative information on long-term behavior and strength of the lithosphere. In structured *viscous-elastic-plastic* media, all rheological properties are interrelated, and various time- and scale-dependent factors may cause variations in the effective elastic, brittle, and ductile deformation. Consequently, observations of long-term deformation are of primary importance for assessment of the effective mechanical properties of the lithosphere. These and other sources of information are summarized below:

1. Observations of long-term ($t > 1{,}000$ year) lithosphere response to tectonic loading, in the order of importance:
   - Observations of *regional isostatic compensation*: *gravity-flexural studies* providing estimates for the *equivalent elastic thickness* of the lithosphere, T_e
   - Vertical motions due to *postglacial rebound*, *lake, and volcanic island* loading
   - Observations of *lithosphere folding* (*folding wavelength* is a function of plate strength)
   - Field observations of ductile and brittle deformation in the outcrops
   - Interpretations of *deformational microstructures* and *paleo-stresses*
   - *Seismic tomography* (e.g., evidence for more or less strong slabs at depth)
   - *Borehole* stress measurements
2. Intermediate timescale observations ($t > 3$–10 year):
   - *Slow-rate rock mechanics experiments* (strain rates 10^{-9} to 10^{-4} s^{-1}, for ductile properties)
   - Geodetic (GPS-INSAR) data over >5 year time spans (strains, viscoelastic properties)
   - *Inter-seismic deformation*; *slow earthquake* data
3. Observations of deformation in response to short-term loading ($t < 1$ year):
   - Short-term *rock mechanics experiments* (elastic and brittle properties)
   - Distribution of intraplate seismicity (brittle properties)
   - Tidal deformation (viscoelastic properties)
   - Post-seismic relaxation (viscoelastic properties)
   - Geodetic (GPS-INSAR) data (strains, viscoelastic properties)
   - Attenuation of S waves (proxy to low-viscosity zones)
   - Magnetotelluric sounding (reduced electrical resistivity is proxy to low-viscosity zones)
4. Physical considerations and self-consistent thermo-mechanical models:
   - Estimates of the minimal *integrated strength* of the lithosphere required for lifetime stability of geological structures, *subduction* or transmission of tectonic stresses, and forces over large spatial scales, including horizontal pressure gradients caused by lateral variations in lithospheric density structure and topography (*gravity potential energy* theory). For example, lithosphere must be strong enough to transmit *ridge push* and *slab pull* forces on the order of 10^{11}–10^{13} N per unit length.
   - Lithosphere scale *numerical thermo-mechanical models* of tectonic processes integrating multidisciplinary data, which allows for testing the validity of data and hypotheses on lithosphere rheology.

Observations of long-term deformation provide key parameters such as the *integrated strength* of the lithosphere. These parameters are needed to constrain *rock mechanics data* obtained at laboratory conditions, because they are too far from geological conditions: short timescales ($t < 5$ years), small spatial scales ($l \sim 0.1$ m), high strain rates ($\dot{\varepsilon} > 10^{-9}\ s^{-1}$), small strains ($\varepsilon < 10$), high temperatures, simple deformation, largely mono-phase samples. *Rock mechanics data* allow only for assessment of general form of rheology laws, their sensitivities, and relative strengths of different kinds of rocks. Their extrapolation to geodynamic scales ($t > 10^6$ years, $l \sim 1{,}000$ km, $\dot{\varepsilon} < 10^{-14}\ s^{-1}$, $\varepsilon > 100$, cold temperatures, aggregate rocks) needs re-parameterization and validation using real-scale observations and models.

Interpretation of short-term data (seismic, satellite geodesy, INSAR) is not straightforward. In particular, interpretation of *intraplate seismicity* (see entry *Seismicity, Intraplate*) and *post-seismic relaxation* data is questioned due to the lack of evidence that mechanisms of this short-term deformation can be physically linked to those of long-term deformation.

Observations of flexural behavior and effective long-term strength of the lithosphere

Observations of *regional isostatic compensation* (e.g., Watts, 2001; see entry *Isostasy*) have shown that the lithosphere has substantial long-term *elastic rigidity* that allows for transferring and maintaining intraplate stress levels (10 MPa–1 GPa) over geodynamic time spans (>several Myrs). Studies of gravity anomalies observed

over mountain ranges and subduction zones have demonstrated that lithospheric plates bend like *thin elastic plates* of finite stiffness in response to tectonic, topography, and sedimentary loads. With improvement of geophysical measurement techniques during the second half of twentieth century, multiple studies (specifically, forward *flexural models*) have produced robust estimates of *flexural rigidity*, D, and *equivalent elastic thickness*, T_e, of lithospheric plates. These data arguably present a major source of information on the long-term *mechanical properties* of lithosphere. *Flexural studies* reveal strong variations of lithosphere strength, from near zero at ocean ridges to 100–150 km thick quasi-elastic cores detected within old and cold cratons (Burov and Diament, 1995; Kirby and Swain, 2009). In *flexural models* of *regional isostatic compensation* (Watts, 2001), D is varied until the model-predicted basement or *Moho topography* provides optimal fit to observations. *Gravity data* (see entry *Gravity Anomalies, Interpretation*) are used when basement topography is hidden (e.g., by sediments) or not representative of flexure (e.g., modified by erosion). The most robust *gravity models* are *forward models*. *Inverse models* are widespread due to the ability to cover large zones, but their results are more sensible to errors and should be cross-checked with *forward models*. For example, if not properly formulated, *gravity admittance* techniques may generate spurious results in continents (e.g., Kirby and Swain, 2009).

D provides a direct measure for the *integrated long-term strength* of the lithosphere and is linked to the *equivalent elastic thickness* of the lithosphere, $T_e : D = E\,T_e^3\left(12(1-\nu^2)\right)^{-1}$, where E and ν are Young's modulus and Poisson's ratio, respectively. *Plate bending or flexure* is characterized by its vertical deflection, $w(x)$ and local radius of curvature, $R_x(x)$ or curvature, $K(x) = -R_x^{-1} = \partial^2 w / \partial x^2$ (Figure 1). The flexural equation, when expressed using *bending moment* $M_x(x)$ is rheology independent and is valid for all, elastic and inelastic plates:

$$\frac{\partial^2}{\partial x^2}\overbrace{\left(\underbrace{\frac{E\,T_e^3}{12(1-\nu^2)}}_{D(x)}\underbrace{\frac{\partial^2 w(x)}{\partial x^2}}_{K(x)}\right)}^{M_x(x)} + \frac{\partial}{\partial x}\left(F_x\frac{\partial w(x)}{\partial x}\right) + \Delta\rho g w(x) = \rho_c g h(x) + p(x) \tag{1}$$

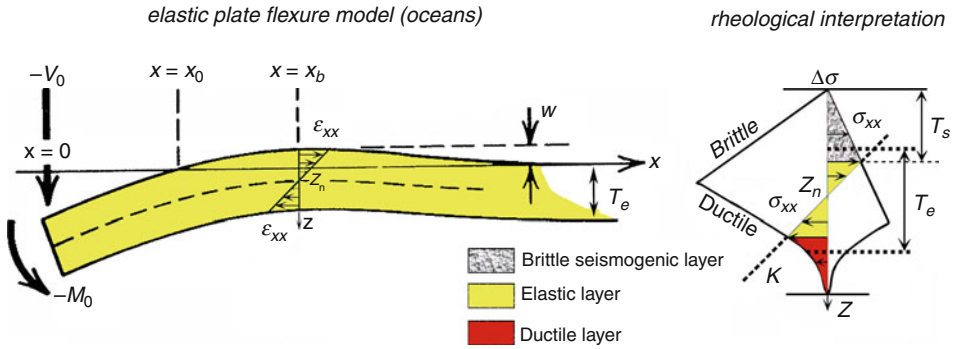
where F_x is horizontal *fiber force*, $\Delta\rho$ is the density contrast between surface and subsurface material (i.e., between mantle density ρ_m and topography/sediment/water density ρ_c), $h(x)$ is topography elevation, and $p(x)$ is additional surface or subsurface load. The elasticity is used just as simplest rheological interpretation of bending strength: elastic bending stress is linear function of curvature and depth: $\sigma_{xx}(x,z) \approx (0.5T_e - z)EK(1-\nu^2)^{-1}$. T_e is thus effective parameter and should not be automatically related to any real layer within the lithosphere. For inelastic plates, T_e and D have a sense of "condensed" plate strength and are direct proxies for the long-term *integrated strength*, B, of the lithosphere (Watts, 2001). For example, for a single-layer oceanic plate ($T_e = T_{e\_ocean}$, Figure 1):

$$B = \int_0^{h_m} \sigma^f(x,z,t,\dot{\varepsilon})dz$$

$$\text{while } T_{e\_ocean} = \left(12\left(\frac{\partial \sigma_{xx}^f}{\partial y}\right)^{-1}\overbrace{\left(\int_0^{h_m}\sigma_{xx}^f(z-Z_n)dz\right)}^{M_x(x)}\right)^{\frac{1}{3}};$$

$$T_{e\_ocean} < h_m \tag{2}$$

where $\sigma_{xx}^{\ f}$ is brittle-elastic-ductile bending stress (Burov and Diament, 1995); T_e is usually much smaller than



Lithosphere, Mechanical Properties, Figure 1 Classical flexural model of oceanic lithosphere (*left*). *Right*: brittle-elastic-ductile *yield stress envelope* (YSE) and interpretation of the *equivalent elastic thickness* T_e of the lithosphere. $\varepsilon_{xx}(z)$ is *flexural* strain, $\sigma_{xx}(z)$ is *flexural stress*, $K(x)$ is local plate curvature, $\Delta\sigma$ is *differential stress*, $Z_n(x)$ is *neutral plain*, and T_s is brittle seismogenic layer. $w(x)$ is vertical plate deflection. V_0 and M_0 are boundary cutting force and *flexural moment*, respectively.

h_m. M_x and, hence, D can be obtained from depth integration of σ_{xx}^f. D and T_e may spatially vary due to their dependence on local bending that leads to localized plate weakening (called *plastic or ductile hinging*) in the areas of utmost flexure, for example, near *subduction zones* or below mountains and islands.

Rheological properties of lithosphere according to rock mechanics data

The long-term mechanical behavior of rocks is represented by *extended Maxwell solid*, in which total strain increment equals a sum of elastic, viscous (ductile), and plastic (brittle) increments while the elastic, viscous, and plastic stresses are mutually equal. The weakest rheological term thus defines the effective behavior of this solid. Goetze and Evans (1979) used *Maxwell solid* and rock mechanics data to introduce the *yield stress envelope* (YSE) of the lithosphere (Figures 1 and 2). This approach consists in predicting, for a representative *background strain* rate and depth–pressure–temperature profile, the maximal *yield strength* $\Delta\sigma_{\max}(z)$ as function of depth, z. If elastic *differential stress* estimate $\Delta\sigma^e(z) < \Delta\sigma_{\max}(z)$, the deformation is elastic and *differential stress* $\Delta\sigma(z) = \Delta\sigma^e(z)$. If $\Delta\sigma^e(z) \geq \Delta\sigma_{\max}(z)$, then $\Delta\sigma(z) = \Delta\sigma_{\max}(z)$ and the deformation is brittle or ductile depending on z.

Elastic properties

The *elastic behavior* is described by linear Hooke's law:

$$\sigma_{ij} = \lambda\varepsilon_{ii}\delta_{ij} + 2G\varepsilon_{ij} \tag{3}$$

where λ and G are Lame's constants. Repeating indexes mean summation, δ is Kronecker's operator. For most rocks $\lambda \approx G = 30$ GPa (Turcotte and Schubert, 2002).

Brittle-plastic properties

Brittle resistance, τ, is a linear function of normal stress, σ_n and, that is, of pressure (Byerlee, 1978):

$$\tau = 0.85\sigma_n, \quad \sigma_n \leq 200 \text{ MPa} \tag{4}$$

$$\tau = 50 \text{ MPa} + 0.6\sigma_n, \quad 1,700 \text{ MPa} > \sigma_n > 200 \text{ MPa}$$

Byerlee's law is equivalent of *Mohr-Coulomb* (*Coulomb-Navier*) *plastic failure criterion* (e.g., Burov, 2007):

$$\tau = C_0 + \tan(\phi)\sigma_n \tag{5}$$

where C_0 is *cohesive strength* (<20 MPa) and ϕ is the *internal friction angle* (30–33°).

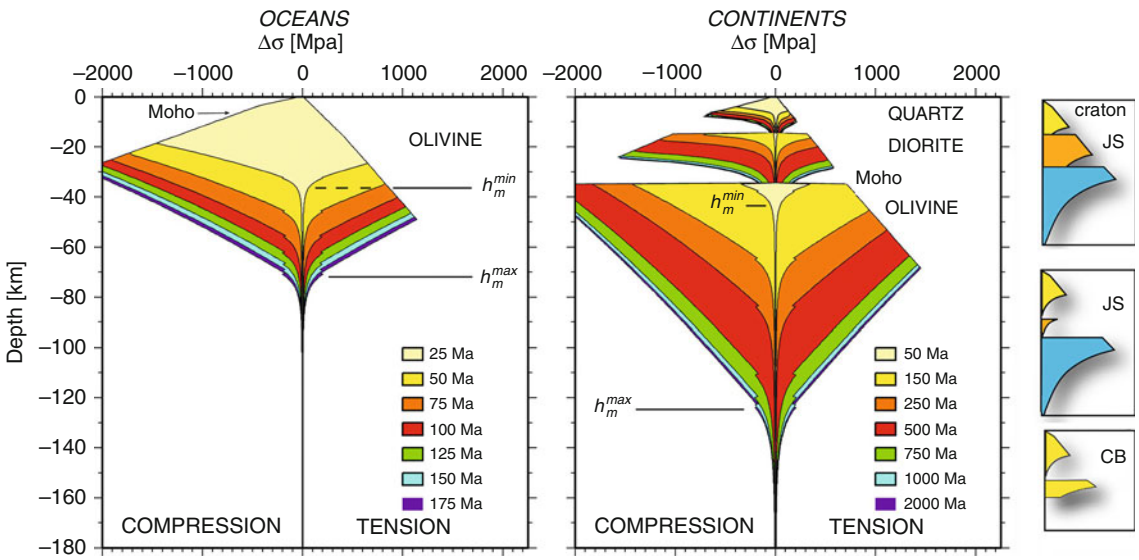
The Griffith criterion extends (5) to pure tensile domain:

$$\tau^2 = 4T_0^2 + 4T_0\sigma_n \tag{6}$$

where $T_0 = C_0/2$ is tension cut-off.

Viscous-ductile properties

Effective viscosity of the lithosphere varies from 10^{19} Pa s at the lithosphere–asthenosphere boundary to $>10^{26}$ Pa s in cold mantle near Moho depth. Several ductile mechanisms such as *diffusion creep*, *grain boundary sliding* (*GBS*), *pressure solution*, and *cataclastic flow* may play



Lithosphere, Mechanical Properties, Figure 2 Commonly inferred (brittle(Byerlee)-elastic-ductile) rheological *yield stress envelopes* (YSEs) as function of *thermotectonic age* for oceans and continents. In mantle, maximal strength can be limited by *Peierls* or *GBS* law instead of *Byerlee's law*. For continents, variations in crustal composition and fluid content result in various Jelly Sandwich, *Jelly Sandwich*, and more rare *Crème Brûlée*. After Burov (2007).

important role at appropriate conditions, but the leading part belongs to the *dislocation creep* (Kohlstedt et al., 1995):

$$\dot{\varepsilon}^d = Af_w\Delta\sigma^n \exp\left(-H(RT)^{-1}\right) \quad \text{for } \Delta\sigma < 200 \text{ MPa (Power law)}$$
$$\dot{\varepsilon}^d = Af_w \exp\left(-H\left(1-\Delta\sigma/\sigma_p\right)^2/RT\right) \quad \text{for } \Delta\sigma > 200 \text{ MPa (Harper}-\text{Dorn law)} \tag{7}$$

where $\dot{\varepsilon}^d$ is shear strain rate, A is material constant, n is power law constant, f_w is water fugacity factor, $\Delta\sigma$ is differential stress, R is universal gas constant, $H = Q + PV$ is creep activation enthalpy, Q is activation energy (100–600 kJ/mol), P is pressure, V is activation volume, T is temperature in K. σ_p is Peierls-like stress ($\sigma_p \sim$ several GPa). For tectonically relevant $\Delta\sigma/\sigma_p$ ratios (<0.1), Harper-Dorn's flow refers to *Peierls plasticity* that limits rock strength in high stress regime. *Dislocation creep* is strongly nonlinear non-Newtonian viscous flow ($n \sim 2$–4).

The second most important deformation mechanism, low-stress *diffusion creep* (Nabarro-Herring, Coble), results from directional diffusivity of rocks under applied stress (Kohlstedt et al., 1995):

$$\dot{\varepsilon}^d = Aa^{-m}f_w\Delta\sigma^n \exp\left(-H(RT)^{-1}\right) \tag{8}$$

where a is grain size and m is diffusion constant. For olivine, $m \sim 3$ and $n \sim 1$ and the constitutive law is linear Newtonian.

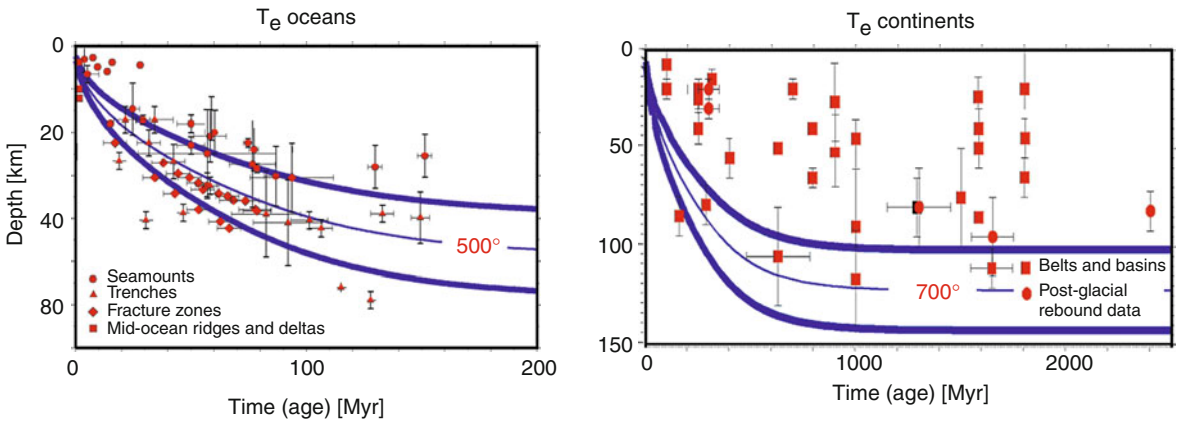
The *Peierls* super-plasticity is likely to replace *Byerlee's law* in the mantle below 30–40 km depth (Kameyama et al., 1999), that is, at high confining pressures (>1 GPa). Since *Peierls* plasticity includes strong water-weakening ($H = Q + p(V - \beta\Delta V_w)$), where ΔV_w is molar volume change due to incorporation of hydroxyl ions in the main rock and β is experimental parameter, it may play an important role in localization of deformation in subduction zones. This role may be shared with *GBS* creep, which might be responsible for aseismic localization of deformation in the lithosphere mantle.

Mechanical properties of oceanic lithosphere versus continental lithosphere

Due to its temperature dependence, ductile strength limits the depth to the base of mechanical lithosphere, h_m, to that of the isotherm 500–600°C in oceans and 700–800°C in continents (in continents, higher pressure for given temperature increases yield strength via the H term in Equation 7).

Oceanic YSEs ($0 < T_e <$ 50–70 km) derived from (3) to (8) predict brittle seismogenic behavior in the upper parts of lithosphere, where depth of *brittle-ductile transition*, BDT, varies from few km near ridges to 40 km near subduction zones (Figures 2 and 3). The competent domains below the *elastic core* are dominated by aseismic ductile creep, where important ductile strength is preserved down to the depths of 80–100 km near subduction zones.

Continental YSEs ($0–5 < T_e <$ 150 km) reflect strong *rheological stratification* between the upper, intermediate, lower crust and mantle lithosphere (Figures 2 and 3). There might be several BDT depths, typically at 15–25 and 30–45 km (Watts and Burov, 2003; Burov, 2007, 2010). The depth to the mechanical bottom of strong ductile lithosphere may vary from 30 to 200 km.



Lithosphere, Mechanical Properties, Figure 3 Observed T_e distribution in oceans and continents as function of *thermotectonic age* of the lithosphere, compared to depth of geotherms of 500 and 700°C defining, respectively, the mechanical base of oceanic and continental lithosphere (Burov and Diament, 1995; Watts, 2001; see also *Isostasy*). Assumed thermal thickness of continental lithosphere is 250 km. Depending on assumed thermal thickness and boundary conditions, predicted depth to 500–700°C may shift up or down by up to 25%. In continents, crust-mantle decoupling leads to ~50% T_e reduction compared to the depth to its mechanical base at 700°C. This explains occasional correlation of T_e with 400°C depth.

T_e estimates (Figure 3) reveal important difference between the mechanical properties of oceanic and continental lithosphere. In oceans, T_e (<70 km) follows the geotherm of 500 ± 100°C and correlates with age. It fits into the mechanically competent core (Figure 1).

In continents, maximal T_e values in cratons are 110–150 km, while T_e distribution reveals complex "bimodal" behavior (clustering around 25–30 and 70 km), which does not allow for geometrical interpretation of T_e. This complexity is explained by *rheological stratification* of continental plates, specifically of their 35–70 km thick crust (see entry *Lithosphere, Continental*). Continental plates consist of several layers of contrasting strength (upper, lower, intermediate crust, mantle) that can be mechanically coupled or uncoupled. For plates younger than 200–400 Ma, T_e grows with age yet exhibiting large scatter of values between 0 and 70 km. Older plates (>500–700 Ma) are in stationary thermal regime; the age dependence of T_e is expectedly small, but T_e scatter is still strong revealing influence of factors other than thermal age. Using rock mechanics data (Figure 4), it can be shown that continental T_e should be strongly controlled by the crustal thickness and mechanical state of crust-mantle interface (Burov and Diament, 1995). When the lower crust is mechanically strong, crust and mantle are mechanically coupled yielding a single strong layer with high T_e:

$$T_e \approx h_1 + h_2 \ldots = \sum_n h_i \tag{9}$$

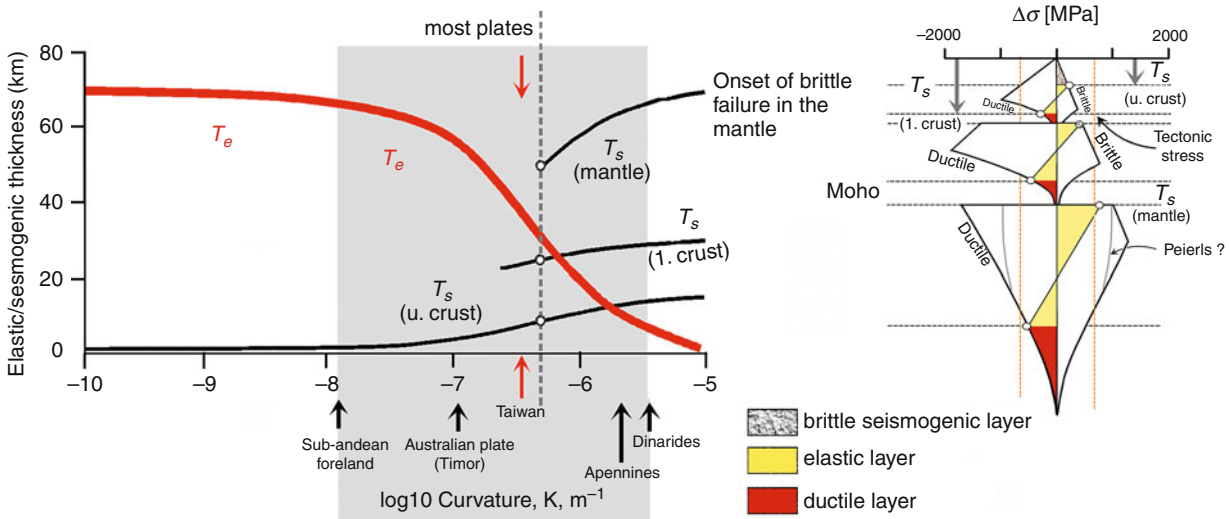
where h_i are thicknesses of crustal and mantle competent layers. Mechanical decoupling between crust and mantle leads to *structural weakening*, that is, dramatically smaller T_e (Burov and Diament, 1995):

$$T_e \approx \left(h_1^3 + h_2^3 \ldots\right)^{1/3} = \sqrt[3]{\sum_n h_i^3} \approx \max(h_i) < \sum_n h_i \tag{10}$$

"Decoupled" T_e roughly equals to the *mechanical thickness* of strongest layer in the lithosphere. In particular, $T_e \geq h_c$ (h_c is crustal thickness) identifies mantle as the strongest layer. Based on T_e data it has been shown that continental crust is systematically mechanically decoupled from mantle, except for cratons, and that mantle lithosphere is generally stronger or at least as strong as the crust (Burov and Diament, 1995). *Numerical thermomechanical models* have later confirmed these assertions (e.g., Burov, 2007, 2010).

Mechanical properties of the lithosphere and four modes of horizontal deformation

In addition to flexure under normal loads, four modes of lithosphere deformation and their combinations can develop under extensional or compressional conditions: (1) *pure shear* (extension: *McKenzie rifting model*; compression: plate *collision*), (2) *simple shear* (extension: *Wernicke rifting model*; compression: *subduction*), (3) *tensional/compressional instability* (extension: *boudinage*; compression: *folding-buckling*), and (4) *gravitationally instable pure shear* (slow *passive margins*, *Rayleigh-Taylor instabilities* due to thickening of a colder and thus denser lithosphere mantle overlying hotter and thus lighter asthenosphere). Ductile properties largely control these



Lithosphere, Mechanical Properties, Figure 4 Mechanical and rheological interpretation of continental T_e and *seismogenic layer thickness* T_s (*right*) computed for a representative YSE as function of local plate curvature K (*left*). After Burov (2010). Homogeneous horizontal stresses will shift T_e and T_s up or down in opposite directions, but will not change the character of their dependence on plate curvature.

deformation modes, including brittle fault spacing observed at surface (e.g., slow oceanic spreading centers vs. fast spreading centers). Brittle properties play an important role as well, specifically in localization of small and large-scale deformation, dyking, stabilization of cratons, transmission of tectonic stresses, and seismicity. Some of brittle mechanisms such as *low-angle faulting*, localization of large-scale *strike-slip* or *transform faults* and large-scale *dyke propagation* remain enigmatic.

Rheological stratification of continents is likely to influence the response of the lithosphere in all deformation modes. *Crust-mantle coupling* results in formation of narrow deep rifts while decoupling at different crustal levels may produce a variety of basins, from shallow large basins to narrow rifts. Under compression, *crust-mantle decoupling* plays a major role for continental subduction, UHP rock exhumation, slab detachment, formation of plateaus, or *lithosphere folding*. Lateral lower crustal flow is believed to be one of the key mechanisms affecting orogenic building and rift/passive margin formation or assuring feedback between surface and subsurface processes. For *mantle-continental lithosphere interactions*, such as *plume impingement*, the presence of ductile lower crust results in damping of the long-wavelength *dynamic topography* at surface, which is replaced by short-wavelength tectonic-scale deformation.

Uncertainties of data on the mechanical properties of the lithosphere

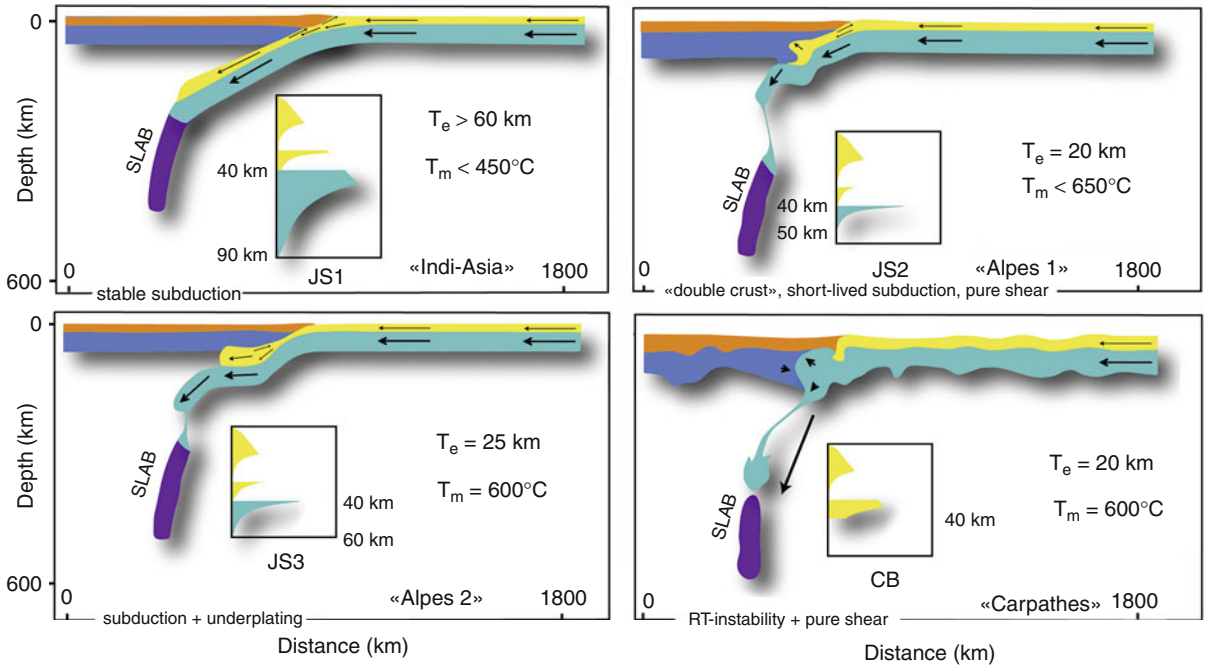
T_e studies play a key role in interpretation of *rock mechanics data*, allowing for parameterization of laboratory-based *flow laws* to geological time and spatial scales. Indeed, due to the large differences between time (10 orders) and spatial scales (6 orders) of the laboratory experiments and nature, *rock mechanics data* cannot be solely used to quantify the long-term *mechanical properties* of the lithosphere. Yet, since the *integrated strength B* of the lithosphere is directly related to T_e, one can use inelastic *flexural models* in combination with *rock mechanics data* (Equations 1–10) to constrain, parameterize, and validate rheology parameters for geological timescales. In particular, one can derive YSEs that are compatible with the host of different observations such as T_e, heat flow, seismic imagery, gravity, and so on. There is generally a good correlation between the observed T_e and T_e predicted from experimental YSE (Watts, 2001, Figure 4). In oceans, T_e values scale well with *rock mechanics data* and *thermal models of the lithosphere* (e.g., McNutt and Menard, 1982, Figure 3). In continents, T_e rheology parameterization is delicate but can be done if one respects the physics of underlying processes (Burov and Diament, 1995; Burov, 2010). Three types of *rheological envelopes* can be proposed for continental lithosphere: *Cratonic Jelly Sandwich*, *Jelly Sandwich*, *Crème brûlée* (with its variant hydrated *Banana Split* rheology) (Watts and Burov, 2003; Burov, 2010, Figure 2). The first two types evoke strong mantle lithosphere that can be either mechanically coupled with the crust (*Cratonic type*) or decoupled from it by weak lower or intermediate crustal layer (pure *Jelly Sandwich*). These first two rheologies dominate and correspond to 80–90% of all cases, as confirmed by matching T_e predicted from YSEs with the observed T_e and mechanical models. The *Crème brûlée* rheology infers weak lithosphere mantle, and applies to few cases of hot lithospheres (e.g., *Basins and Range Province*).

Mechanical properties of the lithosphere and styles of tectonic deformation

Important constraints on lithosphere rheology and properties come from physical considerations including *thermo-mechanical numerical models* of geodynamic processes. One can verify the validity of rheological assumptions by comparing predictions of thermo-mechanical models based on these assumptions with the observed tectonic evolution in various contexts such as collision and mountain belts (Figure 5), rifts and passive margins, and subduction zones and lithosphere folding. The analytical and *numerical thermo-mechanical models* of geological processes show that predicted evolution is highly sensitive to rheology choices (Figure 5). It has been demonstrated that in most cases mantle lithosphere has to be rheologically strong to insure stability of geological structures and the observed styles of deformation. In particular, stability of mountain ranges, continental *subduction* (e.g., Figure 5, Burov, 2010) and most observed *rifting modes* (Buck, 1991; Bassi, 1995) need initially strong mantle lithosphere with *mechanical thickness of the mantle part* more than 30–50 km. Collision models (Figure 5) quantify relationships between the collision styles and deep lithosphere structure and rheology. They show that tectonic evolution depends not only on the integrated strength (T_e) but as much on strength distribution with depth, that is, on which particular lithological layer (upper crust, lower crust, or mantle lithosphere) provides main contribution to T_e. Rift models show that *narrow rifts* are associated with initially strong mantle lithosphere, while metamorphic core complexes "need" warm lithospheres with negligible mantle strength. Wavelengths of *boudinage* and deca-kilometric upper crustal fault spacing observed in rift zones also appear to be a direct function of the ductile properties of the underlying lower crust. In compressional settings, observed *folding wavelengths* provide constraints on plate rheology since these wavelengths (50–1,000 km) are proportional to 5–10 times thickness of competent layers in the lithosphere and correlate with T_e values (e.g., Burov, 2007).

Links between different timescales: Burger's rheology model

The abundance of short-timescale observations such as *earthquake distributions* or *post-seismic relaxation* explains attempts to interpret these data in terms of long-term rheology. Yet, these attempts are typically based on



Lithosphere, Mechanical Properties, Figure 5 Numerical thermo-mechanical models of long-term geological processes (here, continental collision) allow for validation of inferred rheology profiles (*Jelly Sandwich*, JS1–JS3, and *Crème Brûlée*, CB) by testing the compatibility of model-predicted and observed tectonic evolution. Models JS2, JS3, and CB have similar T_e (20–25 km) but different mantle strength, leading to strong differences in resulting collision styles. The models show that strong mantle lithosphere is needed to drive continental subduction while strong crust cannot play same role (its low density prohibits subduction if there is no mantle drag). T_m – temperature at Moho. Colors: *orange*, *yellow* – crust; *blue*, *azure* – mantle; *purple* – oceanic slab. Modified from Burov (2010).

misinterpretation of Goetze's YSEs that are valid only for geodynamic strain rates. Earthquakes and relaxation occur at locally and temporarily high strain rates that are not representative of long-term rates. Hence, the fact that continental lithosphere below Moho depth is mainly aseismic cannot be interpreted as a sign of weak ductile behavior (Figure 4). It rather indicates that either mantle stress levels are insufficient to induce brittle sliding, or that seismogenic *Byerlee's law* does not operate at high pressure (>36–70 km depth), or that the frictional behavior is strain-rate dependent. At seismic strain rates (10^{16} times higher than geodynamic strain rates), the entire lithosphere acts as a brittle-elastic body and no ductile flow can occur. Strictly speaking, T_e and *maximal seismogenic depth* T_s should rather anti-correlate (Watts and Burov, 2003, Figure 4). However, T_s cannot be used to constrain T_e without knowing intraplate stress level. The fact that T_s distributions in oceans and continents are similar, while their rheological and mineralogical structures are different, adds to the argument that seismicity is primarily related to stress levels.

Post-seismic relaxation data provide controversial results yielding effective viscosities of deforming domains about 1–2 orders of magnitude smaller than *postglacial rebound data* that provide minimal viscosity (10^{19} to 5×10^{19} Pa s) of the Earth's weakest layer – asthenosphere. Since these estimates are based on inversion of *surface deformation*, they strongly depend on initial assumptions on lithosphere structure and properties, while it is impossible to determine which layer in the lithosphere relaxes post-seismic deformation. Nevertheless, whatever is this layer, the estimated viscosities appear too low for long-term properties. To explain this controversy, one can consider *Burger's model* of solids. According to this model, lithospheric behavior is described by two independent serially connected terms, one of which is *Kelvin solid* responsible for the primary creep (seismic, post-seismic) and the second one is *Maxwell solid* responsible for secondary long-term creep (geodynamic timescale). The viscosity of the first term is independent of that of the second term. In other words, the physics of deformation mechanisms activated at seismic-scale strain rates is different from that of the mechanisms acting at geodynamic strain rates.

Summary

Lithosphere has important long-term ductile, elastic, and brittle-plastic strength and is capable of maintaining (not relaxing) differential stresses at geological timescales. These properties can be accessed mainly from estimates of the *equivalent elastic thickness*, T_e, combined with *rock mechanics data*, validated by analytical and *numerical*

thermo-mechanical models of geological and geodynamic processes. In oceans, T_e fits in the mechanical core of the lithosphere, but in continents, it generally does not represent any particular layer due to strong *rheological stratification*. The brittle-plastic properties of the lithosphere are governed by *Byerlee's law* to 30–40 km depth and likely by *Peierls* plasticity and/or *GBS* creep below. The ductile properties are dominated by *dislocation power-law creep*, which is strongly nonlinear and rock-type dependent. Highest *integrated strength* ($T_e \sim$ 100–150 km) is detected in cratons (*Cratonic Jelly Sandwich rheology*) where strong crust is mechanically coupled with strong mantle. In warmer continental lithospheres, mechanical decoupling between crust and mantle leads to *structural weakening* resulting in ~50% reduction of T_e (*Jelly Sandwich rheology*). In most cases, mantle lithosphere provides main contribution to the integrated plate strength. The cases where mantle is weaker than crust (*Crème-Brûlée rheology*) refer to hot lithospheres such as metamorphic core complexes. There is probably no exploitable link between short-term deformation such as earthquake data or post-seismic relaxation and long-term properties of the lithosphere.

Bibliography

Bassi, G., 1995. Relative importance of strain rate and rheology for the mode of continental extension. *Geophysical Journal International*, **122**, 195–210.

Buck, W. R., 1991. Modes of continental extension. *Journal of Geophysical Research*, **96**, 20161–20178.

Burov, E. B., 2007. Plate rheology and mechanics. In Schubert, G. (ed.), *Treatise on Geophysics,* Vol. 6: *Crust and Lithosphere Dynamics,* **99**(152), 611 pp. (Vol. ed.: Watts, A.B.). Elsevier: Amsterdam. TOGP00102. ISBN:978-0-444-51928-3.

Burov, E., 2010. The equivalent elastic thickness (T_e), seismicity and the long-term rheology of continental lithosphere: time to burn-out "crème brûlée"? Insights from large-scale geodynamic modeling. *Tectonophysics*, **484**, 4–26.

Burov, E. B., and Diament, M., 1995. The effective elastic thickness (T_e) of continental lithosphere: what does it really mean? *Journal of Geophysical Research*, **100**, 3895–3904.

Byerlee, J. D., 1978. Friction of rocks. *Pure and Applied Geophysics*, **116**, 615–626.

Goetze, C., and Evans, B., 1979. Stress and temperature in the bending lithosphere as constrained by experimental rock mechanics. *Geophysical Journal of the Royal Astronomical Society*, **59**, 463–478.

Kameyama, M., Yuen, D., and Karato, S.-I., 1999. Thermal-mechanical effects of low-temperature plasticity (the Peierls mechanism) on the deformation of a viscoelastic shear zone. *Earth and Planetary Science Letters*, **168**, 159–172.

Kirby, J. F., and Swain, C. J., 2009. A reassessment of spectral T_e estimation in continental interiors: the case of North America. *Journal of Geophysical Research*, **114**, B08401, doi:10.1029/2009JB006356.

Kohlstedt, D. L., Evans, B., and Mackwell, S. J., 1995. Strength of the lithosphere: constraints imposed by laboratory experiments. *Journal of Geophysical Research*, **100**, 17587–17602.

McNutt, M. K., and Menard, H. W., 1982. Constraints on yield strength in the oceanic lithosphere derived from observations of flexure. *Geophysical Journal of the Royal Astronomical Society*, **71**, 363–395.

Schmeling, H., Babeyko, A. Y., Enns, A., Faccenna, C., Funiciello, F., Gerya, T., Golabek, C. J., Grigull, S., Kaus, B. J. P., Morra, G., Schmalholz, S. M., and van Hunen, J., 2008. A benchmark comparison of spontaneous subduction models – toward a free surface. *Physics of the Earth and Planetary Interiors*, **171**, 198–223.

Turcotte, D. L., and Schubert, G., 2002. *Geodynamics*. Cambridge: Cambridge University Press, 456 pp.

Watts, A. B., 2001. *Isostasy and Flexure of the Lithosphere*. Cambridge: Cambridge University Press, 458 pp.

Watts, A. B., and Burov, E., 2003. Lithospheric strength and its relationship to the elastic and seismogenic layer thickness. *Earth and Planetary Science Letters*, **213**, 113–131.

Cross-references

Earth's Structure, Global
Gravity Anomalies, Interpretation
Isostasy
Lithosphere, Continental
Lithosphere, Continental: Thermal Structure
Lithosphere, Oceanic: Thermal Structure
Seismicity, Intraplate

LITHOSPHERE, OCEANIC

James McClain
Department of Geology, University of California, Davis, CA, USA

Definition and introduction

The term "lithosphere" is used or defined in a variety of formal and informal ways. All are fundamentally linked to a comparison of near-surface (lithosphere) mechanical and/or thermal properties with those properties deeper in the Earth (the asthenosphere). And all are linked to the concept of a thermal boundary layer between the bulk of the Earth and its atmosphere/ocean above. In this entry, we deal with the oceanic lithosphere, and we will define it as that part of the oceanic crust and the uppermost mantle that is usually created at mid-ocean spreading centers and that generally is consumed in subduction zones. For space purposes, we will specifically exclude oceanic sediments, island or continental arcs, and continental margins.

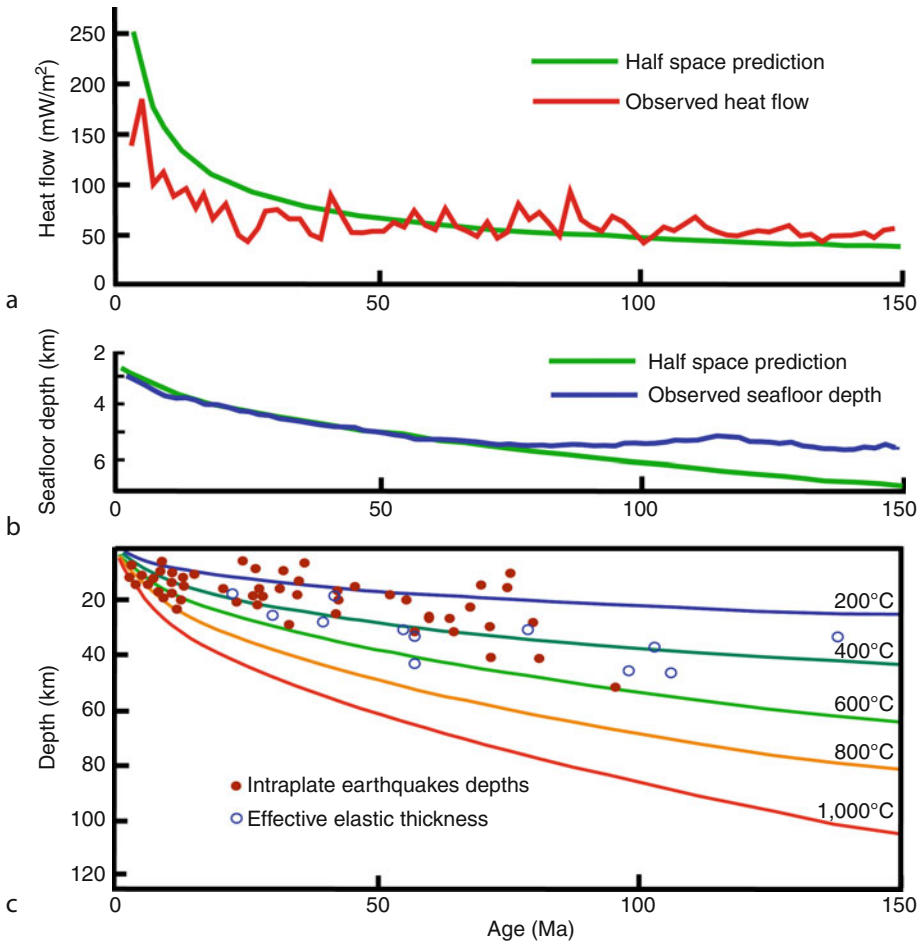
The ocean lithosphere is generally considered the outer carapace of the Earth that is cool, with an elastic rheology, and exhibiting brittle failure (e.g., failure by earthquake). This contrasts with the asthenosphere, which exhibits viscous rheology and higher temperatures. The concept of lithosphere was first developed in 1914 by Barrell (1914), who realized that the outer part of the Earth must be able to support tectonic loads over geological timescales. Since that time, workers additionally have defined the lithosphere in terms of plates that move coherently in plate tectonics. Alternatively, the oceanic lithosphere may be defined as the relatively high velocity or low attenuation "seismic lid," or as that part of the earth that can sustain earthquakes. These various definitions are not

necessarily mutually exclusive, but it is important to consider how the term lithosphere is used (Figure 1).

The uppermost mantle lithosphere

The mantle component of the lithosphere is composed of peridotite, and the thermomechanical transition from lithosphere to asthenosphere is defined on the basis of the data used to detect that transition. Many have used the seismic low-velocity zone often inferred for the upper mantle as the essential (e.g., Forsyth, 1977; Nishimura and Forsyth, 1989). Typically exploiting surface wave observations, these studies have been very useful in defining variations in lithospheric thickness.

Another approach to estimating lithospheric thickness is derived from observations of the response of the lithosphere to tectonic loading (e.g., Calmant et al., 1990). Specifically, the addition of a large seamount or island should cause the lithosphere (including the seafloor) to deflect downward as the load is accommodated by the asthenosphere. The amount of deflection or flexure, and the regional distribution of the deflection, can be considered a function of the thickness of the strong lithospheric plate. The topography of the seafloor around a seamount can be used to calculate this "effective thickness." A compilation of thicknesses is shown in Figure 1c, and reveals the thickness of the lithosphere coincides with isotherms between 400°C and 600°C in the mantle. Similar values for lithospheric thickness are obtained if one uses the depths of intraplate earthquakes, which presumably are indicators of brittle failure confined to the lithosphere (Wiens and Stein, 1983) see (Figure 1c).



Lithosphere, Oceanic, Figure 1 Different observables are used to constrain models for the oceanic lithosphere, primarily its mantle component. (**a**) Observed heat flow data (in *red*) averaged in 2-million-year bins (Stein and Stein, 1992) compared to predictions for a conductively cooled half space (in *green*). Heat flow is too low and scattered for ages less about 40 Ma, indicating nonconductive (hydrothermal circulation) cooling must be taking place. (**b**) Observed seafloor depth (in *blue*) compared to predictions for a conductively cooled half space (in *green*). Depths are too shallow beyond 80 Ma, indicating temperatures have stabilized and heat is being supplied from below. (**c**) Plot of earthquakes depths and effective elastic thickness on top of expected isotherms for a half-space model.

Regardless of the thermomechanical observable used, the lithosphere is observed to thicken with age as it spreads away from the mid-ocean ridges, a result of cooling of the mantle. Independent evidence for this cooling is the progressive deepening of the oceans with age, and the coincident decrease in heat flow. Simple, conductively cooled, half-space models (e.g., Davis and Lister, 1974) yield a prediction that depth must increase linearly with the square root of age ($t^{1/2}$), and heat flow must decrease proportionally with the one over the square root of age ($t^{-1/2}$) (see Figure 1a and b, modified from Stein and Stein, 1992).

Even the earliest studies pointed out that heat flow and depth data did not perfectly fit the half-space model. Parsons and Sclater (1977) and Stein and Stein (1992) made detailed analysis of the global compilations of depth and heat flow measurements. They document that, at great ages (e.g., >80 Ma), the seafloor is too shallow, and heat flow values are too high compared to predictions (Figure 1a and b). They point out that a plate model, where heating from below restricts continued cooling of the plate would explain the discrepancy. The source of this heating could be small-scale convection in the asthenosphere below the plate, radioactive decay in the mantle, or mantle plumes. Parsons and Slater derived a plate (i.e., lithosphere) thickness of 125 km at great ages. Using a larger data set and more refined analysis, Stein and Stein (1992) achieved a better fit for a lithosphere that thickens to 95 km. We note these thicknesses are substantially greater than those that would be derived from the earthquake depths and flexural loading; this illustrates the fact that the characteristics determined for the lithosphere are obervation-dependent.

The second discrepancy of global data with the half-space model was the fact that heat flow values were highly scattered and substantially lower for seafloor less than 50 million years old. Using their conductively cooled plate model, and the misfit between the predicted and measured heat flow, Stein and Stein (1994) calculate that some 34% of the oceans heat loss is accommodated by hydrothermal circulation. They further conclude that about 70% of the advective heat loss occurs in crust from 1 to 65 million years, and not through vigorous hydrothermal activity occurring at the mid-ocean ridges.

The crustal component of the oceanic lithosphere

The top few kilometers of lithosphere comprise the ocean crust. To a first approximation, the crust is viewed as the mafic member of the lithosphere, and most is generated by magmatic, tectonic, and hydrothermal processes at the mid-ocean ridges, and modified by the ongoing hydrothermal circulation noted above. Our knowledge of the crust is dominated by three lines of evidence, all of which provide key constraints on the nature of oceanic lithosphere. These include direct sampling of seafloor rocks where they are exposed or from boreholes, seismic experiments that reveal the seismic velocity structure of the crust, and the study of crustal analogues exposed above sea level. These studies led to a convergence on the so-called Penrose or "ophiolite" model (defined below). While this model is highly successful in its predictions about the crust and upper mantle, the research in the last 10 years has led to recognition, at least for some crust generated at ridges with slow to ultraslow spreading rates, that the ophiolite models must be revised to include newly discovered features.

Samples of crustal rocks

Ocean crustal rocks are generally recovered from near the mid-ocean ridges before sediments have a chance to accumulate on the aging seafloor. Samples are collected using dredges and other techniques from surface ships, as well as exploiting manned submersibles and remotely operated vehicles (ROVs).

The vast majority of rocks recovered from the seafloor are mafic basalts, although other types, such as gabbros and serpentinites, are sometimes exposed at transform faults or on rifted mid-ocean ridges. However, it is the striking uniformity of seafloor basalts that reveal the similarity processes that form the mantle and crustal components of the lithosphere.

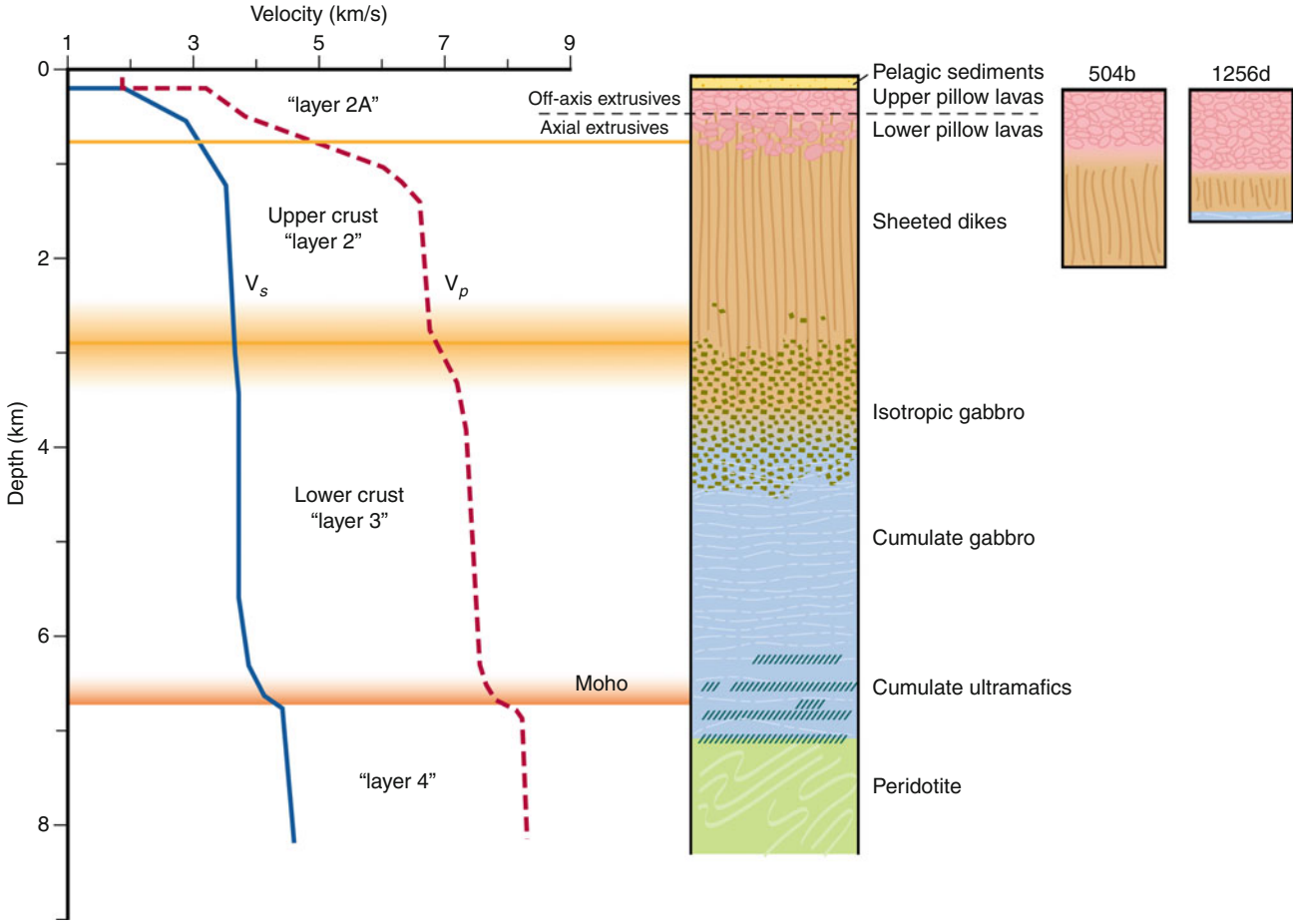
Most of the mafic rocks recovered on the mid-ocean ridges are so-called MORBs or mid-ocean ridge basalts. The composition of MORBs indicates that the mid-ocean ridge basalts originate from partial melting of a depleted upper mantle beneath the ridges. MORBs have been further parsed into N-MORB (normal-MORB) and E-MORB (enriched-MORB), and intermediate compositions (transitional or T-MORB) have been observed. These observations generally are interpreted to indicate they originated in a heterogeneous, but still depleted mantle source. It is this depleted mantle that will, upon cooling with age, form the mantle lithosphere.

An additional inference made from analysis of seafloor basalts was that the detailed composition required partial fractionation at shallow (i.e., lower crustal) pressures; that is, a magma chamber exists beneath most ridge axes.

Ophiolites, ocean drilling, and the ocean crust

Ophiolite sequences have been studied for over 100 years, and have been the source of substantial controversy. Complete ophiolites see (Figure 2) include a coherent stratigraphy extending from submarine lava flows into sheeted dikes, the latter representing the feeders for the lava flows. Beneath the sheeted dikes are gabbros, often divided between isotropic gabbros (no preferred fabric) at the top, and layered (cumulate) gabbros at the base. Gabbros represent the frozen magma chamber, originally inferred from the composition of MORB. Beneath the gabbros are ultramafic peridotites (inferred to be the mantle), varying between harzburgites, lherzolites, and dunites. In general, the peridotites are serpentinized to a greater or lesser extent, but is unclear whether the serpentinization occurs during emplacement and uplift onto land, or on the seafloor.

Early on, ophiolites were thought to originate as ocean crust and uppermost mantle, and to have been thrust up on land by convergent processes (see Moores, 1982). Substantial debate centered on whether particular ophiolites



Lithosphere, Oceanic, Figure 2 Comparison of seismic velocities with an idealized ophiolite, and with drilling results for two Pacific ODP (504b) and IODP (1256d) drill holes that successfully penetrated to mid-crustal levels (Modified from McClain, 2003). The seismic structure is for a specific profile on 16-Ma crust generated on the fast-spreading southern East Pacific Rise. The mid-crustal inflection is used here as the boundary between Layer 2 and Layer 3 for this particular crust. The inflection could be caused by the lithological boundary between sheeted dikes and gabbros as shown here, or by changes in porosity, or by changes in metamorphic grade. From velocity alone, we cannot tell. However, all three causes could occur at the same depth.

originated in arc or backarc settings, or whether some may have originated on mid-ocean ridges. Regardless of the origin of any particular ophiolite, it is clear that ophiolites provide an important analog for ocean crust generated on mid-ocean ridges. The presence of sheeted dikes to deliver magma from a crustal magma chamber, and the presence of gabbros originating from the freezing of the chamber, leads to a model that is consistent with petrologic models that originate from seafloor sampling.

The Ocean Drilling Project, and its predecessor Deep Sea Drilling Project and successor the Integrated Ocean Drilling Project, also provide support for the "ophiolite" model. In particular, drilling results in the Costa Rican rift (Hole 504b) and off southern Mexico (Hole 1257) have penetrated through the seafloor lavas and into the sheeted dikes. For the latter site, gabbros, intercalated with sheeted dikes have been reached at the base of the sheeted dike column (see Figure 2).

Seismic studies of the ocean crust

Except for studies of subareal ophiolites, rare exposures on the seafloor, and a few deep drilling efforts, seismic studies provide the only information about the ocean crust below the seafloor. The information they reveal about the structure and seismic velocities in the subsurface leads to inferences about the petrology of the upper mantle and crust.

After World War II, marine institutions began to exploit the seismic refraction technique to explore the Earth beneath the sea. These experiments usually yielded compressional or P-wave seismic velocities, although converted shear waves were sometimes observed. In these early studies, theoretical and computational tools necessarily restricted models to laterally homogeneous structures, and a consistent four-layer model for the sediments, crust, and mantle was developed (e.g., Shor et al., 1969). The first layer was that of sediments, with

velocities averaging about 2 km/s. Beneath that was "Layer 2," and because the seafloor was known to reveal lavas in unsedimented areas, it was expected that Layer 2 was the "volcanic layer." The thickest part of the oceanic crust is "Layer 3," with velocities of about 6.8–7.2 km/s and thicknesses on the order of 2–4 km. The lithology of Layer 3, also known as the "oceanic layer," has been controversial, with some workers preferring a gabbroic composition, and others a serpentinite composition. The former model has gained general acceptance, partly because of the presence of gabbros in ophiolites (e.g., Christensen and Salisbury, 1975). Underlying Layer 3 is Layer 4, with compressional wave velocities of about 8 km/s, and corresponds with the upper mantle. The lithology of the mantle is generally taken to be peridotite, consistent with the above-mentioned models for partial melting. Thus, the "crust" comprises Layers 2 and 3, and the boundary between the crust and mantle, Layers 3 and 4, is known as the Moho or Mohorovicic discontinuity.

With better experiments, the original four-layered model was subdivided into more finely divided structures. The earliest refinement was the observation of "Layer 2A." This was the shallowest and lowest velocity portion of the upper crust, and the low seismic velocities were attributed to the high porosities of basalt flows. This interpretation is further supported by the observation that the velocities of Layer 2A tend to increase to normal Layer 2 values and/or that Layer 2A thins over a few million years. This was interpreted as infilling of pores by metasomatic precipitation of minerals as the crust ages (e.g., Houtz and Ewing, 1976). Such an explanation requires fluid circulation in the crust must continue over millions of years after the crust forms, consistent with recent models for the cooling of the lithosphere mentioned above (Stein and Stein, 1994).

In the 1970s, the development of new theoretical and computation tools allowed workers to move beyond the simple layered models, to those that with continuously varying, but usually vertically stratified models. As models grew more sophisticated, it became possible to correlate them with ophiolite structures, using both velocities and thicknesses (e.g., Christensen and Salisbury, 1975; Spudich and Orcutt, 1980; McClain, 2003). As a result, the so-called Penrose model or "ophiolite model" for the oceanic lithosphere has become the accepted model (Figure 2). For the crust, this led to models where Layer 2 was divided into an upper zone in with relatively low velocities (Layer 2A), underlain by a high velocity gradient extending to Layer 3. The high-gradient zone, sometimes labeled Layer 2B, was attributed to a decrease in crustal porosity, and is correlated with the transition to sheeted dikes of the ophiolite model.

The same experiments revealed a slight inflection in seismic velocities at mid crustal levels, and often seismic velocities at mid crustal levels only reach "Layer 3" values beneath that inflection see (Figure 2). This inflection is not always observed, because it is obscured in the data, or because of the analysis techniques used, or because the change in velocity is too gradual. In the latter case, Layer 3 is often interpreted to be at the base of the high-gradient zone. The origin of this slight transition was (and is) attributed either to a metamorphic boundary between lower and higher grade metamorphic changes in lithology, a reduction in residual porosity, or to a transition from diabase dikes to the gabbros in the lower crust. For the drilling results for hole 504B (Figure 2), the velocity boundary was imaged seismically, and then successfully drilled within the sheeted dikes. In that case, it appears the boundary corresponds to a change in metamorphic grade and/or a change in porosity. These explanations are not necessarily mutually exclusive; the penetration of metasomatic fluids requires some porosity, and gabbros in ophiolites typically reveal little or no alteration or relict porosity in comparison to the overlying sheeted dikes.

Finally, the Moho is more gradational in these refined models, and the mantle velocities are well correlated with those expected for peridotites. While the velocity difference between a gabbro and peridotite is large, the apparent, more gradual observed change is most likely because the transition is gradual, with laterally heterogeneous and interleaved gabbros and peridodtites at the Moho.

The mid-ocean ridges

The ocean crust is generated by a complex interaction of magmatic, tectonic, and hydrothermal processes that take place at the mid-ocean ridges where seafloor spreading occurs. As a result, the majority of seismic experiments over the last 2–3 decades have focused on the tectonic plate boundaries, with particular attention on the ridges. Interpretation of the data from these experiments required models that allowed for two- or three-dimensional structure, rather than the laterally homogenous models from earlier. Early ray-tracing techniques have given way to more robust tomographic inversion. In addition, the development of longer seismic hydrophone streamers and new processing algorithms now permits detection of subbasement (crustal and upper mantle) reflections using multichannel reflection experiments.

One of the critical controversies about the structures beneath oceanic spreading centers was the presence of a seismically resolvable axial magma chambers (AMC). As stated above, petrologic data seemed to require some sort of magma chamber in the lower crust, and the ophiolite model includes a thick gabbro layer. However, models resulting from seismic surveys were equivocal, with some revealing crustal AMC (e.g., Orcutt et al., 1975) and some not (e.g., Bratt and Solomon, 1984). McClain et al. (1985) were able to provide convincing evidence for a narrow magma chamber on the East Pacific Rise at 13° north. The seismic velocity required for the chamber was 4.5 km/s, suggesting a partial melt rather than a pure magma. Further evidence for the resolved AMC was found in an extensive tomographic experiment on the East Pacific Rise at 9°30′N (Toomey et al., 1990).

For fast-spreading ridges, the question of magma chamber presence was largely resolved when a large

multichannel experiment was conducted along the northern East Pacific Rise (e.g., Detrick et al., 1987). Using the long hydrophone streamers, researchers were able to resolve reflections from the base of Layer 2A, as well as the Moho. However, most striking was the presence of a reflection at mid-crustal levels along most of the length of the ridge. Analysis of the reflection amplitude and the change in the sign of the reflection requires that the reflection result from a sudden decrease in seismic velocity. The explanation for this velocity inversion was the presence of an axial magma chamber. Furthermore, the large amplitudes of the reflections require a rather low velocity compared to models obtained from refraction experiments. This suggests that reflections result from pure or nearly pure melt sills. The discrepancy results because refraction experiments do not resolve the sills, and instead image an average structure made up of the sills and partial melt zones. Early models have suggested these sills form at the top of the magma chamber, and may be responsible for the isotropic gabbros observed in some ophiolites. The larger partial melt zones would form the cumulate gabbros that make up the bulk of the lower crust.

Subsequent to the initial experiment of Detrick and others (1987), similar magma reflections have been imaged on other segments of the fast or ultrafast spreading East Pacific Rise (e.g., Detrick et al., 1993; Singh et al., 2006) as well as ridges with intermediate spreading rates (e.g., Canales et al., 2009). Most studies suggest that the crustal AMC is generally narrow, the result of vigorous hydrothermal cooling. However, recent studies also have revealed that melt sills are sometimes distributed in depth and laterally.

Hydrothermal activity

One of the most spectacular discoveries in the annals of earth science was the observation of vigorous hydrothermal circulation at the spreading centers (e.g., Corliss et al., 1979). While the presence of hydrothermal circulation was not surprising, the exceptionally high temperatures, the deposition of hydrothermal minerals, and a vigorous chemosynthetic biological community were a major breakthrough in the study of mid-ocean ridges. It is clear that the chemical, biological, and geological contribution of these vents is present throughout the world's oceans. Thus far, over 200 hydrothermal vent fields have been discovered, and as many as 1,000 may lie on the world's mid-ocean ridges (Baker and German, 2004).

Early studies suggested that a substantial portion of the Earth's heat was lost along the high-temperature hydrothermal circulation on mid-ocean ridges. More recent analyses indicate that more heat is lost through lower temperature circulation in older crust. However, it is clear that near-axis circulation is a major component of the crust-forming process, and their discovery has led to a new appreciation for the interdisciplinary nature of mid-ocean ridge science.

High-temperature venting is observed on all ridges, and at all spreading rates. Such systems require a heat source and a permeability structure that permits a focused flow of fluids to the surface. For fast-spreading ridges, heat sources are plentiful, with the presence of axial magma chambers, and frequent eruptive events. Indeed, the narrowness of the axial magma chamber requires that circulation penetrate at least to the top of the chamber and/or into the gabbros at its flanks (e.g., Lister, 1983; Morton and Sleep, 1985). The geometry of the down-going limb of the seawater into the crust remains uncertain. It may be that it penetrates along axis, perhaps along faults or fissures, or it may penetrate from the sides along normal faults.

For slower spreading ridges, the heat source is unclear. Certainly volcanic eruptions do occur, so a heat source is present at least some of the time. However, it may be that water also can penetrate to the deeper hot, but not molten, rocks. The large normal faults often present on slow-spreading ridges may accommodate this deeper circulation.

Along-axis variations in ridge processes and slow-spreading ridges

It has long been known that mid-ocean ridges display a systematic variation in cross-sectional topography, varying from rifted ridges, with rift valleys of 1–2-km deep, and to non-rifted ridges that display topographic highs. In general, rifting is associated with slower spreading ridges, while non-rifted ridges usually exhibit intermediate to rapid seafloor spreading. There are a number of exceptions to the rule. For the Reykjanes Ridge adjacent to the Iceland Hot Spot, the slow-spreading ridge is unrifted. On the Galapagos Ridge, the ridge segments closest to the hot spot there is also unrifted. For ridges spreading at intermediate rates, both rifted and unrifted ridge segments are observed (e.g., the Gorda and southern Juan de Fuca Ridges), and are often separated by only a transform fault.

These observations suggest that the presence or absence of a rift valley is more likely controlled by magma supply rather than spreading rate itself. On slow-spreading ridges, the isotherms are narrowly spaced, and plunge more steeply as the lithosphere ages. Thus, less melt may reach the surface unless excess temperatures and magma are supplied by a nearby mantle plume. For intermediate spreading ridges, local variations in magma supply, and perhaps along-axis flow may influence the formation of a rift valley.

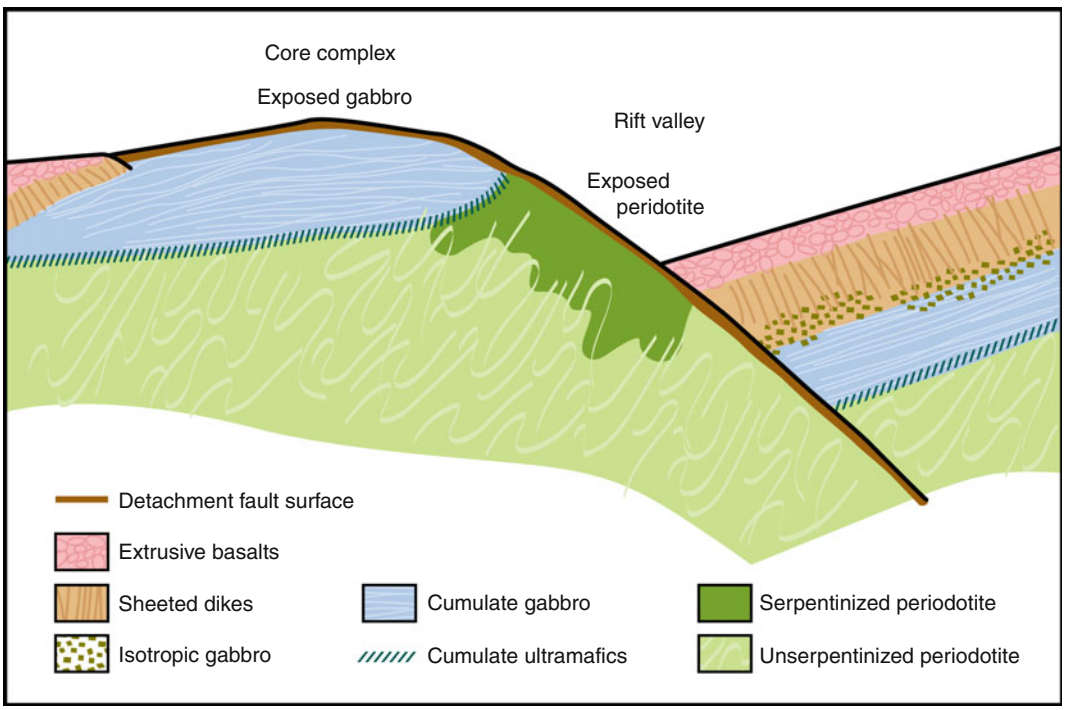
With the advent of multi-beam systems, it was found that ridges are segmented on a variety of scales, with offsets between segments also varying in character and length. In a series of papers, a correlation between segment depths and cross-sectional geometry was pointed out (e.g., Macdonald and Fox, 1988), and it was suggested that these variations were linked to focused magma supply at segment centers and subsequent magma flow along axis to the

segment ends. It is logical to assume that the magma supply would be tied to the resulting crustal structure and thickness, with thicker crust at the segment centers. The seismic reflection results of Detrick et al. (1987) seemed to support that linkage. However, the inferred correlation of ridge topography with crustal thickness has been called into question for fast-spreading ridges by a series of seismic experiments (e.g., Barth and Mutter, 1996; Kent et al., 2000).

Chen (1992) made a global study of crustal thicknesses, and found very little correlation or variation in thickness for spreading rates greater than 30 mm/year, and it appears that segmentation does not have a major influence on crustal thickness at these spreading rates. However, for ridges that spread slowly, or very slowly, crustal thicknesses are far more variable.

With the improvement of techniques, mid-ocean ridge researchers have been able to overcome the complexities caused by topography on rifted ridges. For such ridges, crustal thickness is closely tied to location along a segment, and it is likely that magma and mafic crust production is enhanced at segment centers (e.g., Hooft et al., 2000). Surveys on slow-spreading ridges revealed that gravity anomalies tended toward low values at segment centers where the ridge axis was shallow. At the ridge-transform intersections, the seafloor is deep, and gravity anomalies tend to be high. These gravity "bulls eyes" were shown by Tolstoy et al. (1993) to be correlated with a crust that was thick at the segment centers compared to that adjacent to the transforms.

It appears that thin crust exists at the intersections of rifted ridges and transforms, where we expect the lowest mantle temperatures, and a suppressed magma supply (e.g., Cannat, 1996). The discovery of oceanic core complexes, defined here as exposures of gabbro and/or peridotite on the seafloor, provides observational support models for "new" crustal processes at ridges with slow- to ultraslow-spreading rates (Cann et al., 1997) see (Figure 3). It has been recognized that the ophiolite model may not suffice for such locations. Instead, the model must incorporate substantial rifting that dissects the crust, and exposes the lower crust and upper mantle (Figure 3). Thus, magma intrusions are not continuous beneath oceanic core complexes, and gabbros may not form a continuous layer as it does for the ophiolite model (e.g., Canales et al., 2008). During non-magmatic rifting phases, we would argue that the mafic crustal thickness could be zero; that is, the mantle comes to the seafloor. However, depending on the depth of serpentization in the peridotites, with the consequent decrease of velocities, the seismic crust (velocities less than or equal to 7 km/s) may have a non-zero apparent thickness (Figure 3). The discovery of spectacular, serpentine-hosted hydrothermal circulation at



Lithosphere, Oceanic, Figure 3 Cartoon of rift-dominated processes on a slow- or ultraslow-spreading ridges. Substantial normal faulting along a detachment fault exposes the deeper gabbros and peridotites, as volcanic/magmatic processes cannot "keep up" with the extension. Rock units are the same as those in Figure 2. Serpentinization of exposed peridotites will lower the seismic velocity, perhaps to crustal values. Thus, even where there are no crustal rocks, a thin seismic crust may occur.

least one core complex confirms the importance of rifting and serpentization on the slow- and ultraslow-spreading ridges (Kelley et al., 2001).

Summary

The definition of lithosphere depends on the observations one makes, which in turn depend on the thermomechanical properties on which those observations depend. However, the resulting characteristics are consistent. The basic features of the oceanic lithosphere include an upper mantle component, and it thickens with age. That thickening depends critically on the loss of heat into the ocean, and observations require that cooling in young lithosphere is accommodated by hydrothermal circulation as, well as thermal conduction. Cooling is restricted because of the addition of heat at depths of about 95 km, and that forms the limit for thickening beyond about 80 million years. For ridges spreading at intermediate to fast rates, models for the crustal component of the lithosphere have converged on the Penrose, or ophiolite, model, and this crust is most notable for its striking uniformity. This implies a uniformity of magmatic, hydrothermal, and tectonic processes that act to form that crust. However, recent work on slow and ultraslow ridges shows a great deal more complexity in crustal structure and the processes that form it. It is clear we do not yet understand how models must be changed to accommodate these magma-depleted processes.

Bibliography

Baker, E. T., and German, C. R., 2004. On the global distribution of hydrothermal vent fields. In German, C. R., Lin, J., and Parson, L. M. (eds.), *Mid-Ocean Ridges: Hydrothermal Interactions Between the Lithosphere and Oceans.* Washington, DC: American Geophysical Union. Geophysical Monograph Series, Vol. 148, pp. 245–266.

Barrell, J., 1914. The strength of the Earth's crust. *Journal of Geology*, **22**, 425–433.

Barth, G. A., and Mutter, J. C., 1996. Variability in oceanic crustal thickness and structure: multichannel seismic reflection results from the northern East Pacific Rise. *Journal of Geophysical Research*, **101**, 17951–17975.

Bratt, S. R., and Solomon, S. C., 1984. Compressional and shear wave structure of the East Pacific Rise at 11°20'N: constraints from three-component ocean-bottom seismometer data. *Journal of Geophysical Research*, **89**, 6095–6110.

Calmant, S., Francheteau, J., and Cazenava, A., 1990. Elastic layer thickening with age of the oceanic lithosphere: a tool for prediction of the age of volcanoes or oceanic crust. *Geophysical Journal International*, **100**, 59–67.

Canales, J.-P., et al., 2003. Segment-scale variations in the crustal structure of 150–300 kyr old fast spreading oceanic crust (East Pacific Rise, 8 15 N–10 5 N) from wide-angle seismic refraction profiles. *Geophysical Journal International*, **152**, 766–794.

Canales, J. P., Tucholke, B. E., Xu, M., Collins, J. A., and DuBois, D. L., 2008. Seismic evidence for large-scale compositional heterogeneity of oceanic core complexes. *Geochemistry, Geophysics, Geosystems*, **9**, 1–22, doi:10.1029/2008GC002009.

Canales, J. P., Nedimovic, M. R., Kent, G. M., Carbotte, S. M., and Detrick, R. S., 2009. Seismic reflection images of a near-axis melt sill within the lower crust at the Juan de Fuca Ridge. *Nature*, **460**, 89–94, doi:10.1038/nature08095.

Cann, J. R., Blackman, D. K., Smith, D. K., McAllister, E., Janssen, B., Mello, S., Avgerinos, E., Pascoe, A. R., and Escartn, J., 1997. Corrugated slip surfaces formed at North Atlantic ridge-transform intersections. *Nature*, **385**, 329–332.

Cannat, M., 1996. How thick is the magmatic crust at slow spreading oceanic ridges? *Journal of Geophysical Research*, **101**, 2847–2857.

Chen, Y. J., 1992. Oceanic crustal thickness versus spreading rate. *Geophysical Research Letters*, **19**, 753–756.

Christensen, N. I., and Salisbury, M. H., 1975. Structure and constitution of the lower oceanic crust. *Reviews of Geophysics and Space Physics*, **13**, 57–86.

Corliss, J. B., Dymond, J., Gordon, L. I., Edmond, J. M., von Herzen, R. P., Ballard, R. D., Green, K. L., Williams, D., Brainbridge, A. L., Crane, K., and van Andel, T. H., 1979. Submarine thermal springs on the Galapagos Rift. *Science*, **203**, 1073–1083.

Davis, E. E., and Lister, C. R. B., 1974. Fundamentals of ridge crest topography. *Earth and Planetary Science Letters*, **21**, 405–413. 1974.

Detrick, R. S., Buhl, P., Vera, E., Mutter, J., Orcutt, J., Madsen, J., and Brocher, T., 1987. Multi-channel imaging of a crustal magma chamber. *Nature*, **326**, 35–41.

Detrick, R. S., Harding, A. J., Kent, G. M., Orcutt, J. A., Mutter, J. C., and Buhl, P., 1993. Seismic structure of the southern East Pacific Rise. *Science*, **259**, 499–503.

Elderfield, H., and Schulz, A., 1996. Mid-ocean ridge hydrothermal fluxes and the chemical composition of the ocean. *Annual Review of Earth and Planetary Sciences*, **24**, 191–224.

Forsyth, D. W., 1977. The evolution of the upper mantle beneath mid-ocean ridges. *Tectonophysics*, **38**, 89–118.

Hooft, E. E. E., Detrick, R. S., Toomey, D. R., Collins, J. A., and Lin, J., 2000. Crustal thickness and structure along three contrasting spreading segments of the Mid-Atlantic Ridge, 33.5°–35°N. *Journal of Geophysical Research*, **105**, 8205–8226.

Houtz, R. E., and Ewing, J., 1976. Upper crustal structure as a function of plate age. *Journal of Geophysical Research*, **81**, 2490–2498.

Kelley, D. S., Karson, J. A., Blackman, D. K., Fruh-Green, G. L., Butterfield, D. A., Liley, M. D., Olson, E. J., Schrenk, M. O., Roe, K. K., Lebon, G. T., Rivizzigno, P., and AT3–60 Shipboard Party, 2001. An off-axis hydrothermal vent field discovered near the Mid-Atlantic Ridge at 30°N. *Nature*, **412**, 145–149.

Kent, G. M., Singh, S. C., Harding, A. J., Sinha, M. C., Orcutt, J. A., Barton, P. J., White, R. S., Bazin, S., Hobbs, R. W., Tong, C. H., and Pye, J. W., 2000. Evidence from three-dimensional seismic reflectivity images for enhanced melt supply beneath mid-ocean-ridge discontinuities. *Nature*, **406**, 614–618.

Lister, C. R. B., 1983. On the intermittency and crystallization mechanisms of sub-seafloor magma chambers. *Geophysical Journal of the Royal Astronomical Society*, **73**, 351–365.

Macdonald, K. C., and Fox, P. J., 1988. The axial summit graben and cross-sectional shape of the East Pacific Rise as indicators of axial magma chambers and recent volcanic eruptions. *Earth and Planetary Science Letters*, **88**, 119–131.

McClain, J. S., 2003. Ophiolites and the interpretation of marine geophysical data: How well does the ophiolite model work for the Pacific Ocean crust? *Geological Society of America Special Paper*, **373**, 173–185.

McClain, J. S., and Atallah, C. A., 1985. The structure of young ocean crust near a very fast spreading ridge. *Geophysical Research Letters*, **12**, 680–682.

McClain, J. S., Orcutt, J. A., and Burnett, M., 1985. The East Pacific Rise in cross section: a seismic model. *Journal of Geophysical Research*, **90**, 8627–8629.

Moores, E. M., 1982. Origin and emplacement of ophiolites. *Reviews of Geophysics and Space Physics*, **20**, 735–760.

Morton, J. L., and Sleep, N. H., 1985. A mid-ocean ridge thermal model: constraints on the volume of axial hydrothermal heat flux. *Journal of Geophysical Research*, **90**, 11345–11353.
Nishimura, C., and Forsyth, D., 1989. The anisotropic structure of the upper mantle in the Pacific Ocean. *Geophysical Journal*, **88**, 203–229.
Orcutt, J. A., Kennett, B. L. N., Dorman, L. M., and Prothero, W. A., 1975. Evidence for a low velocity zone underlying a fast-spreading rise crest. *Nature*, **256**, 475–476.
Parsons, B., and Sclater, J., 1977. An analysis of the variation of ocean floor bathymetry and heat flow with age. *Journal of Geophysical Research*, **82**, 803–827.
Shor, G. G., Jr., Menard, H. W., and Raitt, R. W., 1969. Structure of the Pacific Basin. In Maxell, A. E. (ed.), *The Sea*. New York: Wiley, Vol. 4, pp. 3–27.
Singh, S. C., Harding, A. J., Kent, G. M., Sinha, M. C., Combier, V., Bazin, S., Tong, C. H., Pye, J. W., Barton, P. J., Hobbs, R. W., White, R. S., and Orcutt, J. A., 2006. Seismic reflection images of the Moho underlying melt sills at the East Pacific Rise. *Nature*, **442**, 287–290, doi:10.1038/nature04939.
Spudich, P., and Orcutt, J., 1980. Petrology and porosity of an oceanic crustal site: results from wave form modeling of seismic refraction data. *Journal of Geophysical research*, **85**, 1409–1433.
Stein, C. A., and Stein, S., 1992. A model for the global variation in oceanic depth and heat flow with lithospheric age. *Nature*, **359**, 123–129.
Stein, C. A., and Stein, S., 1994. Constraints on hydrothermal heat flux through the oceanic lithosphere from global heat flow. *Journal of Geophysical Research*, **99**, 3081–3095.
Tolstoy, M., Harding, A. J., and Orcutt, J. A., 1993. Crustal thickness on the Mid Atlantic Ridge: Bull's gravity anomalies and focused accretion. *Science*, **262**, 726–729.
Toomey, D. R., Purdy, G. M., and Wilcock, W. S. D., 1990. The three dimensional structure of the East Pacific Rise, near latitude 9°30'N. *Nature*, **347**, 639–645.
Wiens, D. A., and Stein, S., 1983. Age dependence of oceanic intraplate seismicity and implications for lithospheric evolution. *Journal of Geophysical Research*, **88**, 6455–6468.

Cross-references

Earth's Structure, Global
Geodynamics
Heat Flow, Seafloor: Methods and Observations
Lithosphere, Mechanical Properties
Lithosphere, Oceanic: Thermal Structure
Magnetic Methods, Principles

LITHOSPHERE, OCEANIC: THERMAL STRUCTURE

Earl E. Davis[1], David S. Chapman[2]
[1]Pacific Geoscience Centre, Geological Survey of Canada, Sidney, BC, Canada
[2]Department of Geology and Geophysics, The University of Utah, Salt Lake City, UT, USA

Definition

Lithosphere. The outermost shell of the Earth that is sufficiently strong to support long-term geologic loads and transmit stress. These properties are attributed to low temperature and possibly composition.

Asthenosphere. The mobile layer underlying the lithosphere where heat transfer is dominated by convection. Its properties are attributed to high temperature and trace quantities of water.

Plate tectonics. Lateral motion of large intact pieces of lithosphere (plates) carrying both continents and oceans. Mountain building, volcanoes, and earthquakes are the consequences of the relative motion of plates at their convergent, divergent, and transform boundaries.

Boundary-layer cooling. Conductive cooling of an initially hot semi-infinite medium, applied to the vertical cooling of aging oceanic lithosphere created at a seafloor spreading center.

Plate cooling. The description of the cooling of an initially hot layer of uniform thickness having a lower boundary maintained at a uniform temperature.

Rayleigh number. A dimensionless number reflecting the ratio of the forces driving and resisting convection.

Nusselt number. A dimensionless number reflecting the vigor of convection, equal to the ratio of the total heat transport relative to the transport that would take place by conduction alone across the same temperature differential.

Introduction

The thermal structure of the oceanic lithosphere is of profound importance. One of the dominant mechanisms of heat loss from the Earth is the creation at seafloor spreading centers and subsequent cooling of oceanic lithosphere. Ocean lithosphere constitutes roughly 60% of the Earth's surface, and the total heat flow through its surface comprises more than 80% of the global mantle heat flow (i.e., excluding the contribution from radiogenic heat production in continental crust; Jaupart et al., 2007). The thermal structure of the lithosphere plays key roles in constraining the rigidity and ultimate strength (seismic rupture potential) of the outermost layer of the Earth, and influencing many geodynamic processes including the Earth's most fundamental and defining one, plate tectonics. Driving forces for the motion of tectonic plates, and numerous consequent geological processes, are derived from the integrity and density of the oceanic lithosphere as controlled to a large degree by its thermal structure. And the volume of ocean basins, and hence the continental freeboard relative to sea level, are influenced directly by the thermal structure of oceanic lithosphere.

"Lithosphere" has multiple meanings. It was first defined as the outer layer of the Earth, including the crust and upper mantle, having sufficient strength to support loads such as those imposed by mountains, volcanoes, and major river deltas (Barrell, 1914). It was later identified as the high seismic velocity lid that overlies a deeper low-velocity zone in the mantle; its base was inferred to correspond to the closest approach of the geotherm to the mantle melting temperature (Gutenberg, 1959). More recently and most fundamentally, it has become a geodynamic term, which refers to the part of

the mantle that forms relatively rigid plates, and within which heat is transported by conduction. Flexural strength, long-term immobility, seismic properties, and possibly composition are properties that are used to define the lithosphere and its thickness, and all are tied to temperature. Beneath the lithosphere is the relatively mobile asthenosphere, where heat transport is dominated by vigorous (high Rayleigh number, high Nusselt number) convection of near-solidus mantle. The oceanic lithosphere forms a thermomechanical upper boundary layer to a "conveyor-belt" convection system, in which the Earth's interior heat is liberated through the seafloor as the layer moves laterally away from the locus of formation at a seafloor spreading center, and is absorbed by the cooled layer when it sinks back into the asthenosphere at a subduction zone. Convection in the asthenosphere is to a large extent decoupled from the lithosphere. It is characterized by Rayleigh and Nusselt numbers of the order of 10^7 and 200, respectively, and is driven both by cooling at the top by heat lost through the lithosphere, and by heating from the bottom and internally by deep and distributed sources derived from secular cooling and radiogenic heat production.

Oceanic lithosphere includes a 7-km-thick basaltic oceanic crust, produced by partial melting of the upper mantle at a seafloor spreading center, and underlying upper mantle that has been cooled by an amount that gives the material its short- and long-term mechanical and elastic properties that contrast with the underlying mobile asthenosphere. Contrasting composition, a consequence of the refractory process of producing the oceanic crust, may also play a part in defining the properties of the oceanic lithosphere. These various aspects are summarized in Figure 1. Previous reviews can be found in Davis (1989) and Jaupart and Mareschal (2007).

Key observations

Because the creation and cooling of ocean lithosphere is a transient process, observations used to constrain its thermal structure are most commonly placed in the context of lithospheric age, and considered in the context of a model for transient cooling. Various observations are discussed in this section, and theoretical models are discussed in the next.

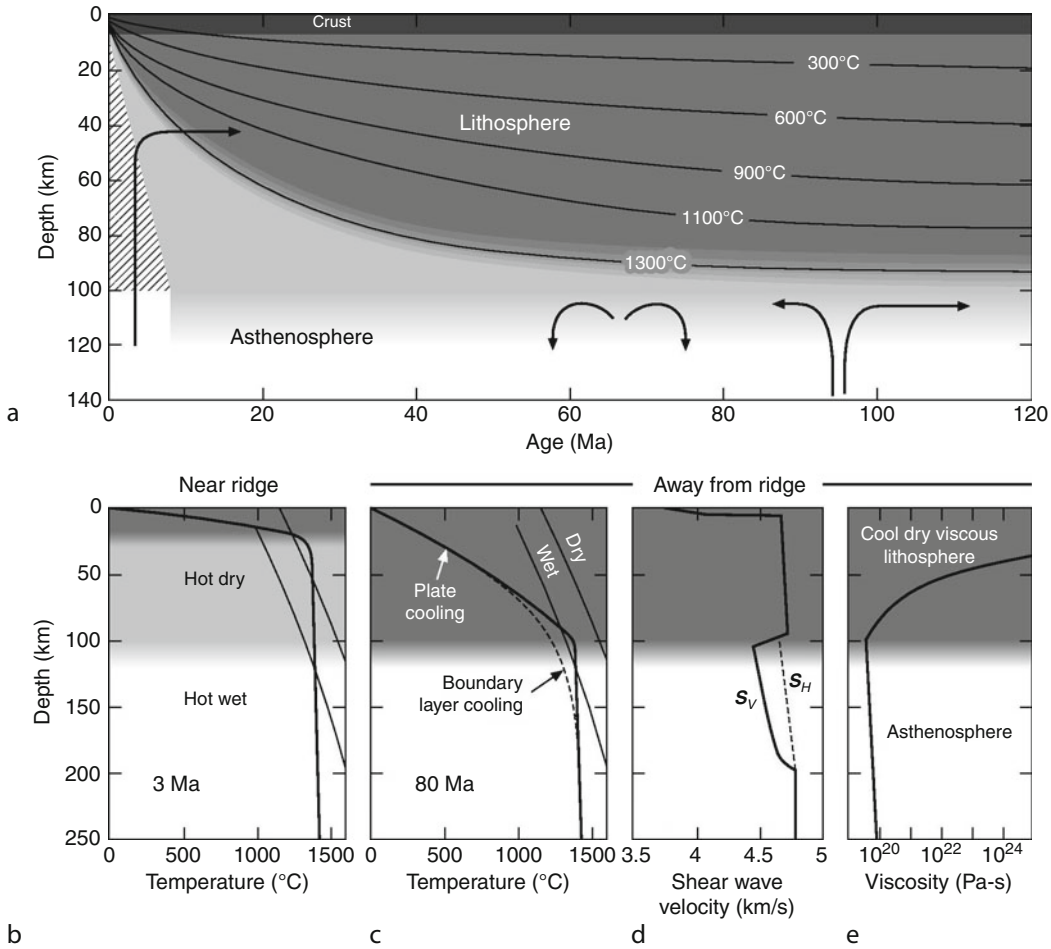
Heat flux

Heat flux through the seafloor provides the most direct constraint on the thermal structure of the underlying lithosphere, although observational uncertainties are often large (Figure 2a). Detailed studies of the thermal regime just below the seafloor commonly reveal a hydrologic regime that in young areas is dominated by high Rayleigh- and Nusselt-number pore-water convection (see discussions in Harris and Chapman, 2004, and *Heat Flow, Seafloor: Methods and Observations* entry, this volume), with high permeabilities in the uppermost few hundred meters of young igneous crust resulting from the volcanic and tectonic processes involved in crustal creation. Hydrothermal alteration of the crust slowly reduces the permeability, but the effects of convective heat transport are observed to persist for tens of millions of years (Von Herzen, 2004). Heat passing advectively through the seafloor is virtually impossible to assess quantitatively, and often goes unrecognized. Hence, wholesale compilations of seafloor heat flux observations are strongly biased, particularly in young, sparsely sedimented areas where igneous crustal outcrops, serving as permeable ventilation points, are frequent. Where thick and extensively continuous accumulations of low-permeability sediment are present, advective ventilation ceases, and the heat flux becomes fully conductive and thus measurable. If circulation within the igneous crust persists, however, heat-flux variations can still be present, and large numbers of measurements must be made over a large area to obtain a meaningful average of the conductive seafloor heat flux at any given age.

To avoid the biasing effects of advective heat loss, two approaches have been used. Values from older data sets have been selected to be far from areas of basement outcrop (relying on statistics for a reliable mean to emerge; e.g., Sclater et al., 1976). Newer surveys have been carried out with large numbers of observation points positioned carefully in context of local and regional sediment/igneous hydrologic structure to rule out effects of ventilation and properly account for the effects of sub-sedimentary fluid flow that produces scatter (e.g., Lister et al., 1990; Davis et al., 1999; Fisher et al., 2003). From such studies, a reliable relationship between lithospheric heat flux and age has emerged, with heat flux decreasing proportionately with $\text{age}^{-1/2}$ out to an age of roughly 100 Ma (Figure 2b), and becoming fairly constant at greater ages at roughly 48 mW m^{-2}.

Seafloor depth

The most reliably measured consequence of lithospheric cooling is seafloor subsidence (Figure 3a). Regardless of the potentially unquantifiable means by which heat passes through the seafloor, the depth-integrated heat loss and accompanying thermal contraction are reflected by seafloor subsidence, and hence bathymetry provides a valuable constraint for understanding lithospheric thermal structure. As in the case of seafloor heat flux, seafloor depths must be considered with some care, with attention given to several potential perturbations. These include checks on isostatic compensation, sediment loading, crustal thickness variations, and sub-lithospheric density anomalies. Early data compilations were made with only modest regard to these factors (e.g., Parsons and Sclater, 1977; Davis and Lister, 1974); later studies, such as those of Schroeder (1984), Johnson and Carlson (1992), Marty and Cazenave (1989), Hillier and Watts (2005), and Crosby et al. (2006), have put more effort into "filtering" data to define depth/age transects that are relatively free of the effects of known or suspected major mantle

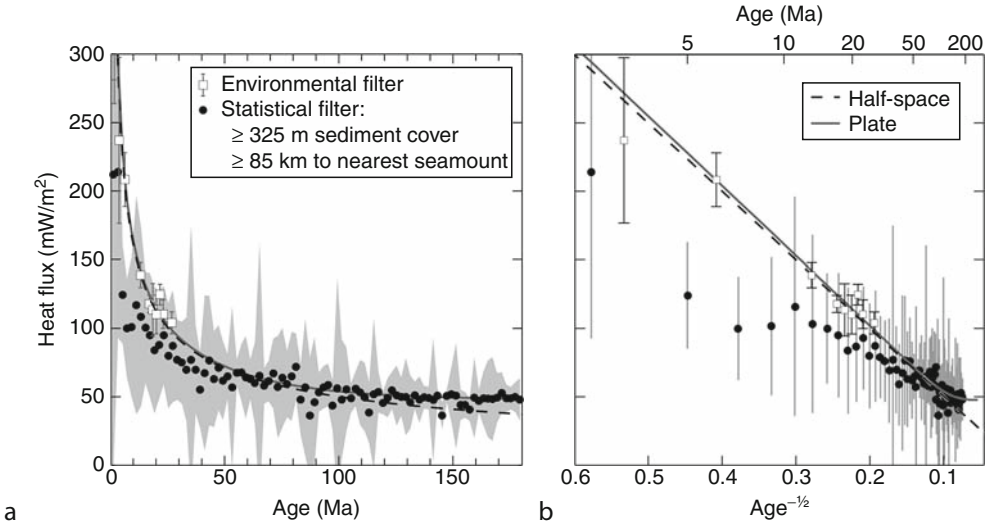


Lithosphere, Oceanic: Thermal Structure, Figure 1 Schematic illustration of the creation and subsequent thermal evolution of oceanic lithosphere (**a**). Dominant control of the mechanical properties of the lithosphere may be thermal (medium gray) or compositional (light gray). Arrows show three modes of convection in the asthenosphere, including upwelling beneath the ridge axis, small-scale convection driven by cooling and associated instability below the base of the lithosphere, and plumes driven by deep-seated heating. In the near-ridge environment, the geotherm (solid line in **b**) intersects a hydrous solidus at a depth of about 100 km; extraction of the partial melt from the rising asthenosphere (hachured zone in (**a**)) produces a basaltic crust about 7 km thick (dark gray), and leaves behind a dried, refractory, and probably stronger residue. Continued conductive cooling at great age (boundary-layer cooling geotherm in (**c**)) may be limited to the compositionally established lithosphere or arrested by regulated convection in the asthenosphere (as shown by the plate cooling geotherm). Properties such as shear-wave seismic velocity (**d**) and effective viscosity (**e**) are controlled by a combination of temperature and water content.

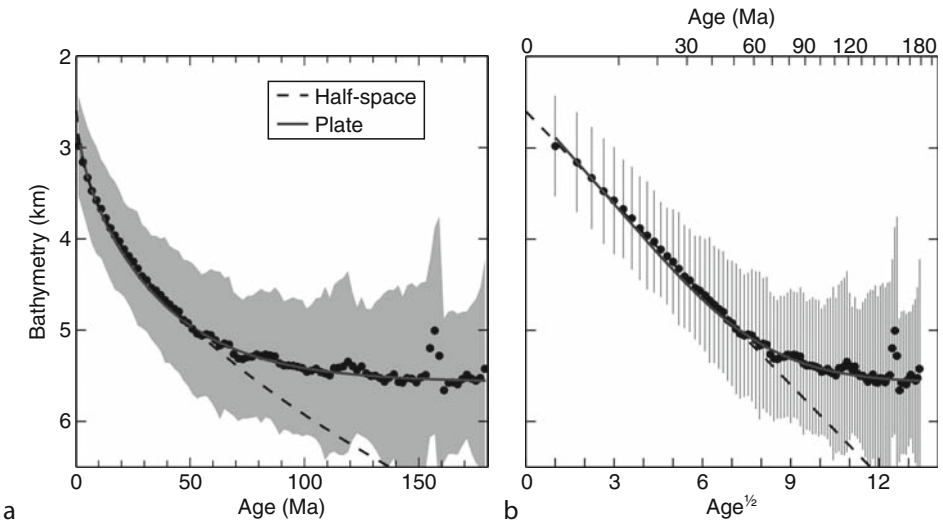
convective upwellings and downwellings, crustal thickness anomalies, and thick sediment accumulations. Despite the various potential sources of "noise" (which, as in the case of heat flux data, add bias, in this case typically causing observed depths to be anomalously shallow), all compilations have shown a similar result, with seafloor depths increasing proportionately with age$^{1/2}$ until 70–80 Ma (Figure 3b). Depths at greater ages fall off of this trend and generally stabilize at about 5.6 km. Some systematic behavior has been gleaned from the compilations, most notably that the rates of subsidence vary both locally and regionally, and that the rate of subsidence is related to the depth of the local ridge axis (Davis and Lister, 1974; Marty and Cazenave, 1989; Hillier and Watts, 2005; Crosby et al., 2006). This is believed to reflect variations in the temperature of the asthenosphere supplying material at spreading centers (Klein and Langmuir, 1987).

Other constraints

Additional constraints on lithospheric thermal structure are provided by observations of shear-wave velocity, flexural rigidity, gravity, and geoid. Gravity data provide little information about the very long wavelength signature of lithospheric cooling except to confirm that the topography is isostatic, but at the shorter wavelengths of abyssal hills, seamounts, and swells, gravity is sensitive to the vertical mass distribution. Hence, it can be used to discriminate



Lithosphere, Oceanic: Thermal Structure, Figure 2 Oceanic heat flux. Data are plotted versus lithospheric age in (**a**) and versus $(\text{age})^{-1/2}$ in (**b**). Because raw heat-flux data in young, sparsely sedimented areas is systematically biased by hydrothermal circulation, the data in these plots have been filtered to exclude sites where sediment thickness is less than 325 m and which are within 85 km of a known seamount. Data are drawn from all oceans. Solid dots show median heat flux for 2 My age bins; gray shading is ± 1 standard deviation for the measurements in a bin. Open symbols represent high-quality data sets drawn from experiments where the environment of the heat flux sites is known from seismic imaging of the sea floor and other geophysical observations. The dashed and solid lines represent heat flux for half-space and plate cooling models, respectively.



Lithosphere, Oceanic: Thermal Structure, Figure 3 Oceanic bathymetry. Data are plotted versus lithospheric age in (**a**) and versus $(\text{age})^{1/2}$ in (**b**). Bathymetry plotted is the position of the top of the igneous crust, 1-D isostatically adjusted for the effect of local sediment loading. Solid dots are median values for 2 My bins of sea floor age; gray shading is ± 1 standard deviation for the measurements in a bin. Dashed and solid lines represent bathymetry for half-space and plate cooling models, respectively.

whether topography is supported or unsupported by lithospheric strength and to constrain the depth of compensation of isostatic topography (e.g., Watts, 2007; Crosby and McKenzie, 2009). The geoid is sensitive to long-wavelength density structure, and provides a useful constraint on the thermally influenced density of oceanic lithosphere. A linear relationship between geoid and age is seen over young seafloor, with a proportionality ranging between 0.1 and 0.14 m Myr^{-1}, whereas the response is flat at ages greater than 80 Ma (e.g., Haxby and Turcotte, 1978; Sandwell and Schubert, 1980; De Laughter et al., 1999).

Laboratory observations and field calibrations show that shear-wave velocity in mantle rocks is a strong function of temperature, with values decreasing rapidly (up to 0.1% K^{-1}) with approach to the melting temperature (Gibb and Cooper, 1998; Priestley and McKenzie, 2006). Hence, shear-wave velocities determined from dispersive surface waves provide a valuable constraint on temperature at depth. Observations show a clear age-dependent structure. A low-velocity zone (Figure 1d), present everywhere in the oceans and centered at roughly 150 km, thickens and rises to the surface at mid-ocean ridges, and extends beneath a high velocity lid (the "seismic lithosphere") that thickens with age out to roughly 70 Ma (Ritzwoller et al., 2004). In addition to temperature, the presence of partial melt and water may be important factors controlling this structure.

Flexural rigidity (usually expressed as effective elastic thickness) and ultimate strength are also strongly controlled by temperature. The first of these related properties is reflected by the long-term load-bearing capacity, and is determined using the wavelength of flexure and the relationship between gravity and topography in the vicinity of seamounts, trenches, fracture zones, and other crustal or tectonic loads. As in the case of the thickness of the high shear-wave-velocity lid, a clear age dependence has been revealed that shows that the effective elastic thickness is limited by temperature. The scatter in thicknesses and inferred maximum temperatures is large, with values falling between 200°C and 600°C. This reflects in part the dependence of this property on the duration of loading.

A limiting constraint on ultimate strength is provided by the occurrence of earthquakes, and when placed in a context of age, a relationship is also seen. The maximum depth of hypocenters increases with increasing age, and the limiting temperatures for seismogenic failure falls at roughly 600°C (McKenzie et al., 2005). Again, the rate of loading that leads to seismogenic failure is an important factor (in addition to temperature) controlling this property.

Ways of understanding lithospheric thermal structure

Boundary-layer cooling

The simplest physical description of the thermal structure of oceanic lithosphere is provided by boundary-layer cooling theory (Parker and Oldenberg, 1973; Davis and Lister, 1974). Except very near the ridge axis, cooling is vertical and the equivalent to cooling of an initially uniform-temperature half space. The lithosphere is defined simply as cooled asthenosphere with the boundary defined by an isotherm that deepens with age and thus distance from the ridge. Heat flux through the seafloor, Q, decreases with age, t, as $Q = \lambda\, T_a\, (\pi\, \kappa\, t)^{-1/2}$, and seafloor depth, h, increases as $h = 2\, \alpha_{eff} T_a\, (\kappa\, t/\pi)\, t^{1/2}$, where T_a = the initial temperature of the asthenospheric material from which the lithospheric forms (relative to seawater at $T \approx 0$°C), κ = thermal diffusivity, $\alpha_{eff} = \alpha\ \rho_0/(\rho_a - \rho_w)$ is the effective volumetric thermal expansivity adjusted for seawater loading, and ρ_a and ρ_w are the densities of the asthenosphere and seawater, respectively. The thermal properties are known to be sensitive to temperature (e.g., Hofmeister, 1999); this must be considered when estimating temperatures at depth (e.g., McKenzie et al., 2005), but it does not change the simple predicted dependence of subsidence and heat flux on age (Davis and Lister, 1974; Lister, 1977). The simplest application of boundary-layer cooling theory produces a heat flux singularity at the ridge axis; this is dealt with by applying a heat-balance boundary condition at the ridge axis. Two important predictions of the theory are that the heat flux, integrated from $t = 0$ to $t = a$ is 2 $Q(a)$, and that the heat flux is related to seafloor depth as $Q = (\rho\ c_P/\alpha_{eff})\ \mathrm{d}h/\mathrm{d}t$, where c_P is the specific heat at constant pressure and α_{eff} is appropriate for 0.7 of T_a (Lister, 1977). Wherever cooling follows boundary-layer behavior, this relationship can be used to obtain a robust estimate of the heat loss from the oceanic lithosphere without a need for relying on sparse, scattered, and potentially biased seafloor heat-flux data.

When plotted as functions of $t^{-1/2}$ and $t^{1/2}$ as they are in Figures 2b and 3b, and considered in light of the heat-flux/subsidence relationship given above, the heat-flux and depth observations clearly demonstrate the efficacy of boundary-layer cooling theory for characterizing lithosphere younger than 70–80 Ma. The same is true for the geoid, which is predicted by boundary-layer cooling theory to decline at a rate of 0.16 m Myr^{-1} (Haxby and Turcotte, 1978). The cessation of simple subsidence and flattening of the geoid beyond 70–80 Ma and the relatively constant heat flux beyond an age of roughly 100 Ma, require something more than the heat supplied via transient cooling, however.

Plate cooling models

That ocean depths at great ages are relatively uniform and consistently shallower than those predicted by boundary-layer cooling provided the fundamental rationale for plate cooling models, which describe the lithosphere as a layer of constant thickness having a fixed lower boundary temperature. This description (McKenzie, 1967) was initially used out of mathematical convenience. There was little physical justification for the lithosphere to have an intrinsic thickness, or for constant heat flux to be supplied by asthenospheric convection beneath old lithosphere, until physical arguments were made for the latter by Parsons and McKenzie (1978) by way of an age-dependent onset of convection in the growing thermal boundary layer below the cooled, mechanically immobile lithosphere. Physical modeling of convection using material having a strongly temperature-dependent viscosity has shown how this phenomenon can serve to stop and possibly even reverse subsidence, and later stop the decline in heat flux with age (Davaille and Jaupart, 1994; Crosby et al., 2006). The key factor in this model is the growth of the

thermal boundary layer beneath the conductively cooled plate. Once this layer reaches a sufficient thickness to become convectively unstable, convection begins to bring heat from the large reservoir of the asthenosphere beneath. A balance is ultimately reached between the heat conducted through the lithosphere and that carried by convection in the asthenosphere.

Alternate means suggested for maintaining an effectively steady-state lithosphere at great age include convection driven from below (plumes, distributed radiogenic heat, and secular cooling), and creation of a compositionally distinct lithosphere at the ridge axis. The first by itself is precluded as a general case by both topographic and heat-flux observations, which show no statistically significant influence of augmented heat flux at intermediate ages. The second may provide a viable alternative to age-initiated top-down convection. The means for creating a compositional constraint on lithospheric thickness can be understood by considering Figure 1a and b. Asthenosphere rising beneath a ridge axis to balance the mass lost to lateral plate motion will begin to melt when it reaches a depth where the adiabat meets the local solidus. The depth at which this occurs depends on the water content of the asthenosphere. Estimates for water content range from 50 to 150 ppm by weight, and the depth at which melting begins is estimated at roughly 115 km, twice the depth for dry conditions (Hirth and Kohlstedt, 1996; Azimow and Langmuir, 2003). The partial melt fraction segregates and rises to produce the oceanic crust, leaving a residue that may differ in composition, and hence in mechanical properties, from normal asthenosphere (Oxburgh and Parmentier, 1977; Morgan, 1997). More importantly, water, being much more soluble in the melt than in the solid fraction, is very efficiently removed from the residue, and this drying will have a strong influence on the viscosity of the residual asthenosphere. Hirth and Kohlstedt (1996) estimate an increase in viscosity by a factor of 500 may result from complete drying. Hence, the mechanical property of the lithosphere that controls convection may indeed be established very near the ridge axis. Lithospheric cooling will initially follow conductive boundary-layer behavior, and temperature-sensitive properties such as shear-wave velocity, flexural strength, and seismic rupture potential, will follow isotherms. Once the depth of the compositional change established near the ridge is reached by cooling, the effects of convection in the normal, relatively wet, and hence invicid asthenosphere away from the ridge axis will be manifest in arrested subsidence, then uniform heat flux. If this is the case, then the thermal state of old lithosphere will simply reflect the combination of its intrinsic thickness and the relatively uniform temperature of the vigorously convecting asthenosphere beneath.

Influence of plumes

Convective plumes in the asthenosphere have been argued to thin, or "reset" the thermal age of oceanic lithosphere (see review in Detrick et al., 1989), with the eventual collective contribution from plumes resulting in the departure from boundary-layer cooling behavior (Heestand and Crough, 1981) and possibly in the supply of much of the heat flux to the base of plates (Malamud and Turcotte, 1999), although these possibilities remain controversial. The influence of plumes can certainly be large; while surface volcanism from plumes is typically highly focused, the associated topographic swells are large and extensive, often exceeding 1 km in amplitude and having a wavelength approaching 1,000 km. A similarly broad thermal anomaly is implied as a source for swells. Given the large thermal time constant of the lithosphere, seafloor heat flux provides an inherently poor constraint on the depth of the thermal source; attempts have been made to detect anomalies with understandably mixed results (e.g., Von Herzen et al., 1989). A study of the geoid signature over the Hawaiian Swell by Moore and Schubert (1997) suggests that the lower part of the lithosphere is involved, but a study of the topography and gravity over swells by Crosby and McKenzie (2009) reveals an admittance (30 mGal km^{-1}) that is consistent with the thermal density anomaly being below the base of the plate. This study also shows no evidence for age dependence of the admittance, which suggests that the conductive lithosphere is established at a very young age. If true, this would support the hypothesis that oceanic lithosphere acquires one of its key properties by virtue of its composition established at the ridge axis.

Summary

A broad suite of observations gathered in context of powerful physical models have brought plate tectonics from a largely geometrical description of the history of the surface of the Earth to a robust theory. Among other things, this theory addresses the mechanisms by which plates form and are driven, the thermal structure of oceanic lithosphere that influences its mechanical and geodynamic behavior, and the dominant modes of heat loss from the Earth. Uncertainties remain, however, and further work is required to resolve several lingering questions about lithospheric creation, cooling, and eventual stabilization. Among these are (1) to what degree is the mechanical behavior of oceanic lithosphere (particularly its resistance to thinning by convection of any form in the asthenosphere) established compositionally by partial-melt extraction at ridge axes, (2) to what degree does asthenospheric convection stimulated by top-down cooling regulate heat loss through and limit the growth of the lithosphere at great age, and (3) to what degree do plumes, the consequence of convective instabilities generated deep in the asthenosphere, contribute to heat flux at the base of plates, and do they convectively thin or merely conductively warm the plates beneath which they ascend? Addressing the first will improve constraints on plate drive generated by gravitational sliding forces, which are maximum for the case of a mechanical lithosphere being

defined purely as a cooling boundary layer. Addressing the second and third will help resolve just what the primary regulating factor is that constrains the rate of heat loss from the Earth, such as the vigor of asthenospheric convection, the quotient of asthenospheric temperature and lithospheric thickness, or the rate of production and destruction of lithosphere at ridges and subduction zones.

Bibliography

Azimow, P. D., and Langmuir, C. H., 2003. The importance of water to oceanic mantle melting regimes. *Nature*, **421**, 815–820.

Barrell, J., 1914. The strength of the Earth's crust. I. Geologic tests of the limits of strength. *Journal of Geology*, **22**, 28–48.

Crosby, A. G., and McKenzie, D., 2009. An analysis of young ocean depth, gravity and global residual topography. *Geophysical Journal International*, **178**, 1198–1219.

Crosby, A. G., McKenzie, D., and Sclater, J. G., 2006. The relationship between depth, age, and gravity in the oceans. *Geophysical Journal International*, **166**, 555–573.

Davaille, A., and Jaupart, C., 1994. Onset of thermal convection in fluids with temperature-dependent viscosity: application to the oceanic mantle. *Journal of Geophysical Research*, **99**, 19853–19866.

Davis, E. E., 1989. Thermal aging of oceanic lithosphere. In Wright, J. A., and Louden, K. E. (eds.), *Handbook of Seafloor Heat Flow*. Boca Raton: CRC press, pp. 145–167.

Davis, E. E., and Lister, C. R. B., 1974. Fundamentals of ridge crest topography. *Earth and Planetary Science Letters*, **21**, 405–413.

Davis, E. E., Chapman, D. S., Wang, K., Villinger, H., Fisher, A. T., Robinson, S. W., Grigel, J., Pribnow, D., Stein, J. S., and Becker, K., 1999. Regional heat flow variations on the sedimented Juan de Fuca Ridge eastern flank: constraints on lithospheric cooling and lateral hydrothermal heat transport. *Journal of Geophysical Research*, **104**, 17675–17688.

De Laughter, J., Stein, S., and Stein, C. A., 1999. Extraction of a lithospheric cooling signal from oceanwide geoid data. *Earth and Planetary Science Letters*, **174**, 173–181.

Detrick, R. S., White, R. S., Courtney, R. C., and Von Herzen, R. P., 1989. Heat flow on midplate swells. In Wright, J. A., and Louden, K. E. (eds.), *Handbook of Seafloor Heat Flow*. Boca Raton: CRC press, pp. 69–190.

Fisher, A. T., Stein, C. A., Harris, R. N., Wang, K., Silver, E. A., Pfender, M., Hutnak, M., Cherkaoui, A., Bodzin, R., and Villinger, H., 2003. Abrupt thermal transition reveals hydrothermal boundary and role of seamounts within the Cocos plate. *Geophysical Research Letters*, **30**, 1550, doi:10.1029/2002GL016766.

Gibb, T. T., and Cooper, R. F., 1998. Low-frequency shear wave attenuation in polycrystalline olivine: grain boundary diffusion and the physical significance of the Andrade model for viscoelastic rheology. *Journal of Geophysical Research*, **103**, 27267–27279.

Gutenberg, B., 1959. *Physics of the Earth's Interior*. New York: Academic.

Harris, R. N., and Chapman, D. S., 2004. Deep-seated oceanic heat flux, heat deficits, and hydrothermal circulation. In Davis, E. E., and Elderfield, H. (eds.), *Hydrogeology of the Oceanic Lithosphere*. Cambridge: Cambridge University Press, pp. 311–336.

Haxby, W. F., and Turcotte, D. L., 1978. On isostatic geoid anomalies. *Journal of Geophysical Research*, **83**, 5473–5478.

Heestand, R. L., and Crough, T. S., 1981. The effect of hot spots on the oceanic age-depth relation. *Journal of Geophysical Research*, **86**, 6107–6114.

Hillier, J. K., and Watts, A. B., 2005. Relationship between depth and age in the North Pacific Ocean. *Journal of Geophysical Research*, **110**, 1–22, doi:10.1029/2004JB003406.

Hirth, G., and Kohlstedt, D. L., 1996. Water in the oceanic upper mantle: implications for rheology, melt extraction and the evolution of the lithosphere. *Earth and Planetary Science Letters*, **144**, 93–108.

Hofmeister, A., 1999. Mantle values of thermal conductivity geotherm from phonon lifetimes. *Science*, **283**, 1699–1709.

Jaupart, C., and Mareschal, J.-C., 2007. Heat flow and thermal structure of the lithosphere. In Watts, A. (ed.), *Treatise on Geophysics*. Oxford: Elsevier, Vol. 6, pp 217–251.

Jaupart, C., Labrosse, S., and Mareschal, J.-C., 2007. Temperatures, heat, and energy in the mantle of the Earth. In Bercovici, D. (ed.), *Treatise on Geophysics*. Oxford: Elsevier, Vol. 7, pp 253–303.

Johnson, H. P., and Carlson, R. L., 1992. Variations of sea floor depth with age: A test of models based on drilling results. *Geophysical Research Letters*, **19**, 1971–1974.

Klein, E. M., and Langmuir, C. H., 1987. Global correlations of ocean ridge basalt chemistry with axial depth and crustal thickness. *Journal of Geophysical Research*, **92**, 8089–8115.

Lister, C. R. B., 1977. Estimators for heat flow and deep rock properties based on boundary layer theory. *Tectonophysics*, **41**, 157–171.

Lister, C. R. B., Sclater, J. G., Davis, E. E., Villinger, H., and Nagihara, S., 1990. Heat flow maintained in ocean basins of great age: investigations in the north-equatorial west Pacific. *Geophysical Journal International*, **102**, 603–630.

Malamud, B. D., and Turcotte, D. L., 1999. How many plumes are there? *Earth and Planetary Science Letters*, **174**, 113–124.

Marty, J. C., and Cazenave, A., 1989. Regional variations in subsidence rate of oceanic plates: a global analysis. *Earth and Planetary Science Letters*, **94**, 301–315.

McKenzie, D., 1967. Some remarks on heat flow and gravity anomalies. *Journal of Geophysical Research*, **72**, 6261–6273.

McKenzie, D., Jackson, J., and Priestley, K., 2005. Thermal structure of oceanic and continental lithosphere. *Earth and Planetary Science Letters*, **233**, 337–349.

Moore, W. B., and Schubert, G., 1997. Lithospheric thinning and chemical buoyancy beneath the Hawaiian Swell. *Geophysical Research Letters*, **24**, 1287–1290.

Morgan, J. P., 1997. The generation of a compositional lithosphere by mid-ocean ridge melting and its effect on subsequent off-axis hotspot upwelling and melting. *Earth and Planetary Science Letters*, **146**, 213–232.

Oxburgh, E. R., and Parmentier, E. M., 1977. Compositional and density stratification in oceanic lithosphere – causes and consequences. *Journal of the Geological Society*, **133**, 343–355.

Parker, R. L., and Oldenberg, D. W., 1973. Thermal model of ocean ridges. *Nature*, **242**, 137–139.

Parsons, B., and McKenzie, D., 1978. Mantle convection and the thermal structure of the plates. *Journal of Geophysical Research*, **83**, 4485–4496.

Parsons, B., and Sclater, J. G., 1977. An analysis of the variation of ocean floor bathymetry and heat flow with age. *Journal of Geophysical Research*, **82**, 803–827.

Priestley, K., and McKenzie, D., 2006. The thermal structure of the lithosphere from shear wave velocities. *Earth and Planetary Science Letters*, **244**, 285–301.

Ritzwoller, M. H., Shapiro, N. M., and Zong, S.-J., 2004. Cooling history of the Pacific lithosphere. *Earth and Planetary Science Letters*, **226**, 69–84.

Sandwell, D., and Schubert, G., 1980. Geoid height versus age for symmetric spreading ridges. *Journal of Geophysical Research*, **85**, 7235–7241.

Schroeder, W., 1984. The empirical age-depth relation and depth anomalies in the Pacific Ocean basin. *Journal of Geophysical Research*, **89**, 9873–9883.
Sclater, J. G., Crowe, J., and Anderson, R. N., 1976. On the reliability of ocean heat flow averages. *Journal of Geophysical Research*, **81**, 2997–3006.
Von Herzen, R. P., 2004. Geothermal evidence for continuing hydrothermal circulation in older (>60 M.y.) oceanic crust. In Davis, E. E., and Elderfield, H. (eds.), *Hydrogeology of the Oceanic Lithosphere*. Cambridge: Cambridge University Press, pp. 414–447.
Von Herzen, R. P., Cordery, M. J., Detrick, R. S., and Fang, C., 1989. Heat flow and the thermal origin of hot spot swells: The Hawaiian swell revisited. *Journal of Geophysical Research*, **94**, 13783–13799.
Watts, A. B., 2007. Crust and lithosphere dynamics: An overview. In Watts, A. (ed.), *Treatise on Geophysics*. Oxford: Elsevier, Vol. 6, pp 1–48.

Cross-references

Heat Flow, Seafloor: Methods and Observations
Lithosphere, Continental
Lithosphere, Mechanical Properties
Lithosphere, Oceanic

M

MAGNETIC ANISOTROPY

Leonardo Sagnotti
Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy

Definition

The magnetic properties of rocks and minerals are generally anisotropic, that is, they are directionally dependent. Magnetic anisotropy of minerals arises from either fundamental anisotropy in the crystal structure or from the shape of nonspherical ferromagnetic grains. The most common cause of magnetic anisotropy in rocks is the preferential distribution-orientation of the constituting minerals, in other words the rock fabric.

Introduction

Pioneering works on the magnetic anisotropy of rocks were carried out during the 1940s and 1950s (Ising, 1942; Graham, 1954). These authors first realized that magnetic methods may be used to characterize the preferred orientation of minerals within the rock samples. Ising studied varved clays in Sweden and noticed that the magnetic susceptibility was higher on the bedding plane than orthogonally to it. Graham recognized that the anisotropy of magnetic susceptibility (AMS) may be regarded as a petrofabric element; he later extended the analysis to various sedimentary rocks of the Appalachian Mountains and pointed out the existence of distinct and systematic relationships of the magnetic properties with structural setting (Graham, 1966). The studies progressively developed in the following decades and a first comprehensive review on magnetic anisotropy and its application in geology and geophysics was published by Hrouda (1982). Over the past 20–30 years, researches on magnetic anisotropy gained widespread use and were extended to examine the fabric in a variety of sedimentary, igneous, and metamorphic rocks (e.g., see reviews by Jackson, 1991; Jackson and Tauxe, 1991; Rochette et al., 1992; Tarling and Hrouda, 1993; Borradaile and Henry, 1997; Borradaile, 2001; Borradaile and Jackson, 2004; Tauxe, 2005; Lanza and Meloni, 2006; Hrouda, 2007). Presently, the study of the magnetic anisotropy of rocks is still one of the most promising research issues in the field of rock magnetism.

Theoretical principles

Two principal mechanisms control the magnetic anisotropy of rocks: (1) lattice alignment of crystals with magnetocrystalline anisotropy, (2) shape alignment of ferromagnetic grains.

Magnetocrystalline anisotropy

This is an intrinsic property of all minerals since it arises from the ordered structure of elementary particles in the crystal lattice; it is defined as the energy per atom needed to move the crystal magnetization from one crystallographic direction to another. Magnetocrystalline anisotropy thus reflects the dependence of the free energy of a magnetic system on the direction of magnetization. In the atomic structure of a crystal, the spin–orbit coupling between the magnetic spins and the lattice establish "easy" and "hard" directions of magnetization and determine the orientation of the spin moments with respect to the crystal structure. Within ferromagnetic crystals (the term ferromagnetic will be used in the following in its broader sense, that is to indicate all minerals with ordered magnetic spin structures, including ferrimagnetic and antiferromagnetic ordering), the exchange energy in the 3d electronic orbitals of transition elements (e.g., Fe, Co, Ni) is, unlike the *s* orbitals, anisotropic. Hence, within these crystals

Harsh K. Gupta (ed.), *Encyclopedia of Solid Earth Geophysics*, DOI 10.1007/978-90-481-8702-7,

electron spins will be easier to coordinate in some directions than in others. Along these directions, the elemental electron spins align more readily and the magnetization acquired in an applied field is greatest. In magnetite the direction of the lowest magnetocrystalline anisotropy energy (i.e., the easy axis of magnetization) is along the body diagonal, that is the [111] crystallographic direction. In hematite, magnetocrystalline anisotropy is very weak in the basal plane, but it is very high along the crystallographic *c*-axis, perpendicular to this plane. As a consequence, hematite crystals may be easily magnetized by magnetic fields applied along the basal plane and are almost unaffected by fields applied perpendicular to it.

Shape anisotropy

Shape anisotropy mostly occurs in ferromagnetic minerals with high intrinsic susceptibility and low degree of magnetocrystalline anisotropy (e.g., magnetite, maghemite), where just a slight deviation from the isometric shape produces notable magnetic anisotropies, in order to minimize the magnetostatic energy due to the formation of surface magnetic poles at the extremities of the mineral grains. In ferromagnetic minerals, the alignment of electron spins produced by exchange energy or by the application of an external magnetic field generates an external magnetic field that is proportional to the spontaneous or induced magnetic moment and may be represented as a field produced by an apparent distribution of north and south magnetic poles on the surface of the crystal. In turn, the surface magnetic poles generate an internal field whose field lines start on the north magnetic poles on the surface and end on the south poles (Figure 1). This internal field is, therefore, antiparallel to the alignment of the magnetic moments within the crystal and it is known as the demagnetizing field $\mathbf{H}_d$. Such a demagnetizing field is proportional to the magnetization $\mathbf{M}$ of the crystal and is sensitive to its shape:

$$\mathbf{H}_d = -N\mathbf{M}$$

where N is a demagnetizing factor which is determined by the shape of each crystal.

Shape anisotropy arises from the dependence of the demagnetization factor N on the direction of magnetization with respect to the particle shape; different directions resulting in different arrays of surface poles. By analogy with electrostatics, magnetostatic forces decrease with distance and the internal demagnetizing field is reduced when the surface poles are furthest apart. For a spherical grain $N = 1/3$. In an ellipsoid grain with principal axes $a \geq b \geq c$, $N_a + N_b + N_c = 1$. In the case of an ellipsoid magnetized parallel to the elongation axis a the demagnetizing field, which depends on $1/r^2$, is weaker than in the case of a sphere (i.e., $N_a < 1/3$). Similarly, if the ellipsoid is magnetized along the shortest axis c, the demagnetizing field is stronger (i.e., $N_c > 1/3$). In general, the demagnetizing factor N is minimum for very elongated ("pencil-shaped") grains and maximum for platy grains ("disk-shaped"). In an ellipsoid of revolution, N_a approaches 0 for an infinite needle ($N_b = N_c = ½$), and N_c approaches 1 for a disk of infinite thinness ($N_a = N_b = 0$).

The demagnetizing factor N determines the intensity of the magnetic field which effectively acts ($\mathbf{H}_{eff}$) within a grain subjected to an external applied field $\mathbf{H}_{ex}$:

$$\mathbf{H}_{eff} = \mathbf{H}_{ex} - N\mathbf{M}$$

The magnetization M_j of a grain along an arbitrary direction j is given by

$$M_j = k_0\mathbf{H}_{ex} = k_i\mathbf{H}_{eff} = k_i\left(\mathbf{H}_{ex} - N_jM_j\right)$$

where k_0 is the observed magnetic susceptibility along the direction j, k_i is the intrinsic magnetic susceptibility of the grain and $\mathbf{H}_{ex}$ is the intensity of the external magnetic field.

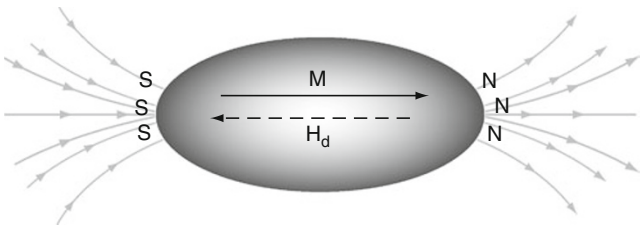
The observed magnetic susceptibility along a given direction of a grain is related to the intrinsic magnetic susceptibility of the mineral by

$$k_0 = k_i/(1 + Nk_i)$$

The degree of magnetic susceptibility shape anisotropy in a grain is defined as

$$P = (1 + k_iN_c)/(1 + k_iN_a);$$

where the intrinsic susceptibility k_i is isotropic and dependent only on the chemical composition of the mineral. Therefore, the net effect of the shape anisotropy is to favor the orientation of the induced magnetization along the long axis of a magnetic grain, where the internal demagnetizing field is minimized.



Magnetic Anisotropy, Figure 1 Surface magnetic poles and internal demagnetizing field H_d (dashed arrow) for an ellipsoidal grain with uniform axial magnetization M.

Magnetostriction – stress anisotropy

The application of an external stress to a rock may cause reversible as well as irreversible changes to its magnetic anisotropy. In fact, straining a ferromagnetic crystal can rotate its spontaneous magnetization away from the easy axis given by magnetocrystalline anisotropy. Also, the application of a magnetic field that rotates the spontaneous magnetization of a ferromagnetic crystal away from its intrinsic preferred direction can strain the lattice and modify the crystal shape, producing changes in its demagnetization factors and shape anisotropy. Both processes are

expression of the phenomenon known as magnetostriction and are due to changes in the exchange energy when modifications in the magnetization or in the crystal shape alter the spin–orbit coupling in neighboring atoms (e.g., O'Reilly, 1984).

Magnetostriction is defined as the spontaneous change in crystal dimensions that accompanies the process of magnetization or as the change in the magnetization of a crystal as the result of the application of stress. The linear saturation magnetostriction constant λ is defined as the fractional change in the length $\Delta L/L$ of a demagnetized ferromagnetic crystal as its magnetization increases from zero to saturation (e.g., Moskowitz, 1993). Magnetostriction is positive when the crystal expands in the direction of magnetization. In magnetite at room temperature, the magnetostriction depends crucially on the direction, being positive for a magnetization along a <111> easy axis but negative for magnetization along a <100> hard axis.

In a cubic crystal, the anisotropy of λ is given by:

$$\lambda = 3/2\lambda_{100}(\alpha_1{}^2\gamma_1{}^2 + \alpha_2{}^2\gamma_2{}^2 + \alpha_3{}^2\gamma_3{}^2 - 2/3) + 3\lambda_{111}(\alpha_1\alpha_2\gamma_1\gamma_2 + \alpha_2\alpha_3\gamma_2\gamma_3 + \alpha_3\alpha_1\gamma_3\gamma_1),$$

where α_i, ($i = 1, 2, 3$) are the direction cosines of saturation magnetization (**Ms**) with respect to the <100>, <010>, and <001> crystallographic axes, respectively, and γ_i are the equivalent direction cosines of the direction along which the magnetostriction is being measured (Dunlop and Özdemir, 1997). λ_{100} and λ_{111} are the crystal's magnetostriction constants along the <100> and <111> crystallographic axes for **M** aligned in the same direction.

In the case of isotropic magnetostriction (i.e., $\lambda_s = \lambda_{100} = \lambda_{111}$), the magnetoelastic energy caused by the application of a stress σ to a magnetic crystal is given by:

$$E = 3/2\lambda_s\sigma\sin^2\theta$$

where θ is the angle between **Ms** and σ.

In this simple case, the energy is minimum when **Ms** is parallel to σ and $\lambda_s\,\sigma > 0$ or when **Ms** is perpendicular to σ and $\lambda_s\,\sigma < 0$. Thus, the application of a uniaxial stress will cause **Ms** to rotate toward the stress axis if $\lambda_s\,\sigma > 0$ (i.e., under the combination of a compression and a negative λ_s or of a tension and positive λ_s).

In the general case, the relationship between the change in the magnetic anisotropy and the increase of strain induced by a stress field is not simple, reflecting the complex behavior due to the superposition of the strain effects upon an already magnetically anisotropic specimen.

Exchange anisotropy

Exchange anisotropy results from superexchange interaction across the interface of two different magnetically ordered phases (Meiklejohn, 1962; Berkowitz and Takano, 1999). In the simplest model, the origin of exchange anisotropy is considered to be the coupling of a ferromagnetic spin system to an antiferromagnetic spin system, separated by a planar interface. To obtain a preferred direction of the coupling it is necessary that Curie temperature T_C of the ferromagnet be greater than the Néel temperature T_N of the antiferromagnet. When a magnetic field is applied at a temperature $T_N < T < T_C$, the spins of the ferromagnet will orient parallel to the applied field. Then, as the material is cooled through T_N, the spins of the antiferromagnetic lattice closest to the ferromagnet will align in the same direction as the ferromagnet and subsequent spin planes in the antiferromagnetic lattice will orient antiparallel to each other. When the antiferromagnet is fully ordered, with high magnetocrystalline anisotropy, it holds the magnetization of the ferromagnetic material in the direction of the applied field, giving rise to a unidirectional magnetic anisotropy.

Exchange anisotropy can be detected from a shifted hysteresis loop when the material is cooled in a magnetic field and from the presence of a $\sin\theta$ term in torque curves measured in high fields due to a rotational hysteresis loss.

Exchange anisotropy in natural minerals has been found in titanomagnetite or in the intergrowth of maghemite with hematite and has been invoked to explain self-reversals in the direction of magnetization of natural rocks (e.g., Nagata and Uyeda, 1959). In the last decades there has been an increase of interest about this magnetic phenomenon because of the importance of its technological applications in spin-valve devices, such as hard-disk read heads, and in magnetic thin layers.

Anisotropy of magnetic susceptibility (AMS)

The magnetic anisotropy of rocks is often determined by means of the analysis of the anisotropy of the magnetic susceptibility (AMS), which is the property that has found the most applications in geophysical studies so far. The AMS of rocks is controlled by preferentially oriented magnetic mineral grains and, therefore, it contains information about both the grain susceptibilities and their orientation-distribution. Knowing the composition of the rock-forming minerals and their magnetic anisotropy characteristics, it is possible to determine the spatial distribution of the grains and infer the geological processes that originated it. A preferential orientation-distribution of mineral grains is in fact typical of almost all rock types and it develops during various geological processes, such as water flow in sediments, magma flow in igneous rocks, ductile deformation in metamorphic rocks, and even incipient strain in the paramagnetic clay matrix of apparently undeformed fine-grained sediments.

The magnetic susceptibility is the capability of a material to be magnetized under the effect of an external magnetic field; all the materials are "susceptible" to become magnetized in the presence of an applied magnetic field, and the magnetic susceptibility describes this transient magnetism within a material sample. If the magnetic field is relatively weak, the magnetization of a rock is a linear function of the intensity of this field. The low-field magnetic susceptibility is defined as the ratio of the induced magnetization (**M**, dipole moment per unit volume or **J**, dipole moment per unit mass) to the applied low-intensity

magnetic field (**H**). Only for isotropic substances the induced magnetization is strictly parallel to the applied field, and the magnetic susceptibility is a scalar. In the general case of anisotropic media, like minerals and rocks, the induced magnetization is not parallel to the applied field, and the magnetization induced along the direction i is related to the magnetic field acting along the direction j by:

$J_i = \chi_{ij} H_j$ (mass specific), where χ is given in SI units of m^3/kg

or

$M_i = k_{ij} H_j$ (volume specific), where k is dimensionless in SI units

On the basis of the magnetic susceptibility all substances may be classified as diamagnetic, paramagnetic, and ferromagnetic.

In diamagnetic minerals the induced magnetization increases linearly with the increasing field, but in the opposite direction; the magnetic susceptibility is therefore negative and typically very low ($k \sim -10^{-5}$ SI). In paramagnetic minerals the magnetization increases linearly with increasing field, along the same direction, and the susceptibility is positive ($k \sim 10^{-4}$ to 10^{-5} SI). In ferromagnetic minerals, the induced magnetization increases nonlinearly with increasing field and the magnetic susceptibility is typically positive and much higher (k ranges from 10^{-3} to 10 SI) than in diamagnetic and paramagnetic minerals.

If the applied field is so low to approximate as linear the magnetic response of the body, the magnetic susceptibility can be represented by a set of coefficients (k_{ij}) that form a second-order symmetric tensor; each coefficient quantifies how much a body can be magnetized along a certain Cartesian direction according to a Cartesian component of the applied field. The relationship between the components of the magnetization M_i (i = 1, 2, 3) and the components of the magnetic field H_j (j = 1, 2, 3) is expressed by the equations

$$M_1 = k_{11}H_1 + k_{12}H_2 + k_{13}H_3$$

$$M_2 = k_{21}H_1 + k_{22}H_2 + k_{23}H_3$$

$$M_3 = k_{31}H_1 + k_{32}H_2 + k_{33}H_3$$

which can be rewritten in subscript notation as:

$$M_i = k_{ij}H_j$$

where the coefficients k_{ij} define a matrix that has six independent elements, since the second-order tensor must be symmetric to guarantee real eigenvalues, and k_{ij} is imposed equal to k_{ji}. Among all possible Cartesian reference systems, there exists one in which the non-diagonal terms of the tensor are zero so that the above equations simplify to:

$$M_1 = k_{11}H_1$$

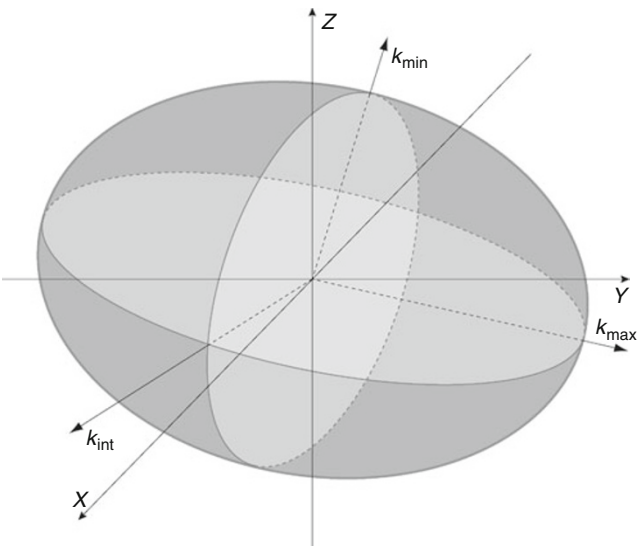
$$M_2 = k_{22}H_2$$

$$M_3 = k_{33}H_3$$

where in this case the three coefficients k_{11}, k_{22}, and k_{33} are the eigenvalues of the magnetic susceptibility tensor: they are called the principal susceptibilities and are generally indicated as $k_{max} \geq k_{int} \geq k_{min}$ (or $k_1 \geq k_2 \geq k_3$), the maximum, intermediate, and minimum susceptibilities

The corresponding eigenvectors are the maximum, intermediate, and minimum principal susceptibility directions that are the directions along which the induced magnetization is strictly parallel to the direction of the applied field. The magnetic susceptibility tensor may be represented geometrically by a triaxial ellipsoid, which is termed the magnitude ellipsoid, whose axes are parallel to the AMS tensor eigenvectors and whose semiaxes lengths are proportional to the AMS tensor eigenvalues (Figure 2). We will refer to this ellipsoid in the following as the AMS ellipsoid. Since for diamagnetic minerals the negative AMS eigenvalues make the magnitude ellipsoid difficult to visualize, the magnetic susceptibility ellipsoid may also be expressed by the representation quadric, in which the length of the radius vector in any direction from the origin is equal to the inverse square root of the susceptibility in that direction. It has however a less direct relationship to the eigenvalues.

In practical measurements, each AMS eigenvector is determined with two uncertainty angles, which define the regions where each principal susceptibility direction lies with a probability of 95% (Figure 3). The definition of a magnetic susceptibility ellipsoid identifies two important elements: a "magnetic foliation" (the plane orthogonal



Magnetic Anisotropy, Figure 2 The magnetic susceptibility ellipsoid. The anisotropy of magnetic susceptibility (AMS) tensor may be geometrically represented by a triaxial ellipsoid, in which the three orthogonal axes correspond to the AMS eigenvectors, k_{max}, k_{int}, and k_{min}, respectively. The orientation of the AMS eigenvectors is then defined in a reference system of Cartesian coordinates (X, Y, and Z).

to the direction of minimum magnetic susceptibility) and a "magnetic lineation" (the direction of maximum magnetic susceptibility).

AMS ellipsoid shapes are classified according to the relationships between the magnetic susceptibility eigenvalues:

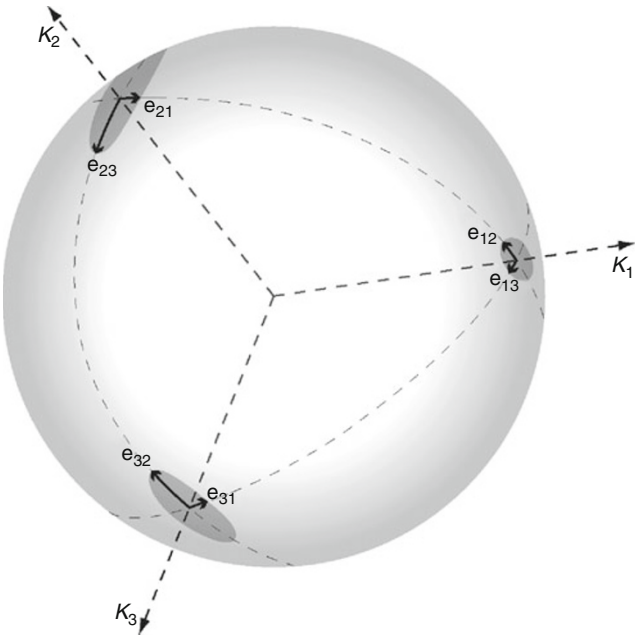
$k_1 \approx k_2 \approx k_3$; isotropic susceptibility, the AMS ellipsoid is a sphere (Figure 4a)

$k_3 \ll k_2 \approx k_1$; the AMS ellipsoid has an oblate shape (i.e., the magnetic fabric is planar, Figure 4b)

$k_3 \gg k_2 \approx k_1$; the AMS ellipsoid has a prolate shape (i.e., the magnetic fabric is linear, Figure 4c)

$k_1 > k_2 > k_3$; the AMS ellipsoid is triaxial (Figure 4d)

The orientations of principal susceptibilities for a set of samples are conventionally visualized in Schmidt equal-area projection, lower hemisphere (Figure 5).



Magnetic Anisotropy, Figure 3 Relationship of the 95% confidence ellipses (e_{12}, e_{23}, e_{13}) to the anisotropy of magnetic susceptibility (AMS) eigenvectors. Around each principal susceptibility axis, the major and minor semiaxes of 95% confidence ellipse lay within the planes defined by the eigenvectors.

The bulk low-field magnetic susceptibility of a rock is a summation of the contribution from all mineral species composing the rock, weighted according to their relative abundance and susceptibilities. Obviously, the contribution of ferromagnetic minerals to the overall magnetic susceptibility of a rock specimen is overwhelming upon those of the paramagnetic and diamagnetic matrix. The contribution of the matrix will be negligible when the percentage of ferromagnetic minerals exceeds 0.1% ($k >$ 3,000–4,000 μSI). Conversely, for rocks in which $k <$ 200–300 μSI, the magnetic susceptibility is almost completely controlled by the contribution of the paramagnetic matrix. For very small or even negative k values (i.e., $k < 50$ μSI), the contribution of the diamagnetic matrix to the bulk susceptibility of a rock specimens is not negligible.

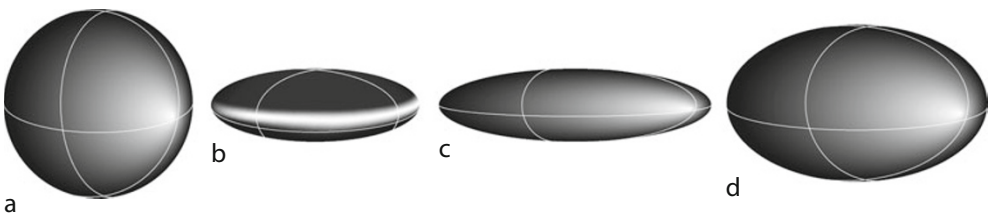
All the rock's constituent mineral fractions contribute to the development of an AMS fabric depending on their intrinsic susceptibility and degree of preferential orientation-distribution.

Among ferromagnetic minerals, hematite and pyrrhotite are the most anisotropic, both with anisotropy degree $P > 100$; whereas for magnetite P shows a typical range of variation between 1.1 for nearly isometric grains and 3 for highly anisometric grains (Table 1). Paramagnetic minerals are characterized by a much smaller AMS degree. In biotite and most phyllosilicates, the k_{max} and k_{min} directions are respectively parallel and orthogonal to the basal plane and their AMS ellipsoid is typically oblate. Finally, the main diamagnetic minerals typically show low magnetic susceptibilities and low anisotropy (Table 1).

In comparison to other methods of petrofabric analysis, the study of AMS has the advantage to be quick, cost-effective, highly sensitive, and above all it can be performed systematically on all rock types. AMS data are usually employed to obtain some information about the geological processes that gave a fabric to a rock unit, both during lithogenesis and during any subsequent deformational event.

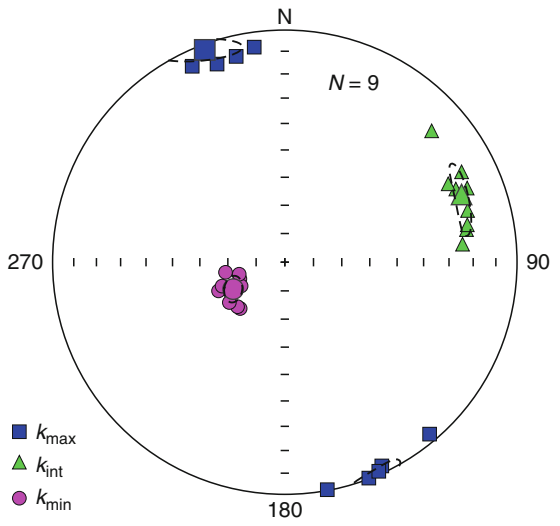
AMS: applications in geology and geophysics

Magnetic anisotropy is an essential petrostructural phenomenon which can be used to study rock fabric, deformation, and various other processes that took place during the geologic history of a rock. The AMS of rocks can be



Magnetic Anisotropy, Figure 4 Shape of anisotropy of magnetic susceptibility (AMS) ellipsoids. (**a**) spherical, (**b**) oblate, (**c**) prolate, (**d**) triaxial.

determined with high accuracy and the method is so sensitive that in rocks with a very weak preferred orientation-distribution of minerals it is the only approach that gives reasonable results. The study of AMS has been therefore used as a tool of structural analysis for almost all rock types and AMS analyses were applied to investigate a wide variety of geological processes capable to produce a preferential orientation-distribution of minerals in rocks. Moreover, AMS studies are an important complement to paleomagnetic researches, as a way to assess the reliability of paleomagnetic data and to determine possible deflections in the natural remanent magnetization components induced by deformation and compaction. For these reasons, the analysis of the AMS has experienced broad use in many branches of the Earth Sciences.



Magnetic Anisotropy, Figure 5 Equal area projection of the principal anisotropy of magnetic susceptibility (AMS) directions and the 95% confidence ellipses for a single sampling site. Squares: k_{max}; triangles: k_{int}; circles: k_{min}. Small symbols indicate individual specimens and large symbols indicate the mean tensor.

A brief overview of the wide spectrum of potential applications of the AMS research is provided below.

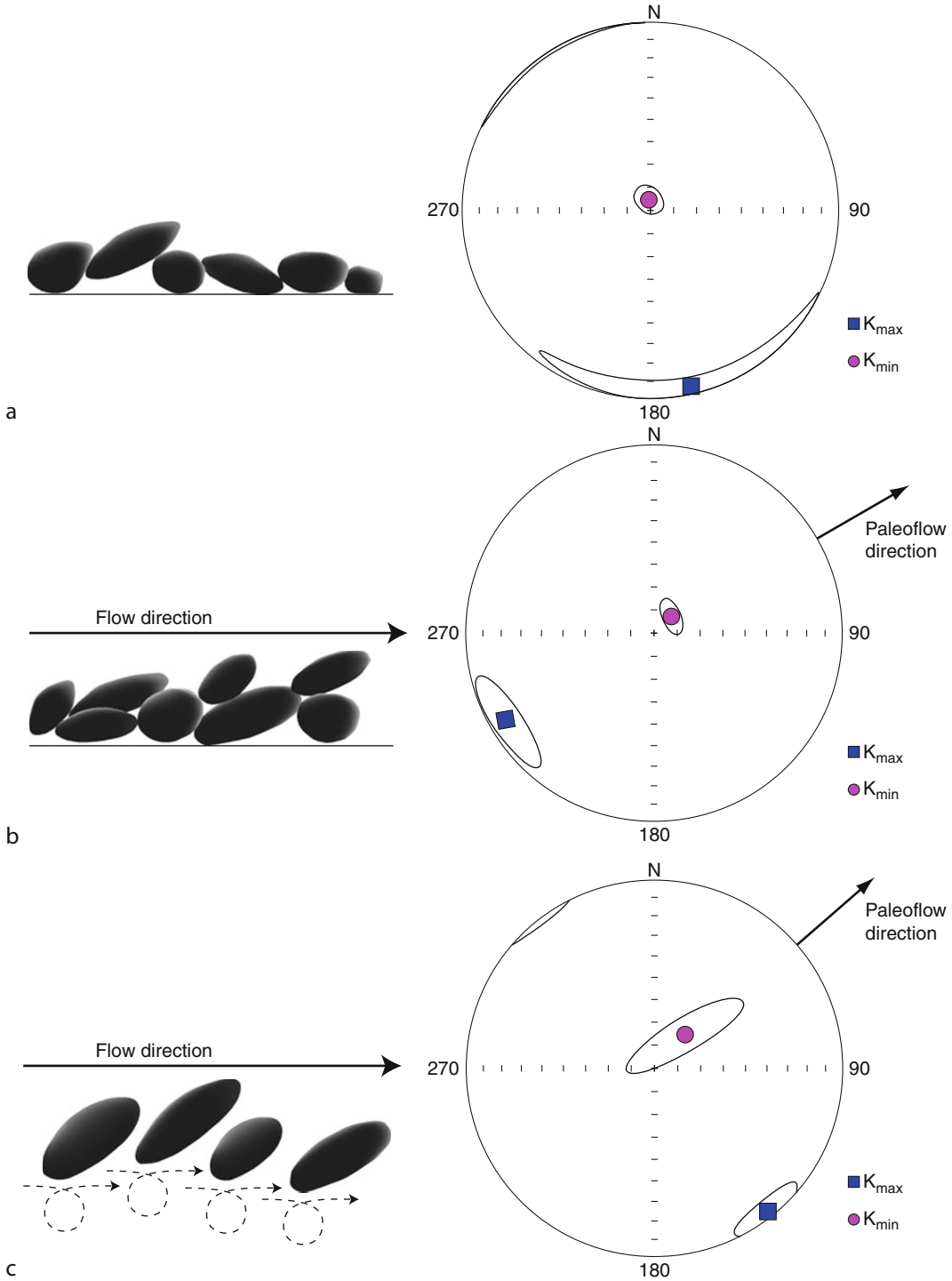
In undeformed *sediments*, the magnetic anisotropy may indicate the orientation of the sedimentation paleo-horizontal and eventual paleo-flow directions. The primary magnetic fabric of sediments is strongly affected by the processes that govern deposition and compaction. In quiet water deposition, such as in a lagoon or lake environment, gravity-driven sedimentation brings platy grains to lay with their longer dimensions statistically parallel to the bedding-compaction plane. Then, with further sediment burial, the effect of diagenetic compaction on platy minerals by pressure and water expulsion reinforces the parallelism between the magnetic foliation and the bedding plane.

Thus, if the magnetic fabric is purely depositional, or related only to compaction loading, the AMS ellipsoid shows an oblate shape, with the k_{min} axis perpendicular to the bedding plane (which can be determined through AMS analyses with high accuracy) and the k_{int} and k_{max} axes scattered in the bedding plane itself (Figure 6a). The action of flow currents may cause the maximum susceptibility axis to be aligned either parallel or perpendicular

Magnetic Anisotropy, Table 1 Grain anisotropy of magnetic susceptibility (AMS) determinations reported for various common minerals (Adapted from Hrouda, 1993, 2007; Tarling and Hrouda, 1993; Borradaile and Jackson, 2004)

Mineral	k (SI)	P	T	AMS type
Ferromagnetic (s.l.)				
Magnetite	2.0–5.7	1.1 – 3.0	Variable	Shape
Hematite	$1 \times 10^{-3} \div 5 \times 10^{-2}$	>100	~1	Magnetocrystalline
Pyrrhotite	$5 \times 10^{-2} \div 3 \times 10^{-1}$	100 – 10,000	~1	Magnetocrystalline
Paramagnetic				
Biotite	$1.0 \div 2.8 (\times 10^{-3})$	1.2 ÷ 1.6	0.9 ÷ 1	Magnetocrystalline
Muscovite	$0.12 \div 0.17 (\times 10^{-3})$	1.3 ÷ 1.4	0.4 ÷ 0.7	Magnetocrystalline
Phlogopite	$0.3 \div 1.2 (\times 10^{-3})$	1.3 ÷ 1.5	0.78 ÷ 0.95	Magnetocrystalline
Chlorite	$0.07 \div 1.55 (\times 10^{-3})$	1.2 ÷ 1.7	0.3 ÷ 0.7	Magnetocrystalline
Hornblende	$8.92 (\times 10^{-3})$	1.67	−0.51	Magnetocrystalline
Tourmaline	$1.69 (\times 10^{-3})$	1.005	~1	Magnetocrystalline
Siderite	$3.98 (\times 10^{-3})$	1.56	−0.9	Magnetocrystalline
Diamagnetic				
Quartz	$-13.4 \div -15.4 (\times 10^{-6})$	1.01	–	Magnetocrystalline
Calcite	$-13.8 (\times 10^{-6})$	1.11	1	Magnetocrystalline
Aragonite	$-19 (\times 10^{-6})$	1.15	0.8	Magnetocrystalline

k = mean magnetic susceptibility
$P = k_1/k_3$ – degree of anisotropy
$T = (2\eta_2 - \eta_1 - \eta_3)/(\eta_1 - \eta_3)$ – shape parameter for the AMS ellipsoid (Jelínek, 1981) where $\eta_1 = \ln k_1$, $\eta_2 = \ln k_2$, $\eta_3 = \ln k_3$, $\eta = (\eta_1 + \eta_2 + \eta_3)/3$



Magnetic Anisotropy, Figure 6 Schematic representations of the anisotropy of magnetic susceptibility (AMS) development for undeformed sediments deposited in quiet conditions (**a**), under the action of a weak or moderate water currents (**b**), and under the action of strong currents, sufficient to entrain particles and to induce the rolling of elongated grains with their long axis perpendicular to the water flow (**c**). (Modified from Tauxe, 2005.)

to the paleocurrent direction (e.g., Rees, 1961, 1966, 1983; Hamilton and Rees, 1970; Rees and Woodall, 1975; Ellwood and Ledbetter, 1979; Liu et al., 2001; Pàres et al., 2007). In moderate currents, platy particles tend to imbricate and elongated particles tend to align with their greatest axis parallel to the flow. On the AMS results obtained from a group of specimens, the imbrication induces a slightly vertical offset of the mean k_{min} and the clustering of the k_{max} axes in a direction antiparallel to the paleo-flow (Figure 6b); the overall AMS fabric is still characterized by an oblate ellipsoid. In such a case, imbrication yields a better estimate of the flow direction than magnetic lineation, which may wander within the foliation plane and even be orthogonal to the flow direction. In fact, when deposition occurs under high current flow (traction transport mode of sedimentation) the distribution of k_{min} axes from various samples is streaked, the k_{max} axes cluster in a direction perpendicular to the flow (Figure 6c), and the fabric is characterized by prolate or triaxial AMS ellipsoids.

In general, sedimentary depositional fabrics are comparatively low in both susceptibility ($<10^{-4}$ SI) and in anisotropy ($P < 1.05$).

In *igneous rocks*, AMS analyses are mostly used to determine the orientation of the plane and flux direction of the magma flows. In these rocks, the viscosity of the magma plays a fundamental role for the genesis of a rock fabric. The movement/rotation of magnetic minerals is hindered when viscosity reaches a level that impedes any further ductile flow.

It is useful to distinguish the case of extrusive and shallow intrusive rocks (lava flows, ignimbrites, and dykes) from plutonic rocks. In volcanic rocks, the processes that induce a preferential orientation of the magnetic minerals (mostly titanomagnetites) are connected to the plane and direction of magma flow, both for lavas from effusive eruptions, for pyroclastic flows from explosive volcanism, and for dykes and other shallow tabular intrusive rocks (e.g., Khan, 1962; Incoronato et al., 1983; Knight et al., 1986; MacDonald and Palmer, 1990; Tauxe et al., 1998; Cañon-Tapia, 2004; Gurioli et al., 2005). Titanomagnetite grains are oriented in the flow plane with their greatest dimension parallel to the magma flow, so that for lava flows the magnetic foliation matches the (paleo)topographic surface, whereas for dikes or sills it develops parallel to the walls of the hosting rocks. The magnetic lineation is mostly parallel to the magma flow directions, even though perpendicular or oblique relationships can also be rarely found. The efficiency of the volcanic flow to impart a preferential orientation on magnetic grains is usually relatively low, causing rather low anisotropy degrees, comparable to those typical of undeformed sediments.

A random fabric may be an indication of turbulent motion, which for dikes and sills occurs in the case of injection along wide enough fractures and for ignimbrites and surges occurs at a short distance from the vent where the eruption energy causes a chaotic motion of the transported solid particles. A coherent fabric may develop however at farther distance from the eruption centers where the kinetic energy decreases and the particles deposit on the morphological paleo-surfaces.

As for undeformed sediments, an eventual imbrication of the magnetic grains may be used to reconstruct the sense of the flow. In fact, the normal fabric of pyroclastic flows is considered the one with the foliation plunging upflow, so that the flow direction is given by the azimuth of the foliation pole and the lineation is close to it. AMS analyses in volcanic rocks have been used to constrain the location of the eruption centers.

In plutonic rocks, the analysis of the AMS may provide information on the processes of magma injection, differentiation, and crystallization (e.g., Heller, 1973; Hrouda et al., 1971; Jover et al., 1989; Hrouda and Lanza, 1989; Bouchez, 1997). The AMS fabric may therefore be affected by the flow of the magma, the changes in its viscosity and composition, and the finite deformation it undergoes under regional stress before complete crystallization. Generally, the AMS of plutonic rocks whose magnetic fabric was produced through magma flow are characterized by a low anisotropy degree which overlaps with that of volcanic rocks and undeformed sediments, and the shape of their AMS ellipsoid ranges from oblate to prolate according to the local character of the magma flow; the magnetic foliation is parallel to the flow plane and the magnetic lineation is parallel to the flow direction.

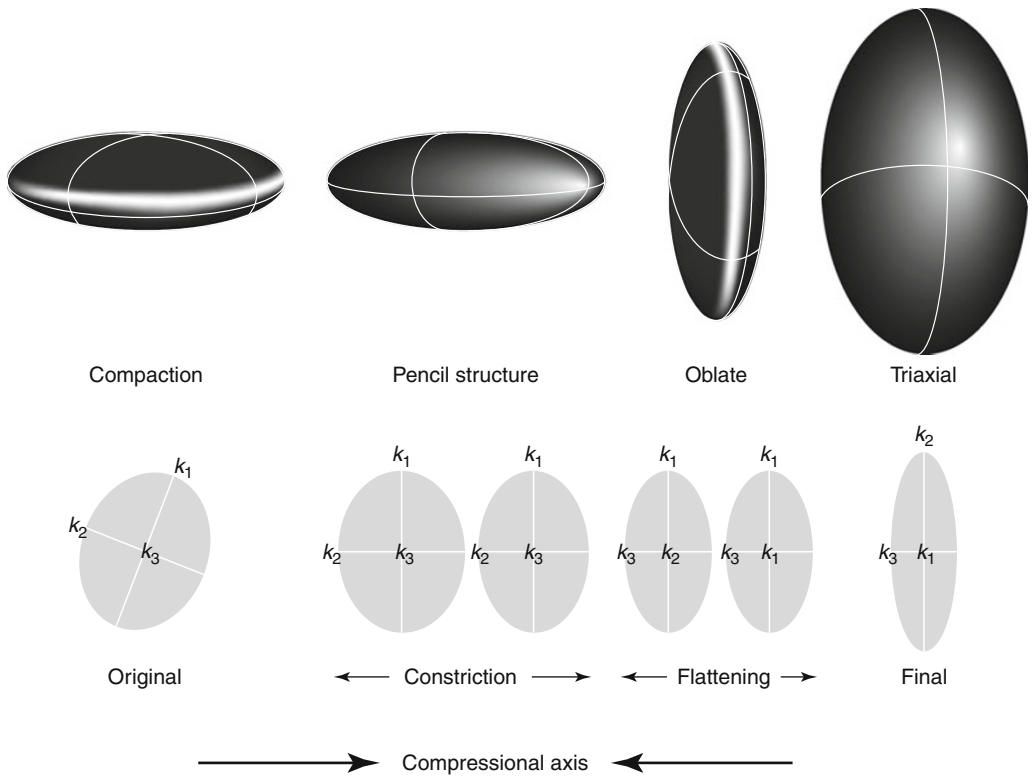
Conversely, the AMS of plutonic rocks which cooled under high temperature and pressures in the lower levels of the crust and are associated with metamorphic rocks may have a high anisotropy degree mostly determined by ductile deformation, which is a relatively efficient mechanism for the reorientation of magnetic minerals. The magnetic fabric in these rocks usually indicates a preferred orientation of magnetite minerals that nucleated at a late stage of the geological history of the pluton, and are therefore younger than the surrounding silicates. These magnetite grains grew in highly anisometric intergranular spaces and show therefore a high degree of shape anisotropy. This magnetic fabric is usually termed as mimetic magnetic fabric.

In *deformed rocks*, AMS analyses are used to evaluate the orientation of the strain ellipsoid (e.g., Borradaile, 1991; Hrouda, 1993; Borradaile and Henry, 1997; Borradaile and Jackson, 2004). In fact, AMS analysis is a very sensitive tool for the detection of rock strain, even in weakly deformed sediments and where other strain markers are scarcely developed or even totally absent.

AMS studies in nearly undeformed sediments have also been used as a tool for recognizing intervals affected by significant coring-induced deformation and assess the reliability of paleomagnetic data derived from such sediments (e.g., Rosenbaum et al., 2000).

AMS is extremely sensitive to incipient strain and it develops well before other macro- and mesoscopic strain features, such as cleavage, can be observed. As a matter of fact, the preferred orientation of phyllosilicate grains

in mudstones may reflect the strain produced by compaction and tectonic processes. If there is a stress field acting on a sediment, causing progressive strain, the primary magnetic fabric is modified according to the nature and the extent of the deformation (e.g., Graham, 1966; Lee et al., 1990; Sagnotti and Speranza, 1993; Sagnotti et al., 1998; Parés et al., 1999; Parés, 2004) (Figure 7). The first effect of strain upon such undeformed-compactional AMS fabric is to induce the clustering of the k_{max} axes in the direction of maximum stretching, which is roughly perpendicular to the maximum compressive stress. The rotation of platy grains around the intermediate strain axis makes the planar bedding fabric less pronounced. Anyway, at such very first stage of deformation the k_{min} axes are still normal to the bedding. AMS studies carried in clayey and weakly deformed sediments pointed out a significant relationship between the magnetic fabric of apparently undeformed clays and the tectonic history of the sedimentary basins in which they were deposited, including compressional as well as extensional tectonic settings (e.g., Sagnotti et al., 1994; Mattei et al., 1997). In such sediments the magnetic foliation generally indicates the sedimentation-compaction plane, whereas the magnetic lineation is perpendicular to the shortening direction in the compressional settings and parallel to the stretching direction in the extensional settings. The magnetic lineation found in macroscopically undeformed fine-grained sediments from extensional settings develops from the spatial distribution of the basal planes of the paramagnetic phyllosilicates of the clay matrix and is parallel to the common axis of differently oriented basal planes (Cifelli et al., 2004, 2005). A further increase of strain under compression causes the dispersion (scattering) of the k_{min} axes in such a way that they tend to describe a girdle normal to the k_{max} cluster: the AMS ellipsoid becomes prolate and the primary depositional fabric is progressively obliterated and overprinted by a tectonic fabric (Figure 7). Mineral rotations are sufficient at this stage to produce a fabric strong enough to impart the well-known "pencil structure." Further increasing strain occurs entering in the field of metamorphism and leads to the progressive return of the AMS ellipsoid into the flattening field. Mineral rotation leads to the production of an imperfectly formed cleavage crossing bedding, called embryonic cleavage. A cleavage stage is achieved when the tectonic strain imprint is sufficiently strong to form a dominant planar cleavage fabric, where platy and acicular minerals are oriented in this cleavage plane. The magnetic foliation plane that develops at this stage is parallel to the plane of cleavage or schistosity visible in the rock. With increasing strain, the planar cleavage and lineation are intensified, this last one very close to or parallel to the strain maximum stretching direction (Figure 7).



Magnetic Anisotropy, Figure 7 Schematic representations of the sequence of anisotropy of magnetic susceptibility (AMS) development in progressively deformed sediments in a compressional setting. (Redrawn after Graham, 1966.)

The AMS of *metamorphic rocks* is generally considerably higher than the AMS in undeformed sedimentary and volcanic rocks. As a matter of fact, the mechanisms of ductile deformation and recrystallization operating during metamorphism (e.g., diffusion, crystal plasticity, particulate flow, neomineralization) may be very effective to determine a preferential orientation-distribution of magnetic grains. In rocks with strong tectonic fabrics the principal directions of susceptibility and of finite strain coincide and a quantitative genetic relationship between strain and AMS has been pointed out in some cases (see reviews by Borradaile, 1988 and Borradaile and Jackson, 2004). Nevertheless such relationships are not universal and strain-AMS relationship or correlations found under certain circumstances and lithologies cannot be extrapolated to different situations.

A magnetic fabric is defined normal when the principal magnetic axes correspond one by one to the petrofabric axes. An inverse or intermediate magnetic fabric occurs when the magnetic susceptibility principal axes exchange their position. Inverse AMS fabrics are associated with the prevalence of certain magnetic minerals (see Rochette et al., 1992, 1999). Specifically, some minerals posses an inverse AMS fabric related to their shape (e.g., goethite and tourmaline) or to intracrystalline gliding during ductile deformation (e.g., siderite and other iron-bearing carbonates). Moreover, SD ferrimagnetic particles with prolate shape anisotropy may originate an inverse magnetic fabric as well. As a matter of fact, pseudo single domain (PSD) and multidomain (MD) particles of a ferrimagnetic mineral such as magnetite have a maximum susceptibility parallel to their long (easy) morphological axis, while a stable single domain (SD) magnetite particle, whose magnetization is always saturated along its long axis, has a maximum susceptibility value perpendicular to it (Stephenson et al., 1986; Potter and Stephenson, 1988). Even if inverse and anomalous AMS fabric are relatively rare in nature, in all cases when AMS directions cannot be correlated with obvious structural markers, the mineralogical source of the AMS should be closely investigated and the AMS compared with other types of magnetic anisotropy (i.e., the anisotropy of artificially induced magnetic remanences).

Anisotropy of the magnetic remanence (AMR)

AMS represents the sum of the susceptibility anisotropies of all the mineral components in a rock, including the diamagnetic and paramagnetic fractions. The studies on the AMR are directly relevant to paleomagnetic studies (as early recognized by Fuller, 1960, 1963), since they only measure the anisotropy of the remanence carrying particles. AMR studies also have a large spectrum of application in geological sciences (e.g., see reviews by Jackson, 1991; Potter, 2004; Hirt, 2007).

In fact, AMS may be inappropriate for the purposes of correcting paleomagnetic data, since it depends critically on the size of the remanence carrying particles. One of the advantages of AMR is that it precludes the effect of inverse AMS fabric due to single domain magnetites (Stephenson et al., 1986), since the grain long axis is the axis of maximum remanence irrespective of particle size. The most common AMRs are anisotropy of anhysteretic remanence (AARM), anisotropy of isothermal remanence magnetization (AIRM), and high-field magnetic anisotropy (HFAMS).

AARM: anisotropy of anhysteretic remanence

The anisotropy of anhysteretic remanence (AARM) is the most commonly studied anisotropy of remanence. An anhysteretic remanent magnetization (ARM) is acquired by a sample when it is simultaneously exposed to an alternating magnetic field with decreasing amplitude (usually with a maximum peak in the range 10–100 mT) and a steady DC magnetic field (typically values close to the intensity of the geomagnetic field, e.g., about 50–100 μT). Prior to acquisition of the ARM, the specimen should be demagnetized in an AF peak higher than that used for the ARM acquisition, and its residual remanence measured as a baseline. Then the ARM is imparted and computed as the difference between the measured remanence and the baseline.

There are two principal advantages of using ARM for the characterization of the anisotropy of the remanence: (1) the acquisition curve of ARM vs DC field is expected to be linear in the range of low DC magnetic field commonly employed, (2) a window of AC field can be used to impart ARM. In fact, by measuring different windows of coercivity (partial AARM), one can expect to characterize different subpopulations (size and shape) of ferromagnetic grains.

The ferromagnetic particles that define the AARM tensor may have a broad range of grain sizes.

Since there is a strong empirical correlation between partial anhysteretic remanence (pARM) and magnetite grain size (e.g., Dunlop and Özdemir, 1997), it is possible to isolate the AARM tensor associated with distinct populations of grains by using partial rather than total anhysteretic remanence anisotropy. This can be done by imparting a magnetization to a specific particle fraction during anisotropy determinations. In this regard, Jackson et al. (1988) first documented multiple AARM magnetic fabrics for the same sample by using different windows of coercive field. They called this partial anisotropy of anhysteretic remanent magnetization (pAARM).

Observed AARM indicates possible anisotropy in acquiring NRM. This provides a warning that the rock might not be an accurate paleomagnetic recorder. Measuring the AARM thus provides a fast and reliable method to correct paleomagnetic deviation in rocks. In sedimentary rocks, AARM has been proposed to provide a quantitative correction for inclination error (e.g., Jackson et al., 1991; Kodama and Sun, 1992; Deamer and Kodama, 1990; Collombat et al., 1993; Hodych et al., 1999; Tan and Kodama, 2002; Tan et al., 2003; Vaughn et al., 2005),

and to correctly evaluate the amplitude of paleosecular variation recorded in volcanic rocks (e.g., Gattacceca and Rochette, 2002).

AIRM: anisotropy of isothermal remanent magnetization

Another type of magnetic anisotropy is the anisotropy of isothermal remanent magnetization (AIRM) (Jelínek, 1966). In order to study the IRM anisotropy, a sample is successively subjected to a pulse magnetic field (with intensity generally of the order of 5–60 mT) in several appropriate directions and the acquired IRM is measured. As for the study of the anisotropy of ARM, before the measuring process and after each magnetization step the sample is AF demagnetized. The determination of the IRM anisotropy makes it possible to study the ferromagnetic fraction of a rock separately from the contribution of the paramagnetic and diamagnetic matrix.

Although usually successful, AIRM risks nonlinear susceptibility effects where strong magnetic field is required.

HFAMS: high-field magnetic anisotropy

High-field methods are used to separate the AMS due to ferromagnetic minerals from the AMS due to diamagnetic and paramagnetic minerals (e.g., Hrouda and Jelínek, 1990; Martn-Hernandez and Hirt, 2001). These methods are based on the field dependence of the magnetic susceptibility and rely on the different response of ferromagnetic and paramagnetic minerals to large applied fields, that is, large enough to saturate the ferromagnetic mineral fraction. For rocks that have both ferromagnetic and paramagnetic minerals, changes in magnetization at fields greater than ferromagnetic saturation are due primarily to the paramagnetic susceptibility.

The high-field magnetic anisotropy measurements have been made using torque magnetometers (e.g., Banerjee and Stacey, 1967; Bergmüller et al., 1994) or a high-field, rotating sample, cryogenic magnetometer (Rochette and Fillion, 1988). Although these techniques have great potential to separate paramagnetic and ferromagnetic fabrics, torque magnetometers and high-field cryogenic magnetometers are not available in most magnetic laboratories and HFAMS is not routinely applied.

Summary

Magnetic anisotropy is an essential property of rocks and minerals.

The AMS of rock specimens can be determined with a high degree of precision and the method is sensitive, rapid, and nondestructive. The study of AMS is a fundamental tool of petrofabric analysis, which has the potential to recognize even subtle preferred orientations of mineral lattices or shapes and to infer the causative processes, ranging from the effects of paleocurrents in undeformed sediments to pervasive strain in deformed rocks. As such, it has a wide range of applicability in the Earth Sciences.

The AMR instead refers to the preferential orientation-distribution of the ferromagnetic grains only. The techniques to determine AMR are time consuming because artificial remanences must be applied and then removed along each measurement axis. The study of AMR is more directly relevant to paleomagnetic researches since it may provide a tool to correct paleomagnetic deviations in all cases when the direction of the magnetic remanence is deflected from the direction of the magnetic field in which it was acquired.

Bibliography

Banerjee, S. K., and Stacey, F. D., 1967. The high-field torque-meter method of measuring magnetic anisotropy of rocks. In Collinson, D. W., Creer, K. M., and Runcorn, S. K. (eds.), *Methods in Paleomagnetism*. Amsterdam: Elsevier, pp. 470–476.

Bergmüller, F., Barlocher, C., Geyer, B., Grieder, M., Heller, F., and Zweifel, P., 1994. A torque magnetometer for measurements of the high-field anisotropy of rocks and crystals. *Measurement Science and Technology*, **5**, 1466–1470.

Berkowitz, A. E., and Takano, K., 1999. Exchange anisotropy – a review. *Journal of Magnetism and Magnetic Materials*, **200**, 552–570.

Borradaile, G. J., 1991. Correlation of strain with anisotropy of magnetic susceptibility (AMS). *Pure and Applied Geophysics*, **135**, 15–29.

Borradaile, G. J., 1988. Magnetic susceptibility, petrofabrics and strain. *Tectonophysics*, **156**, 1–20.

Borradaile, G. J., 2001. Magnetic fabrics and petrofabrics: their orientation distribution and anisotropies. *Journal of Structural Geology*, **23**, 1581–1596.

Borradaile, G. J., and Henry, B., 1997. Tectonic applications of magnetic susceptibility and its anisotropy. *Earth Science Reviews*, **42**, 49–93.

Borradaile, G. J., and Jackson, M., 2004. Anisotropy of magnetic susceptibility (AMS): magnetic petrofabrics of deformed rocks. In Martín-Hernández, F., Lüneburg, C., Aubourg, C., and Jackson, M. (eds.), *Magnetic Fabric Methods and Applications*. London: Geological Society. Special Publications 2004, Vol. 238, pp. 299–360.

Bouchez, J. L., 1997. Granite is never isotropic: an introduction to AMS studies of granitic rocks. In Bouchez, J. L., Hutton, D. H. W., and Stephens, W. E. (eds.), *Granite: from Segregation of Melt to Emplacement Fabrics.* Dordrecht: Kluwer, pp. 95–112.

Cañón-Tapia, E., 2004. Anisotropy of magnetic susceptibility of lava flows and dykes: an historical account. In Martín-Hernández, F., Lüneburg, C., Aubourg, C., and Jackson, M. (eds.), *Magnetic Fabric Methods and Applications*. London: Geological Society, Special Publications 2004, Vol. 238, pp. 205–225.

Cifelli, F., Mattei, M., Hirt, A. M., and Gunther, A., 2004. The origin of tectonic fabrics in "undeformed" clays: The early stages of deformation in extensional sedimentary basins. *Geophysical Research Letters*, **31**, L09604, doi:10.1029/2004GL019609.

Cifelli, F., Mattei, M., Chadima, M., Hirt, A. M., and Hansen, A., 2005. The origin of the tectonic lineation in extensional basins: Combined neutron texture and magnetic analysis on "undeformed" clays. *Earth and Planetary Science Letters*, **235**, 62–78.

Collombat, H., Rochette, P., and Kent, D. V., 1993. Detection and correction of inclination shallowing in deep sea sediments

using the anisotropy of magnetic remanence. *Bulletin. Société Géologique de France*, **164**, 103–111.
Deamer, G. A., and Kodama, K. P., 1990. Compaction-induced inclination shallowing in synthetic and natural clay-rich sediments. *Journal of Geophysical Research*, **95**, 4511–4529.
Dunlop, D. J., and Özdemir, Ö., 1997. *Rock Magnetism: Fundamentals and Frontiers*. Cambridge: Cambridge University Press. 573 pp.
Ellwood, B. B., and Ledbetter, M. T., 1979. Paleocurrent indicators in deep-sea sediment. *Science*, **203**, 1335–1337.
Fuller, M. D., 1960. Anisotropy of susceptibility and the natural remanent magnetization of some Welsh slates. *Nature*, **186**, 790–792.
Fuller, M. D., 1963. Magnetic anisotropy and paleomagnetism. *Journal of Geophysical Research*, **68**, 293–309.
Gattacceca, J., and Rochette, P., 2002. Pseudopaleosecular variation due to remanence anisotropy in a pyroclastic flow succession. *Geophysical Research Letters*, **29**, doi:10.10129/2002GL014697.
Graham, J. W., 1954. Magnetic susceptibility anisotropy, an unexploited petrofabric element. *Bulletin of the Geological Society of America*, **65**, 1257–1258.
Graham, J. W., 1966. Significance of magnetic anisotropy in Appalachian sedimentary rocks. In Steinhart, J. S., and Smith, T. J. (eds.), *The Earth Beneath the Continents. Geophysical Monograph Series 10*. Washington, DC: American Geophysical Union, pp. 627–648.
Gurioli, L., Pareschi, M. T., Zanella, E., Lanza, R., Deluca, E., and Bisson, M., 2005. Interaction of pyroclastic density currents with human settlements: evidence from ancient Pompeii. *Geology*, **33**, 441–444, doi:10.1130/G21294.1.
Hamilton, N., and Rees, A. I., 1970. The use of magnetic fabric in palaeocurrent estimation. In Runcorn, S. K. (ed.), *Palaeogeophysics*. London: Academic, pp. 445–463.
Heller, F., 1973. Magnetic anisotropy of granitic rocks of the Bergell massif (Switzerland). *Earth and Planetary Science Letters*, **20**, 180–183.
Hirt, A. M., 2007. Magnetic remanence, anisotropy. In Gubbins, D., and Herrero-Bervera, E. (eds.), *Encyclopedia of Geomagnetism and Paleomagnetism*, New York: Springer, 1054pp, pp. 535–540.
Hodych, J. E., Bijaksana, S., and Pätzold, R., 1999. Using magnetic anisotropy to correct for paleomagnetic inclination shallowing in some magnetite-bearing deep-sea turbidites and limestones. *Tectonophysics*, **307**, 191–205.
Hrouda, F., 1982. Magnetic anisotropy of rocks and its application in geology and geophysics. *Geophysical Surveys*, **5**, 37–82.
Hrouda, F., 1993. Theoretical models of magnetic anisotropy to strain relationship revisited. *Physics of the Earth and Planetary Interiors*, **77**, 237–249.
Hrouda, F., 2007. Magnetic susceptibility, anisotropy. In Gubbins, D., and Herrero-Bervera, E. (eds.), *Encyclopedia of Geomagnetism and Paleomagnetism*, New York: Springer, 1054pp, pp. 546–560.
Hrouda, F., Chlupacova, M., and Rejl, L., 1971. The mimetic fabric of magnetite in some foliated granodiorites, as indicated by magnetic anisotropy. *Earth Science and Planetary Interiors*, **11**, 381–384.
Hrouda, F., and Lanza, R., 1989. Magnetic anisotropy in the Biella and Traversella stocks (Periadriatic Line): implications for the emplacement mode. *Studia Geophysica et Geodaetica*, **56**, 337–348.
Hrouda, F., and Jelínek, V., 1990. Resolution of ferrimagnetic and paramagnetic anisotropies in rocks, using combined low-field and high-field measurements. *Geophysical Journal International*, **103**, 75–84.
Incoronato, A. F. T., Addison, D. H., Tarling, G. N., and Pescatore, T., 1983. Magnetic fabric investigation of some pyroclastic deposits from the Phlegrean Fields, southern Italy. *Nature*, **306**, 461–463.
Ising, E., 1942. On the magnetic properties of varved clays. *Arkiv för Matematik, Astronomi och Fysik*, **29A**, 1–37.
Jackson, M., 1991. Anisotropy of magnetic remanence: a brief review of mineralogical sources, physical origins, and geological applications, and comparison with susceptibility anisotropy. *Pure and Applied Geophysics*, **136**, 1–28.
Jackson, M., Gruber, W., Marvin, J., and Banerjee, S. K., 1988. Partial anhysteretic remanence and its anisotropy: applications and grain size-dependence. *Geophysical Research Letters*, **15**, 440–443.
Jackson, M. J., Banerjee, S. K., Marvin, J. A., Lu, R., and Gruber, W., 1991. Detrital remanence, inclination errors and anhysteretic remanence anisotropy: quantitative model and experimental results. *Geophysical Journal International*, **104**, 95–103.
Jackson, M. J., and Tauxe, L., 1991. Anisotropy of magnetic susceptibility and remanence: developments in the characterization of tectonic, sedimentary, and igneous fabric. *Reviews of Geophysics*, **29**, 371–376.
Jelínek, V., 1981. Characterization to the magnetic fabric of rocks. *Tectonophysics*, **79**, 63–67.
Jelínek, V., 1996. Theory and measurement of the anisotropy of isothermal remanent magnetization of rocks. *Travaux Géophysique*, **37**, 124–134.
Jover, O., Rochette, P., Lorand, J. P., Maeder, M., and Bouchez, J. L., 1989. Magnetic mineralogy of some granites from the French Massif Central: origin of their low-field susceptibility. *Physics of the Earth and Planetary Interiors*, **55**, 79–92.
Khan, M. A., 1962. The anisotropy of magnetic susceptibility of some igneous and metamorphic rocks. *Journal of Geophysical Research*, **67**, 2873–2885.
Kodama, K. P., and Sun, W.-W., 1992. Magnetic anisotropy as a correction for compaction-caused paleomagnetic inclination shallowing. *Geophysical Journal International*, **111**, 465–469.
Knight, M. D., Walker, G. P. L., Ellwood, B. B., and Diehl, J. F., 1986. Stratigraphy, paleomagnetic, and magnetic fabric of the Toba Tuffs: constraints on the source and eruptive styles. *Journal of Geophysical Research*, **91**, 10,355–10,382.
Lanza, R., and Meloni, A., 2006. *The Earth's Magnetism. An Introduction for Geologists*. Berlin: Springer. 278 pp.
Lee, T. Q., Kissel, C., Laj, C., Horng, C. S., and Lue, Y. T., 1990. Magnetic fabric analysis of the Plio-Pleistocene sedimentary formations of the Coastal Range of Taiwan. *Earth and Planetary Science Letters*, **98**, 23–32.
Liu, B., Saito, Y., Yamazaki, T., Abdeldayem, A., Oda, H., Hori, K., and Zhao, Q., 2001. Paleocurrent analysis for the late Pleistocene-Holocene incised-valley fill of the Yangtze delta, China by using anisotropy of magnetic susceptibility data. *Marine Geology*, **176**, 175–189.
MacDonald, W. D., and Palmer, H. C., 1990. Flow directions in ash-tuffs: A comparison of geological and magnetic susceptibility measurements, Tshirege member (upper Bandelier Tuff), Valles caldera, New Mexico, USA. *Bullettin of Volcanology*, **53**, 45–59.
Martín-Hernández, F., and Hirt, A. M., 2001. Separation of ferrimagnetic and paramagnetic anisotropies using a high-field torsion magnetometer. *Tectonophysics*, **337**, 209–222.
Mattei, M., Sagnotti, L., Faccenna, C., and Funiciello, R., 1997. Magnetic fabric of weakly deformed clay-rich sediments in the Italian peninsula: relationship with compressional and extensional tectonics. *Tectonophysics*, **271**, 107–122.
Meiklejohn, W. H., 1962. Exchange anisotropy – a review. *Journal of Applied Physics*, **33**, 1328–1335.
Moskowitz, B. M., 1993. High-temperature magnetostriction of magnetite and titanomagnetites. *Journal of Geophysical Research*, **98**, 359–371.
Nagata, T., and Uyeda, S., 1959. Exchange interaction as a cause of reverse thermoremanent magnetism. *Nature*, **184**, 890.

O'Reilly, W., 1984. *Rock and Mineral Magnetism*. Glasgow: Blackie. 230 pp.
Parés, J.M., 2004. How deformed are weakly deformed mudrocks ? Insights from magnetic anisotropy. In Martín-Hernández, F., Lüneburg, C., Aubourg, C., and Jackson, M., (eds.), *Magnetic Fabric Methods and Applications*. London: Geological Society, Special Publications, Vol. 238, pp. 191–203.
Parés, J. M., van der Pluijm, B. A., and Dinares-Turell, J., 1999. Evolution of magnetic fabrics during incipient deformation of mudrock (Pyrenees, northern Spain). *Tectonophysics*, **307**, 1–14.
Parés, J. M., Hassold, N. J. C., Rea, D. K., and van der Pluijm, B. A., 2007. Paleocurrent directions from paleomagnetic reorientation of magnetic fabrics in deep-sea sediments at the Antarctic Peninsula Pacific margin (ODP Sites 1095, 1101). *Marine Geology*, **242**, 261–269.
Potter, D. K., 2004. A comparison of anisotropy of magnetic remanence methods—a user's guide for application to palaeomagnetism and magnetic fabric studies. In Martin-Hernández, F., Lüneburg, C. M., Aubourg, C., and Jackson, M. (eds.), *Magnetic Fabric: Methods and Applications*. London: The Geological Society of London. Geological Society Special Publications, Vol. 238, pp. 21–36.
Potter, D. K., and Stephenson, A., 1988. Single-domain particles in rocks and magnetic fabric analysis. *Geophysical Research Letters*, **15**, 1097–1100.
Rees, A. I., 1961. The effect of water currents on the magnetic remanence and anisotropy of susceptibility of some sediments. *Geophysical Journal*, **5**, 235–251.
Rees, A. I., 1966. The effect of depositional slopes on the anisotropy of magnetic susceptibility of laboratory deposited sands. *Journal of Geology*, **74**, 856–867.
Rees, A. I., 1983. Experiments on the production of transverse grain alignment in a sheared dispersion. *Sedimentology*, **30**, 437–448.
Rees, A. I., and Woodall, W. A., 1975. The magnetic fabric of some laboratory-deposited sediments. *Earth and Planetary Science Letters*, **25**, 121–130.
Rochette, P., and Fillion, C., 1988. Identification of multicomponent anisotropies in rocks using various field and temperature values in a cryogenic magnetometer. *Physics of the Earth and Planetary Interiors*, **51**, 379–386.
Rochette, P., Jackson, M. J., and Aubourg, C., 1992. Rock magnetism and the interpretation of anisotropy of magnetic susceptibility. *Reviews of Geophysics*, **30**, 209–226.
Rochette, P., Aubourg, C., and Perrin, M., 1999. Is this magnetic fabric normal? A review and case study in volcanic formations? *Tectonophysics*, **307**, 219–234.
Rosenbaum, J., Reynolds, R., Smoot, J., and Meyer, R., 2000. Anisotropy of magnetic susceptibility as a tool for recognizing core deformation: reevaluation of the paleomagnetic record of Pleistocene sediments from drill hole OL-92, Owens Lake, California. *Earth and Planetary Science Letters*, **178**, 415–424.
Sagnotti, L., and Speranza, F., 1993. Magnetic fabric analysis of the Plio-Pleistocene clayey units of the Sant'Arcangelo basin, southern Italy. *Physics of the Earth and Planetary Interiors*, **77**, 165–176.
Sagnotti, L., Faccenna, C., Funiciello, R., and Mattei, M., 1994. Magnetic fabric and structural setting of Plio-Pleistocene clayey units in an extensional regime: the Tyrrhenian margin of central Italy. *Journal of Structural Geology*, **16**, 1243–1257.
Sagnotti, L., Speranza, F., Winkler, A., Mattei, M., and Funiciello, R., 1998. Magnetic fabric of clay sediments from the external northern Apennines (Italy). *Physics of the Earth and Planetary Interiors*, **105**, 73–93.
Stephenson, A., Sadikun, S., and Potter, D. K., 1986. A theoretical and experimental comparison of the susceptibility and remanence in rocks and minerals. *Geophysical Journal of the Royal Astronomical Society*, **84**, 185–200.
Tan, X., and Kodama, K. P., 2002. Magnetic anisotropy and paleomagnetic inclination shallowing in red beds: evidence from the Mississippian Mauch Chunk Formation, Pennsylvania. *Journal of Geophysical Research*, **107**(B11), 2311, doi:10.1029/2001JB001636.
Tan, X., Kodama, K. P., Chen, H., Fang, D., Sun, D., and Li, Y., 2003. Paleomagnetism and magnetic anisotropy of Cretaceous red beds from the Tarim basin, northwest China: evidence for a rock magnetic cause of anomalously shallow paleomagnetic inclinations from central Asia. *Journal of Geophysical Research*, **108**, 2107, doi:10.1029/2001JB001608.
Tarling, D. H., and Hrouda, F., 1993. *The Magnetic Anisotropy of Rocks*. London: Chapman & Hall, p. 217.
Tauxe, L., 2005. Lectures in paleomagnetism, available on line at http://earthref.org/MAGIC/books/Tauxe/2005/
Tauxe, L., Gee, J. S., and Staudigel, H., 1998. Flow direction in dikes from anisotropy of magnetic susceptibility of magnetic data; the bootstrap way. *Journal of Geophysical Research*, **103** (B8), 17775–17790.
Vaughn, J., Kodama, K. P. & Smith, D., 2005. Correction of inclination shallowing and its tectonic implications: the Cretaceous Perforada Formation, Baja California, *Earth and Planetary Science Letters*, **232**, 72–82.

Cross-references

Curie Temperature
Magnetic Domains
Magnetic Methods, Principles
Paleomagnetism, Measurement Techniques and Instrumentation
Paleomagnetism, Principles
Remanent Magnetism

MAGNETIC ANOMALIES, INTERPRETATION

Erwan Thébault
Équipe de Géomagnétisme, Institut de Physique du Globe de Paris, Paris, France

Definition

Magnetic anomaly. On Earth, magnetic anomalies are defined as the vector or scalar difference between the measured magnetic field and an estimate of the main field. Magnetic anomalies are commonly assumed to reflect the magnetic properties of rocks lying in the Earth's crust.

Introduction

The Earth's magnetic field is the vector sum of fields of internal and external origins that operate on a wide range of spatial and temporal scales (Hulot et al., 2007). On Earth, the three major sources are the main or core field, which is generated by electrical currents in the liquid outer core, the lithospheric also referred to as crustal field, which results from magnetized materials in the outer layer of the Earth, and the external fields produced by currents in the ionosphere and the magnetosphere. An anomaly field approximates well the magnetic field of the Earth's lithosphere only if the main and external fields have been corrected to an accuracy comparable to the data noise.

When the magnetic field is known on the surface of a sphere, the magnetic potential field may be expressed in current free regions as the vector $B = -\nabla V(r,\theta,\phi,t)$ with the scalar potential V expanded in terms of Spherical Harmonics (SH)

$$V(r,\theta,\phi,t) = a\sum_{n=1}^{\infty}\sum_{m=0}^{n}\left(\frac{a}{r}\right)^{n+1}\left(g_n^m(t)\cos(m\phi) + h_n^m(t)\sin(m\phi)\right)P_n^m(\cos\theta) + a\sum_{n=1}^{\infty}\sum_{m=0}^{n}\left(\frac{r}{a}\right)^{n}\left(q_n^m(t)\cos(m\phi) + s_n^m(t)\sin(m\phi)\right)P_n^m(\cos\theta). \quad (1)$$

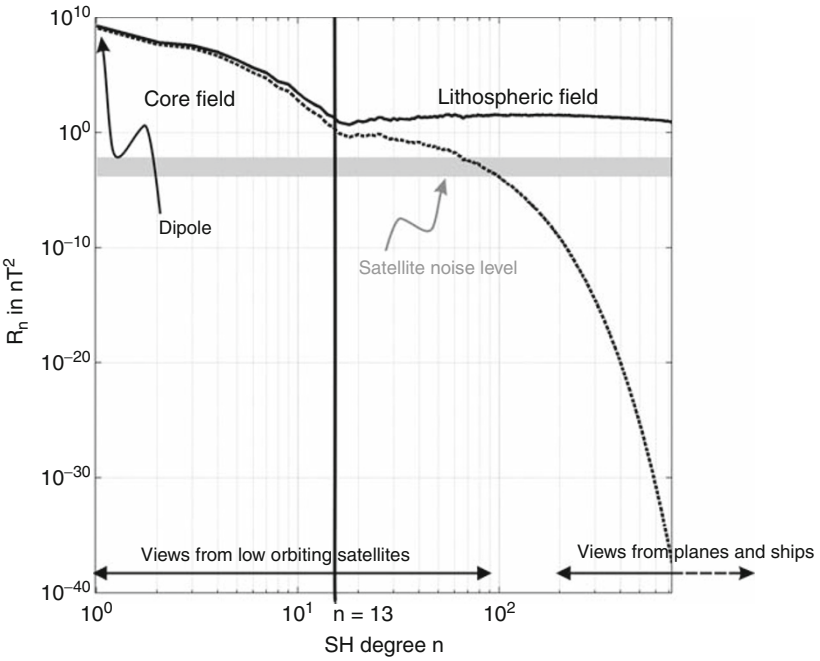
The parameters a, r, θ, and ϕ are the Earth's reference radius (6,371.2 km), the radius, the co-latitude, and the longitude, respectively. The internal (g_n^m, h_n^m) and external (q_n^m, s_n^m) coefficients may be time varying and are expressed in [nT] with respect to the shell of radius $r = a$. In geomagnetism, $P_n^m(\cos\theta)$ is the Schmidt semi-normalized associated Legendre function of integer degree n and order m. The spatial power spectrum of the internal field is defined by (Lowes, 1974)

$$R_n = (n+1)\left(\frac{a}{r}\right)^{2n+4}\sum_{m=0}^{n}\left[\left(g_n^m\right)^2 + \left(h_n^m\right)^2\right], \quad (2)$$

with R_n expressed in nT^2. The power spectrum provides an estimate of the contribution of each SH degree n to the square amplitude of the observed magnetic field of internal origin at the radius r. The horizontal spatial wavelength λ in kilometers [km] at the Earth's mean radius a associated with each degree n, is defined by (Backus et al., 1996, p. 101)

$$\lambda = \frac{2\pi a}{n}$$

Figure 1 displays the power spectrum of an internal model expanded to SH degree $n = 720$ (Maus, 2010). It is derived from satellite, airborne and ship measurements (*Geomagnetic Field, Measurement Techniques*). The break in slope of the power spectrum indicates that long-term dynamo processes (the core field and its secular variation; see *Geomagnetic Field, Secular Variation*) dominate the Earth's lithospheric field at least up to SH degree 13–15. The magnetic field of the Earth's core is thus not accurately represented by its dipole term only. Prior to epoch 2000.0, International Geomagnetic Reference Field (IGRF) models were derived to SH degree 10, and 13 thereafter (*Geomagnetic Field, IGRF*). When processing datasets collected at different epochs, disagreements between, or inaccuracies of, core field models lead to different anomaly field pictures. The geometrical attenuation of SH harmonics with increasing altitudes is important and represents the main impediment to the detection of the lithospheric field at all scales. The fast coalescence of small-scale anomalies with altitude demonstrates that different platforms of magnetic measurements at different altitudes do not detect the same



Magnetic Anomalies, Interpretation, Figure 1 Geomagnetic field power spectrum to SH degree 720 at 400 km altitude (*dashed line*) and (*solid line*) at 0 km above the Earth's mean radius (After Maus, 2010).

properties of the anomaly fields. Techniques for preparing an anomaly map of model at space or near-surface altitudes include removing the best estimates of fields not originating in the Earth's lithosphere, including their slow and rapid time variations, selecting suitable subsets of quiet-time data and averaging a large number of data to better highlight the finest permanent crustal field structures. A magnetic anomaly thus depends on altitude and on the care taken to remove all field contributions not originating from the lithosphere, which is strongly related to the advances made toward analyzing and understanding the Earth's magnetic field as a whole.

The sources of crustal and lithospheric fields

When the errors incurred in correcting the data for core and external fields are mitigated, the anomaly field is assumed to have sources in the crust (*Continental crustal structure*) or sometimes even deeper in the lithosphere (*Continental Lithosphere: Thermal Structure*).

The subject of rock magnetism is complex (Dunlop and Özdemir, 2007) but the general mechanisms are well known. The amount of magnetization carried by a rock depends on many parameters such as size, shape, mineralogy, domains, and temperature (*Magnetic Domains*). Most magnetic minerals on Earth are titanomagnetite and titanohematite, for which the Curie temperatures are between 580°C and 670°C (*Curie temperature*). Curie temperatures are met at depths of about 30 and 7 km, in stable continental and oceanic regions, respectively. Deeper rocks are considered nonmagnetic even though the Curie depth may penetrate the lithosphere. The minerals found in the Earth carry both permanent, $\mathbf{M}_r$ (see *Remanent Magnetism*) and induced magnetizations, $\mathbf{M}_i$. The magnetization $\mathbf{M}_r$, arises when a main field is locked into the material when, for example, the temperature of an igneous rock decreased below the Curie temperature of its magnetic phase. The magnetic field generated by remanent magnetization is a small fraction of the ancient magnetic field, whose direction may be very different from the contemporary field, both because of rock reorientation to their present positions and because of changes of the main field directions over geological times (see *Geomagnetic Field, Polarity Reversals*). The ease with which the magnetization of a material can be induced depends on the magnetic susceptibility k such that $\mu_0\mathbf{M}_i = k\mathbf{B}$, where $\mathbf{M}_i$ is the vector of induced magnetization, $\mathbf{B}$ the inducing ambient field, and μ_0 the magnetic permeability of free space. Susceptibility values vary with rock types. They are smaller for sedimentary than for metamorphic rocks, themselves lower than igneous rocks. In all cases, the induced magnetization produces a field much less intense than the inducing field. Since the Earth's core magnetic field does not vary significantly on a decadal timescale (*Geomagnetic Field, Secular Variation*), the induced lithospheric field is quasi-static over historical times. The total magnetization $\mathbf{M} = \mathbf{M}_i + \mathbf{M}_r$ thus points in a different direction from both the present and ancient fields. The Königsberger ratio M_r/M_i measures the relative importance of remanent to induced magnetization. Since we measure the magnetic fields resulting from magnetization rather than magnetization itself, it is difficult to resolve it into its induced and remanent components without collecting rock specimens and measuring their magnetic properties. The Königsberger ratio is generally large for oceanic rocks. Thanks to the rather homogeneous nature and structure of the oceanic crust (see *Lithosphere, Oceanic*), inferring the remanent magnetization over an entire oceanic chrons by interpolating the magnetic properties of a few distant samples is achievable (see *Paleomagnetism, Principles*). In contrast, it is almost impossible to predict the remament magnetization over continents. The Königsberger ratio is there anyway small because rock materials contain coarse magnetite grains. Common practice is thus to disregard remanent magnetization on continental areas and to consider induced magnetization negligible on oceans.

Important theorems show that not all magnetization distributions, in particular the uniform distribution (Runcorn, 1975) or contrasts oriented North–south near the Equator, produce magnetic fields above the Earth's surface. The spatial variation of the crustal magnetic field reflects lateral contrasts of magnetization rather than magnetization itself. Two adjacent geological units having both strong but comparable susceptibilities may produce magnetic field variations comparable to a body with lower magnetization but surrounded by sediments, for instance. Therefore, obvious outcropping geological features do not necessarily have a detectable magnetic signature and interpretations of magnetic maps are nonunique.

Measuring magnetic anomalies from planes and ships

Inland measurements are available but cover very small areas only (a few tens of meter squares). Near-surface surveys are routinely undertaken for mineral and hydrocarbon exploration and sometimes for basic research. Airborne magnetic surveys are generally carried out to support mineral and petroleum exploration and aid in subsurface geological mapping (*Magnetic Methods: Airborne*). These surveys are generally executed in a regular pattern at some hundred meters altitude above mean terrain clearance, with a horizontal distance about twice as much as the aircraft altitude (Reeves, 2005). The flight lines are oriented perpendicular to the geological strike directions to detect better the magnetic contrasts. For reconnaissance mapping of large regions (e.g., states, countries) typical line spacing is 1-km kilometer at altitudes above the sea level ranging between 1 and 5-km. More widely spaced segments crossing the flight lines at right angles are then used to correct the scalar data for offsets caused, either by magnetic field time variations, or by measurement errors. Local and regional surveys do not detect the same magnetic anomalies (Figure 1). Low altitude surveys provide locally detailed profiles but little

information about the magnetic structures exceeding the survey dimension. Likewise, high altitude surveys provide a synoptic view of the magnetic anomalies but generally miss the features with length scales smaller than the altitude and larger than the survey area. Near-surface measurements are carried out during days. Marine and aeromagnetic surveys provide raw total field intensity data and more rarely vector measurements because the reference frame is not sufficiently stable.

The anomaly field is defined by correcting the raw data for a main field model, in general an International Geomagnetic Reference Field model (*Geomagnetic Field, IGRF*), or simply a regional trend, and then for external field variations that occurred during the survey. Current practice is to use the magnetic records of magnetic observatories or a nearby ground station purposefully set up for the survey to monitor the Earth's magnetic field, to identify the disturbances and eliminate spikes, and to correct ultimately for the diurnal magnetic field variation. The final product is a map of anomaly values, reaching up to a few thousands nanotesla in amplitude, built by interpolation between data points and leveled to a common altitude onto a regular grid. An anomaly map build upon the compilation of various surveys displays crustal features with typical spatial scales from kilometers to hundred kilometers (Figure 1). They mostly highlight the magnetic properties of the upper crust.

Satellite measurements

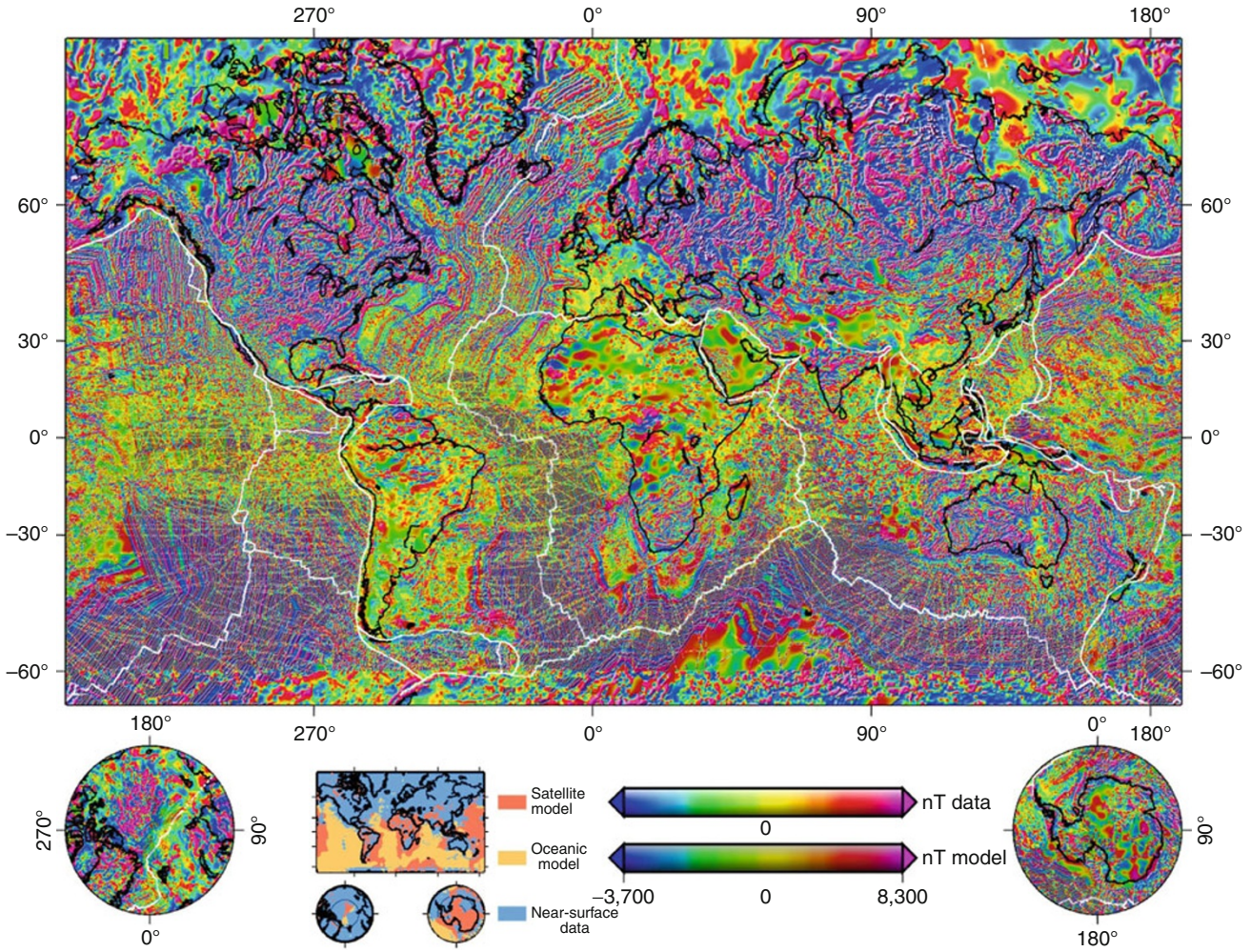
Large geologic structures with dimensions of a few hundred kilometers can be inferred from Low Earth Orbiting (LEO) satellites. Satellites provide scalar and vector homogeneous measurements at the Earth's scale (*Magnetic methods, satellite*). The lithospheric field is one of the weakest detectable source at the satellite altitudes. It represents less than 0.1% of the full signal, on average (Figure 1), and magnetic anomalies rarely exceed 20 nT at 400-km altitude (*Magnetic Methods, Satellite*). Satellites are also closer to (and sometimes fly through) the ionospheric sources and are more sensitive to the external field than the ground-based measurements. Measurements of satellites may be exploited to co-estimate the core and the lithospheric field in SH. This considerably reduces the error introduced by correcting the data with an independent core field model such as IGRF. Satellites also allow to select nighttime measurements, which permits to disregard the diurnal external magnetic field; the counterpart being that a stacking of systematic errors may be caused by local external fields unidentified on Earth on the nightside. The corrected data represent vector magnetic anomalies expressible in terms of SH. Such models are easily upward or downward continued to any spherical surface.

A satellite is constantly moving along its orbit and such displacements may be interpreted in terms of spatial or time-varying magnetic fields. Attitude uncertainties of the spacecraft, time errors, misalignment of magnetometers, and the absence of baseline control introduce a total error whose magnitude is enhanced by the spacecraft velocity. At about 7.5 km s^{-1} at 400 km altitude for the CHAMP satellite (Reigber et al., 2002), such uncertainties can convert into a few nT error. The absence of traverse lines also introduces East–west artificial offsets that may easily be confused with genuine anomalies (Thébault et al., 2010). For this reasons, only long periods of observations allow improving the signal-to-noise ratio. After a few years of measurements, it is generally accepted that lithospheric field models are robustly described to about SH degree 60 only because the signal drops into the noise somewhere between SH degree 60 to 90 (Figure 1). However, recent measurements acquired at 350-km altitude improved our view of the Earth's lithospheric field down to 350-km spatial wavelenghts (Maus et al., 2008) although these finer structures have not yet been ascertained by independant approaches. Magnetic anomaly maps obtained from space provide a statistical view of the entire crust down to the lithosphere (Langel and Hinze, 1998).

Aeromagnetic and satellite data offer a complementary view of the lithospheric field, yet incomplete. We remain partly blind for the wavelengths between 200 and 400 km. Such wavelengths could be in principle registered by stratospheric balloon at 20–30 km altitude but controlling the airship route is challenging. Thus, current efforts focus instead on ways to improve the resolution achievable from space by considering gradient measurements and constellation of satellites (Friis-Christensen et al., 2006).

World magnetic anomalies – WDMAM

For scientific purposes, it is essential to map the data over long distances, in order to evaluate the crustal structures, to understand the geological processes, to establish tectonic models of continental and oceanic crust, to enable paleo-reconstructions, and to contribute to geodynamic models. The merging of ground, near-surface, and satellite measurements is a way to access this multi-scale view of the Earth's lithospheric magnetic field. The World Digital Magnetic Anomaly Map (WDMAM) project is an international effort directed toward compiling magnetic anomaly data available in the public domain (Khorhonen et al., 2007). WDMAM 1.0 is based upon more than 50 years of aeromagnetic and marine surveys and provides a three arcmin spacing grid at 5 km altitude (Figure 2). It relies on published continental-scale compilations derived by independent groups (e.g., the North Atlantic Magnetic Anomaly Group NAMAG, or the Antarctic Digital Anomaly Map, ADMAP), as well as upon independent contributions from countries and individuals. Several mergers were tested and led to different candidate grids for the WDMAM 1.0. This difference illustrates the difficulty of blending such a variety of datasets. Continuous efforts are being made to improve the quality and the data coverage and will benefit WDMAM 2.0, which should be released in 2011.



Magnetic Anomalies, Interpretation, Figure 2 World magnetic anomaly compilation using aeromagnetic, marine, and satellite data. The magnetic signal in some oceanic part is computed via a magnetization model (After Khorhonen et al., 2007; courtesy of the Commission for the Geological Map of the World, CGMW).

Qualitative interpretations

Most susceptibility contrasts may give rise to a magnetic field detectable by a magnetometer (*Geomagnetic Field, Measurement Techniques*). An exhaustive review of qualitative magnetic anomaly map interpretation is impossible. Magnetic surveys are traditionally acquired for resource exploration in order to detect ore bodies, but also oil and natural gases, which are trapped in sedimentary basins. Magnetic maps are also useful to track anthropogenic buried materials, like unexploded bombs, hidden weapons, abandoned oil wells, but may also be used for guiding archaeological excavations, for instance.

At a regional scale, mapping the patterns of magnetic anomalies is a very effective way of reconnoitering large areas of geology virtually everywhere (Purucker and Whaler, 2007), in remote regions where rock outcrop is absent and bedrock covered by glacial, sand, vegetation, or water overburdens. In particular, the interpretation of magnetic anomalies over oceans was crucial to the development of the plate tectonics paradigm (Vine and Matthews, 1963). Over continents, tectonic processes are detailed in a magnetic anomaly map because they are responsible for the creation, destruction, and movement of magnetic materials within the lithosphere of the Earth that modify preexisting magnetic signatures. New magnetic contrasts are created by volcanism and related igneous processes such as dike emplacement, rifting, and faulting.

At a continental scale, magnetic anomalies are particularly prominent over the main old and cold continental masses in Africa, Antarctica, Asia, Australia, Europe, Greenland, India, and North and South Americas (Figure 2), and comparatively weaker over younger and hotter oceans. The largest magnetic anomaly features correlate well with cratons, Precambrian and Paleozoic provinces. Within each structure, magnetic anomalies of

smaller extension like mobile belts, collision zones, and volcanic areas can be characterized. Crustal thickness variations at continent ocean boundaries and large subduction zones are also sometimes outlined. Large magnetic anomalies often correlate with plate boundaries and seismotectonic activity around the world. Reconstructions of wide separated geological units or rotated blocks make extensive use of the seafloor stripes and transform faults.

Quantitative geological interpretations

Interpretation of magnetic maps is nonunique because potential theory shows that a single magnetic anomaly can be caused by an infinite variety of magnetized bodies. Quantitative analyses such as depth, shape, and position of a magnetized body require drastic assumptions or significant a priori information derived from other geological and geophysical results. Locally, on continents, a quantitative description of a magnetized body is performed (1) by planar geometry and by using processing techniques inherited from Fourier analyses; (2) by setting a distribution of equivalent magnetic poles in the subsurface and interpreting the observed anomaly in terms of inducing field; (3) by analyzing only simple regular anomalies (*Magnetic Anomalies of Simple Geophysical Models*). The picture of a magnetic anomaly is generally more understandable after a pole reduction that transforms an observed magnetic anomaly into the anomaly that would be measured at the north magnetic pole. The pole reduction method relocates the magnetic anomalies so that its maximum lies in the middle above the magnetized body, thus making magnetic interpretation easier. The flank of anomalies may also be clarified by upward and downward continuation and quick depth estimations are empirically obtained by computing the map of the magnetic anomaly derivatives. Baranov (1975) summarizes these techniques.

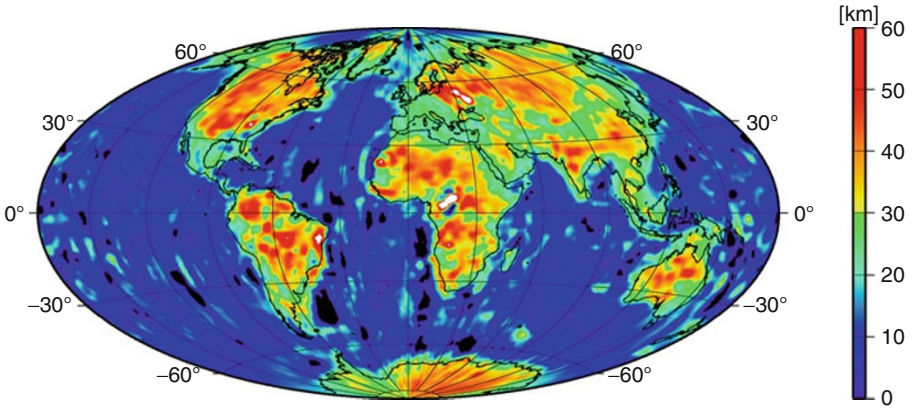
When the geometrical parameters are bounded, the magnetization of the body can be found by trials and errors and its rock type can be subsequently inferred from a table of rock magnetization (see *Magnetic Domains*). The magnetic field anomaly created by a simple geometric body at depth (*Magnetic Anomalies of Simple Geophysical Models*) is compared to the measured anomaly field until the residuals between both is comparable to measurement errors. In some areas, interpreting individually each magnetic anomaly is not feasible, particularly if the geophysical body has no simple geometry. In such cases, the ambiguities in interpreting the magnetic anomaly maps can be reduced only by requesting guidance from other disciplines like, gravity, seismic, geology, etc.

The systematic recourse to integrated geophysical information led to some successes in interpreting measurements from space at the global scale. Simplifying assumptions are easier to justify from space because of the measurements altitude and their horizontal spatial resolution. In particular, the magnetization can be assumed to vary laterally and be constant in a crustal layer of varying thickness because at 400 km altitude, the Earth's crust appears as a thin layer within which we cannot distinguish between the different sources depths and the problem can be solved in two dimensions only.

Maus and Haak (2003) verified that, statistically, the lithospheric magnetization was indeed mostly induced over continent and remanent over oceans at satellite altitudes. For this reason, it makes sense to assume over continents that the anomaly source is due to an array of equivalent dipole with moments aligned along the main core field. The bare thickness contrast between continental (~40 km) and oceanic (~7 km) crusts of uniform, but different, susceptibility values, generates magnetic anomaly field that are seen in magnetic field measurements (Counil et al., 1991). This shows that magnetic crustal thickness plays a significant role in the dichotomy between oceanic and continental magnetic signals observed at satellite altitudes (see *Magnetic Methods, Satellite*, Figure 1). Cohen and Achache (1994) went a step further and introduced a model of oceanic topography to estimate a more realistic crustal thickness in some oceanic areas; they could explain the magnetic anomalies observed over oceanic plateaus. On ocean bottoms, remanent magnetism predicted by Dyment and Arkani-Hamed (1998) explains well large magnetic signatures over superchrons in the North Atlantic and Indian oceans. Hemant and Maus (2005) synthesized these induced and remanent studies in a unique model in order to predict the lithospheric magnetic field anomalies at the altitude of the CHAMP satellite and could show that our understanding of the Earth's magnetic lithosphere was statistically reasonable.

Following the same scheme, the magnetic crustal thickness can be estimated by trials and errors by iterative comparisons between such predicted anomaly fields with real satellite magnetic anomalies. The solution shown in Figure 3 provides an estimate of the magnetic crustal thickness of the continents, thus a statistical estimate of the Curie depth (Purucker et al., 2002). The magnetic lithosphere is bounded at depth by temperatures and the magnetic crust can be related to thermal processes. Such picture shows, for instance, that despite a thick crust under the Tibetan plateau, the magnetic crust is there thinner because of enhanced heat flows. To some extent, the average heat flow (Fox Maule et al., 2005) may be estimated from the magnetic crustal thickness. However, it requires simplifying assumptions about the thermal state of the continental crust and a starting model to constrain the longest wavelengths of the lithospheric field, in excess of SH degree 15 (~2,900 km), which are difficult to assess because they are dominated by core field processes as is shown in Figure 1.

Quantitative interpretations of magnetic anomalies are thus possible although non-unique. Unfortunately, there is currently no alternative between analyzing aeromagnetic data assuming planar geometries and studying satellite data using spherical harmonics and equivalent dipole techniques. This lack of joint processing is explained by the poor compatibility between near-surface data and satellite data.



Magnetic Anomalies, Interpretation, Figure 3 Estimation of the world magnetic crustal thickness of continents from the magnetic anomaly map based on satellite data (After Purucker et al., 2002; Mollweide projection).

Summary

On Earth, a magnetic anomaly is defined with respect to the core and external fields. Magnetic anomaly maps give the essence of the crustal magnetic field that in turn provides useful geological information about the magnetization of rocks in the subsurface. Characterizing the anomaly field and depicting all spatial dimensions is not simple as data processing leading to anomaly maps differ with altitudes and epochs. Nowadays, the main sources of discrepancies are the rapid temporal changes of the external magnetic field that are poorly understood. The lack of data at intermediate altitudes is the main impediment to the reconciliation between near-surface and satellite-based anomaly maps. There is indeed little overlap between the wavelengths measured by satellite and airborne surveys and even between the communities involved in their analysis. Current efforts are being made to reprocess ancient aeromagnetic and marine data, and to acquire unprecedented high-resolution magnetic data. This should ultimately provide us with a complete view of the magnetic crust and thus prompt for novel joint analysis techniques.

Bibliography

Backus, G., Parker, R., and Constable, C., 1996. *Foundations of Geomagnetism*. Cambridge: Cambridge University Press.

Baranov, V., 1975. *Potential Fields and Their Transformations in Applied Geophysics.* Berlin: Borntraeger, ISBN:3443130089.

Cohen, Y., and Achache, J., 1994. Contribution of induced and remanent magnetisation to long-wavelength oceanic magnetic anomalies. *Journal of Geophysical Research*, **99**, 2943–2954.

Counil, J., Cohen, Y., and Achache, J., 1991. The global continent–ocean magnetisation contrast: spherical harmonic analysis. *Earth and Planetary Science Letters*, **103**, 354–364.

Dunlop, D. J., and Özdemir, O., 2007. *Magnetisations in Rocks and Minerals*, Chapt. 8, Vol. 5, Kono, M. (ed.), Elsevier, Treatise on Geophysics, pp. 278–331.

Dyment, J., and Arkani-Hamed, J., 1998. Contribution of lithospheric remanent magnetisation to satellite magnetic anomalies over the world's oceans. *Journal of Geophysical Research*, **103**, 15423–15441.

Fox Maule, C., Purucker, M., Olsen, N., and Mosegaard, K., 2005. Heat flux anomalies in Antarctica revealed by satellite magnetic data. *Science*, **309**, 464–467.

Friis-Christensen, E., Lühr, H., and Hulot, G., 2006. SWARM: a constellation to study the Earth's magnetic field. *Earth Planets and Space*, **58**(4), 351–358.

Hemant, K., and Maus, S., 2005. Geological modeling of the new CHAMP magnetic anomaly maps using a Geographical Information System (GIS) technique. *Journal of Geophysical Research B*, **110**, B12103, doi:10.1029/2005JB003837.

Hulot, G., Olsen, N., and Sabaka, T. J., 2007. *The Present Field*, Chap. 6, Vol. 5: Geomagnetism. Kono, M. (ed.), Elsevier, Treatise on Geophysics, pp. 33–72.

Khorhonen, J. K., Fairhead, J. D., Hamoudi, M., Hemant, K., Lesur, V., Mandea, M., Maus, S., Purucker, M. E., Ravat, D., Sazonova, T., and Thébault, E., *Magnetic Anomaly Map of the World – Carte des anomalies magnétiques du monde*, Scale: 1:50,000,000, 1st edn. Commission for the Geological Map of the World.

Langel, R. A., and Hinze, W. J., 1998. *The Magnetic Field of the Earth's Lithosphere: The Satellite Perspective*. New York: Cambridge University Press, p. 429.

Lowes, F. J., 1974. Spatial power spectrum of the main geomagnetic field and extrapolation to the core. *Geophysical Journal of Royal Astronomicl Society*, **36**, 717–730.

Maus, S., and Haak, V., 2002. Is the long wavelength crustal magnetic field dominated by induced or by remanent magnetisation? *Journal of Indian Geophysical Union*, **6**(1), 1–5.

Maus, S., Yin, F., Lühr, H., Manoj, C., Rother, M., Rauberg, J., Michaelis, I., Stolle, C., and Müller, R. D., 2008. Resolution of direction of oceanic magnetic lineations by the sixth-generation lithospheric magnetic field model from CHAMP satellite magnetic measurements. *Geochemistry, Geophysics, Geosystems*, **9**, Q07021, doi:10.1029/2008GC001949.

Maus, S., 2010. An ellipsoidal harmonic representation of Earth's lithospheric magnetic field to degree and order 720. *Geochemistry, Geophysics, Geosystems*, **11**, Q06015, doi:10.1029/2010GC003026.

Purucker, M., Langlais, B., Olsen, N., Hulot, G., and Mandea, M., 2002. The southern edge of cratonic North America: evidence from new satellite magnetometer observations. *Geophysical Research Letters*, **29**(15), 8000, doi:10.1029/2001GL013645.

Purucker, M., and Whaler, W., 2007. *Crustal Magnetism*, Chap. 6, Vol. 5, Kono, M. (ed.), Elsevier, Treatise on Geophysics, pp.195–237.

Reeves, C. V., and De Wit, M., 2000. Making ends meet in Gondwana: retracing the transforms of the Indian Ocean and reconnecting continental shear zones. *Terra Nova*, **12**, 272–280, doi:10.1046/j.1365-3121.2000.00309.x.
Reeves, C., 2005. *Aeromagnetic surveys: Principles, practice and interpretation*, published by GeoSoft, p.155, http://www.geosoft.com/resources/news/2009/090723.asp.
Regan, R. D., Cain, J. C., and Davis, W. M., 1975. Global magnetic anomaly map. *Journal of Geophysical Research*, **80**, 794–802.
Reigber, C., Lühr, H., Schwintzer, P., 2002. CHAMP Mission Status. *Advaces in Space Researchi*, **30**, doi:10.1016/S0273-1177(02)00276-4.
Runcorn, S. K., 1975. On the interpretation of lunar magnetism. *Physics of the Earth and Planetary Interiors*, **10**, 327–335.
Thébault, E., Purucker, M., Whaler, K., Langlais, B., and Sabaka, T. J., 2010. The Magnetic field of the Earth's lithosphere, *Space Science Reviews*, **155**(1), doi:10.1007/s11214-010-9667-6.
Vine, F. J., and Matthew, D. H., 1963. Magnetic anomalies over oceanic ridges. *Nature*, **199**, 947–949.

Cross-references

Curie Temperature
Geomagnetic Field, Global Pattern
Geomagnetic Field, IGRF
Geomagnetic Field, Secular Variation
Geomagnetic Field, Theory
Geomagnetic Field, Measurement Techniques
Magnetic Data Enhancements and Depth Estimation
Magnetic Domains
Magnetic Methods, Airborne
Magnetic Methods, Satellite
Magnetic Methods, Surface
Paleomagnetism, Principles

MAGNETIC DATA ENHANCEMENTS AND DEPTH ESTIMATION

Clive Foss
CSIRO, Clayton, South VIC, Australia

Synonyms

Magnetic data processing

Definition

Magnetic data are measurements of the magnetic field (generally of its intensity). Corrections of magnetic data are processing operations designed to remove unwanted features in the data. Enhancements of magnetic field data are processing operations designed to preferentially accentuate the expression of a selected magnetization at the expense of others.

Introduction

Measurement of the earth's magnetic field is used in exploration geophysics as a remote-sensing method to investigate subsurface distributions of magnetization (see *Magnetic Methods, Surface*). The measurements require correction to eliminate artifacts and to remove the primary core field of the earth, thereby improving the expression of local field variations due to upper crustal magnetizations. These corrections are applied to produce clean and robust output suitable for interpretation (see *Magnetic Anomalies, Interpretation*). For specific interpretation objectives, enhancements are also applied to the data to increase the expression of selected sources at the expense of others.

The strong linkage between electricity and magnetic fields allows the development and calibration of magnetometers that supply absolute measurements of the magnetic field (see *Magnetometers*). Most processing corrections are designed to safeguard this integrity of data values and provide data suitable for input to modelling and inversion. Conversely, some enhancement operations distort the data, and care must be taken to ensure that enhanced data is not used inappropriately.

Most magnetic field measurements are of the strength of the field, termed "total magnetic intensity" or TMI. Although these measurements are commonly treated as a scalar, it is important to remember that they are the amplitude of a vector, that the direction of that vector is also significant, and that in regions of strong fields the direction may vary across the survey area. The second most common form of magnetic field measurements are gradients of TMI, acquired by simultaneous measurement with two or more TMI sensors at fixed offsets (see *Magnetic Gradiometry*). Other forms of magnetic field data include measurements of the Cartesian components of the field and of the gradient tensor. Details of corrections and enhancement change only slightly between these different types of magnetic field measurements. The mathematical derivation of many correction and enhancement operators is presented in "Potential Theory in Gravity and Magnetic Application" (Blakely, 1996) which provides an excellent basis to understand this topic. A brief review by Milligan and Gunn (1997) (http://www.ga.gov.au/image_cache/GA2486.pdf) is also useful reading.

Standard magnetic data processing corrections

Measurement of the earth's magnetic field provides a spot value at a specific time and location. Magnetic field data is recorded together with ancillary information such as horizontal position, elevation, and time, as required for processing and interpretation. Corrections of data that are specific to instrumentation are an integral part of data acquisition, and are not discussed in this topic. Common corrections in the processing of magnetic data include diurnal correction, leveling, IGRF correction, gridding, and reduction to pole.

Diurnal correction

The earth's magnetic field is continuously changing (see *Geomagnetic Field, Secular Variation*), and this clearly presents a challenge in mapping spatial variations. Temporal changes of the magnetic field during the progress of a survey are generally termed "diurnal" variation, although this name more correctly describes field variations specifically at the 24 h rotation cycle of the earth.

On normal "quiet" days the strength of the geomagnetic field varies across a range of approximately 50–100 nT. The amplitude and pattern of diurnal variation are functions of latitude, but vary sufficiently day to day that they must be measured at a fixed base station for subtraction from the field data. Surveys presenting the greatest challenges for diurnal correction are marine surveys away from suitable locations for a base station or that including substantial lateral variation in bathymetry or crustal conductivity (which influence diurnal variation through secondary induction fields). If a simple diurnal correction is insufficient, a spatial function can be generated from a distributed network of base stations incorporating time shifts according to the longitude of the base stations and the survey measurements.

There are also periods of rapid, high-amplitude variations in the magnetic field termed "magnetic storms" which follow periods of high sun spot activity. It is not feasible to accurately correct for field variations during magnetic storms, and surveys are generally suspended during these periods.

Network adjustment (leveling)

Most magnetic field measurements are made sequentially from platforms moving along a profile. In particular, airborne magnetic surveys are generally flown on parallel, equally spaced flight lines (see *Magnetic Methods, Airborne*). To reduce errors arising from imperfections in diurnal correction, and to compensate for any residual platform heading effects, a subset of lines, termed "tie lines," is flown perpendicular to the main flight line direction, at a spacing typically five or ten times larger than for the flight lines. The purpose of the tie lines is to provide a set of intersections with the flight lines that can be used for network adjustment or "leveling" of the data. Network adjustment progressively changes base levels, gradients, and curvatures of individual lines and line segments to minimize the misfit of flight line and tie line measurements across the network of intersection points. Leveling errors are primarily manifest as sharp field variations along the axes of the flight lines.

Micro-leveling

The standard network adjustment of data is an empirical, least-squares optimization which leaves residual errors. Micro-leveling or "de-correlation" is a grid-based filtering process designed to attenuate residual level shifts between adjacent lines. These features are recognized as being parallel to the flight line direction and having a spacing related to the flight line spacing. Any valid magnetic field variations with similar characteristics will also be attenuated by micro-leveling (although any features parallel to the flight line direction are always of uncertain validity and difficult to interpret). Micro-leveling commonly has the objective of attenuating noise which becomes evident in the sun-shading of images, or to smooth "herring-bone" patterns in contour lines. Micro-leveling is also used to precondition data for subsequent enhancements that would otherwise include strong expressions of these residual leveling errors. Micro-leveling improves the quality of images of the magnetic field, but it also produces poorly controlled attenuation of the amplitude of field variations, and should not be used for modelling and inversion studies.

Elevation correction

Absolute elevation values are much less significant in magnetic field data processing than in gravity data processing (see *Gravity, Data to Anomalies*). The IGRF definition (see *Geomagnetic Field, IGRF*) includes elevation terms, but these only become significant over ranges much more substantial than those encountered in almost any survey. More significant to the information content of magnetic field data is the elevation separation between the sensor and the shallowest magnetic sources. For this reason, airborne surveys are commonly flown on drape surfaces at a nominal terrain clearance. Rugged terrain clearly increases the problem of drape flying, and rugged areas are best flown on a compromise, smooth drape surface to reduce abrupt elevation steps from line to line. Rugged terrain also poses problems for leveling of data because of difficulties in achieving identical elevations on flight lines and tie lines at each intersection point. The most effective way to correct for these elevation differences is to re-project data from the sensor elevation onto a suitable smooth drape surface by using a set of equivalent sources (Dampney, 1969; Emilia, 1973) to invert the data, and then forward computing the field from those sources onto the reference surface.

Terrain correction, used extensively in processing gravity data, is rarely applied to magnetic field data. It is unusual for terrain to have a consistent magnetization such that the magnetic field of the terrain can be derived by any simple integration. Also, the ground surface is only an effective magnetization contrast over the top of strongly magnetized bodies, and in situations of deep weathering or transported cover, the top of any strong magnetizations may be substantially beneath the ground surface. In areas of rugged terrain the influence of terrain magnetization can sometimes be recognized, but these variations are generally less than those due to variations in separation between the sensor and the terrain. Typical elevation clearance variations include lower clearance over peaks and crests than over the flanking slopes, and increased clearance over deeply incised valleys. It is generally this variation in terrain clearance which carries the danger of incorrect geological interpretation. For instance, an increase in terrain clearance causes attenuation of magnetic field expressions of shallow sources which might be misinterpreted as due to zones of alteration and partial magnetite destruction.

Each magnetic field data channel should be linked to an elevation channel. For the primary field measurements this is generally the sensor elevation determined from

GPS altitude measurements. The smooth drape surface derived or implicit in leveling the magnetic field data is the effective elevation of that data for any modelling or inversion. With some airborne magnetic data no elevation channel is supplied, and in these cases, the best option is to derive a synthetic elevation channel by interpolating the ground elevation (e.g., from SRTM data) onto the lines, adding the nominal terrain clearance reported for the survey, and then low-pass filtering. Another situation sometimes faced by an interpreter is to have only radar altimeter data available. The radar altimeter can be added to the terrain channel to recover sensor elevation, but before doing this, the radar data should be carefully inspected, as it may contain short-wavelength noise which needs to be smoothed, and possibly spikes or abrupt steps which must be eliminated.

De-culturing of magnetic field data

The expressions of man-made features in magnetic field data used for geological studies are termed "cultural" anomalies, and the process of removing these anomalies is termed "de-culturing" the data. Anomalies may be recognized as being man-made by their sharpness, indicating that they are distinctly shallower than other field variations in the area, by distinctive characteristics, or by correlation with mapped man-made features. Automated search algorithms can be used to scan data for possible cultural anomalies, but generally removing cultural anomalies is an interpretive, labor-intensive effort. Removal of cultural features is best done from the primary data. More general removal of sharp variations in the data is known as "despiking."

Gridding of magnetic field data

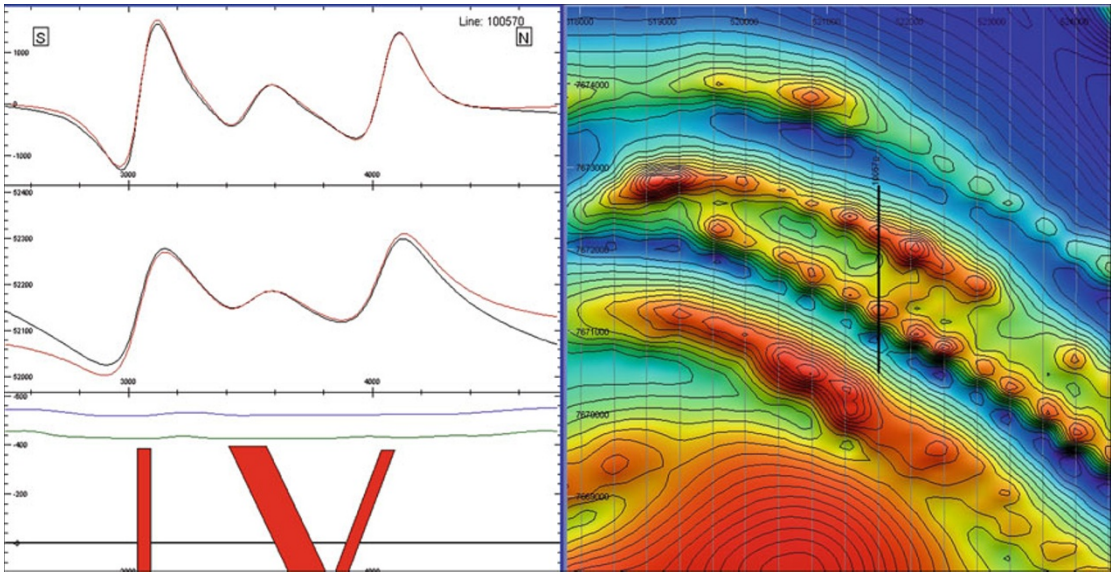
Operations through to leveling of the magnetic data are performed on the primary located data, and key channels are appended to the data so that changes can be tracked through the processing sequence. Once data has been levelled, the next process is generally to grid the data. Gridding is not a correction process, but is crucial to all subsequent processing, enhancement, display, and interpretation. Gridding interpolates the data from the measurement locations to nodes of a regular mesh, creating a new and fundamentally different construct of the data. Magnetic field data is generally gridded using minimum curvature operators to cell sizes of 1/4 or 1/5 of the line spacing. Minimum curvature operators (Briggs, 1974; O'Connell et al., 2005; O'Connell and Owers, 2008) are generally implemented as multi-pass functions with control parameters which can be modified to adjust to different data sets. Minimum curvature is an appropriate imposition on magnetic field data as the magnetic field is continuous with continuous derivatives. More sophisticated gridding algorithms, such as anisotropic diffusion (Smith and O'Connell, 2005), can also be justified as supplying superior interpolation of the field, but minimum curvature is the current industry standard and is adequate in most situations. Gridding breaks down most readily for sharp anomalies due to small, shallow sources. For surveys of regular flight line orientation and spacing, this can occur either where the magnetized unit is particularly shallow or where it trends oblique to the flight line direction. In these situations, the anomaly due to a continuous magnetic unit may break down to a "string of pearls" or "bicycle chain" pattern (Figure 1) because of the insufficiency of the gridding algorithm to interpolate that sharp field curvature from one flight line intersection to the next. These artifacts restrict the use of sun-shading to highlight subtle features in the data and must be removed prior to any attempt to highlight those details by subsequent enhancement processing. These gridding artifacts can be attenuated by re-gridding at a coarser cell-size, or by using a redesigned gridding operator to produce a smoother, lower curvature grid. Interpretation of the depth to the top of a magnetic source is based primarily on the sharpness of curvature of magnetic field variations, and limitations in preserving the sharpest gradients through the gridding process mean that depths estimated from inversion or modelling of grid data may be overestimated, particularly for shallow sources.

IGRF correction

Across survey areas of length several tens of kilometers or more, variations in the earth's primary field may contribute significantly to the measured field variations. In order to emphasize more local field variations due to crustal sources, the International Geomagnetic Reference Field (IGRF), a global model of the earth's magnetic field, is subtracted from the measured data. The IGRF is not a separation of field components from specific sources, but the long wavelength IGRF terms represent mostly core components and broad, lower crustal magnetizations. The IGRF is defined and published by the International Union of Geodesy and Geophysics (IUGG). The 11th generation model released in 2009 (http://www.ngdc.noaa.gov/IAGA/vmod/igrf.html) contains coefficients of degree and order 1–13. (See also entry *Geomagnetic Field, IGRF*).

Reduction to pole (RTP) transform

The external field of a magnetic body varies substantially as a function of the local geomagnetic inclination (see *Magnetic Anomalies, Interpretation*). The simplest relationship between the magnetic field and source magnetizations occurs at the geomagnetic poles where the field orientation is vertical (see *Geomagnetic Field, Global Pattern*). At the geomagnetic poles, induced anomalies due to steeply dipping bodies are essentially positive and are centered above the bodies. The reduction to pole transform is a fast Fourier transform (FFT) filter applied to replicate the field variations expected in a vertical field. The RTP transform is unstable when applied to low geomagnetic inclination fields (this is generally perceived to be a problem at inclinations of less than 30°). The severity of the problems depends on data quality



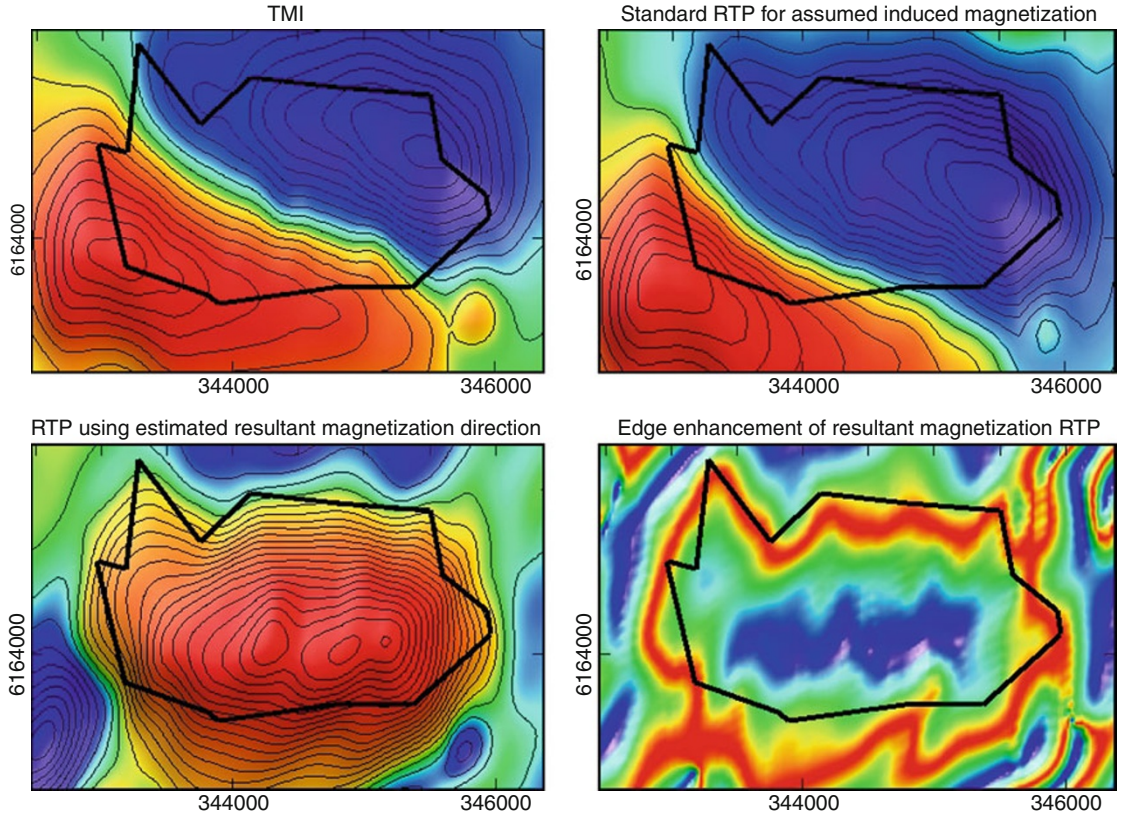
Magnetic Data Enhancements and Depth Estimation, Figure 1 Example aeromagnetic profile over a synclinal fold in the Frew River area of the Northern Territory of Australia (data courtesy of the Northern Territory Geological Survey and Geoscience Australia). The TMI image on the right shows a "string of pearls" pattern. The top track of the profile data on the left shows a vertical derivative enhancement of the measured TMI (*black*) and model computed TMI (*red*). The model shown below was generated by inversion to match these filter enhancements. The mismatch of the measured and model computed TMI in the center track requires addition of a background or "regional" field. The filter enhancement provides higher resolution for the inversion than is available from the primary TMI data.

and sufficiency of sampling, and these problems arise in particular with any apparent North–South trending features in the data. Several methods can be used to overcome low inclination RTP problems, as reviewed by Li (2008). In one approach, the RTP filter can be distorted to reduce the instability. The output of the transform is not strictly valid, but may have the desired characteristics of a steep inclination field. Another approach is to use an equivalent source technique (Guspi and Novara, 2009). The reduction to equator (RTE) transform also provides a partial alternative to RTP at low geomagnetic inclinations. The RTE transform shifts anomalies North–South to be directly above the source magnetizations, but the East–West elongation which is a natural characteristic of anomalies at low geomagnetic inclinations will be slightly amplified. Note that RTP or RTE transforms should not be used in modelling or inversion studies, as forward modelling algorithms readily incorporate the local geomagnetic field direction (although RTP images may help to approximately locate bodies in starting models).

The commonly used RTP transform requires that the direction of magnetization be in the local direction of the geomagnetic field, and for bodies with substantial remanent magnetization (see *Remanent Magnetism*) in a different direction this transform is invalidated. If however the resultant magnetization direction (the vector addition of induced and remanent magnetizations) is known from palaeomagnetic studies, magnetic moment analysis (Helbig, 1963; Schmidt and Clark, 1998; Phillips, 2005), or inversions of the magnetic field data, then an alternative RTP operator can be used which allows for separate specification of the geomagnetic field and the magnetization direction (Figure 2). All FFT filters have potential pitfalls, in part because of the fundamental assumption that the fields repeat infinitely beyond the sampled area. The FFT transform requires some processing to address level shifts between opposite edges of the sample area, and to pad the sample area with a smooth extrapolation. However well this is done, the output around the margins of the data is less reliable than in the center. This problem is of particular concern with the RTP transform, especially at the northern and southern edges of the data, and for longer wavelength field variations that are migrated furthest by the RTP transform.

The pseudo-gravity transform

Gravity and magnetic fields are linked by the common fundamental theory of potential fields (see *Magnetic Methods, Principles*), as described by Poisson's relationship (Baranov, 1957). If the direction of magnetization is consistent, and if there is a constant ratio between density and magnetization, then the RTP magnetic field and the vertical gradient of the gravity field are identical. For any but the simplest geological system, there is no such reliable linkage between magnetization and density, but



Magnetic Data Enhancements and Depth Estimation, Figure 2 The *top right* image is of a TMI anomaly over the Black Hill Norite in South Australia which is dominated by remanent magnetization (data courtesy of the Department of primary Industries and Resources South Australia). The contour interval is 200 nT, and the outline is of the source body derived from a staged inversion of the TMI data using a resultant magnetization direction determined from magnetic moment analysis. The *top right* image is of a standard RTP transform using an incorrect assumption that magnetization is in the geomagnetic field direction. The *bottom left* image is an RTP which correctly incorporates the resultant magnetization direction to produce a compact, positive anomaly centered over the source body. The *bottom right* image is an edge enhancement based on the tilt filter, applied to the resultant magnetization RTP. The maximum in this filter approximately matches the edge of the computed model.

nevertheless Poisson's relationship provides a basis for converting a magnetic field to an equivalent gravity field expression. This transform includes an inverse vertical derivative operation, which amplifies the expression of longer wavelength components of the field. The pseudo-gravity transform is sometimes used to apply analysis and processing techniques developed for gravity data to the magnetic field data, but this must be done with care. For instance, the pseudo-gravity transform accentuates the longer wavelengths of the field, but if the data from a small survey area is transformed, those longer wavelengths are not sufficiently sampled.

Enhancements of magnetic data

Enhancement of magnetic field data applies to any operation designed to highlight field variations due to a selected source of magnetization. These operations include application of various filters and also selection of display options such as histogram normalization and sun-shading.

Image display of magnetic field data

The display of magnetic field data is not an operation in the sense that it changes values of the data, but nevertheless it is highly significant in determining how the data is perceived, and how easily geological information can be recovered from it. Magnetic field data generally has much sharper variations than gravity data across the same area, with much of the amplitude range in only a small subset of the data. For this reason, magnetic field data is often displayed using a histogram equalized color stretch, devoting equal areas of the image to each color band. This form of imaging provides a much improved representation of detail across the complete range of the data. Histogram equalization does however have the drawback that it is difficult to appreciate the true amplitude range of features or to compare the relative amplitudes of different features.

Another important enhancement tool in image display is the use of sun shading. This process involves the addition of shadows and highlights to an image to mimic the illumination of the data surface by a light source. These highlights and shadows reveal features of shorter wavelength and more subtle amplitude variation than can be easily recognized in the color mapping of the amplitude values. With a vertical light source, the highlights appear along the axes of minima and maxima. However, light sources are generally set to moderate or low inclinations to produce a stronger, directional enhancement of features trending perpendicular to the illumination direction. Illumination applied parallel to the flight line direction minimizes the expression of any residual leveling errors.

Vertical continuation of the field

The amplitude of a magnetic field above a source varies with elevation as an exponential function of wavelength. This relationship can be readily exploited with FFT filters to recompute the field at a higher elevation ("upward continuation") or lower elevation ("downward continuation"). Upward continuation attenuates the field through an increased separation from the magnetization. Conversely, downward continuation towards the source or sources amplifies field variations. Upward continuation is inherently stable and acts as a smoothing filter, whereas the amplification of field variations due to downward continuation becomes unstable at depths approaching the source depth, or shallower if there is any shorter wavelength contamination of the data. Vertical continuation of a magnetic field is an enhancement filter because it preferentially accentuates the fields due to sources of different elevation. Methods, including equivalent source computations can be used for continuations in which the input and/or output surfaces are irregular (Hansen and Miyazaki, 1984; Pilkington and Urquhart, 1990). Upward continuation is useful for attenuation of fields from shallow magnetizations to focus on deeper sources, as the effect of the filter can be properly compensated by changing the elevation at which the model field is computed.

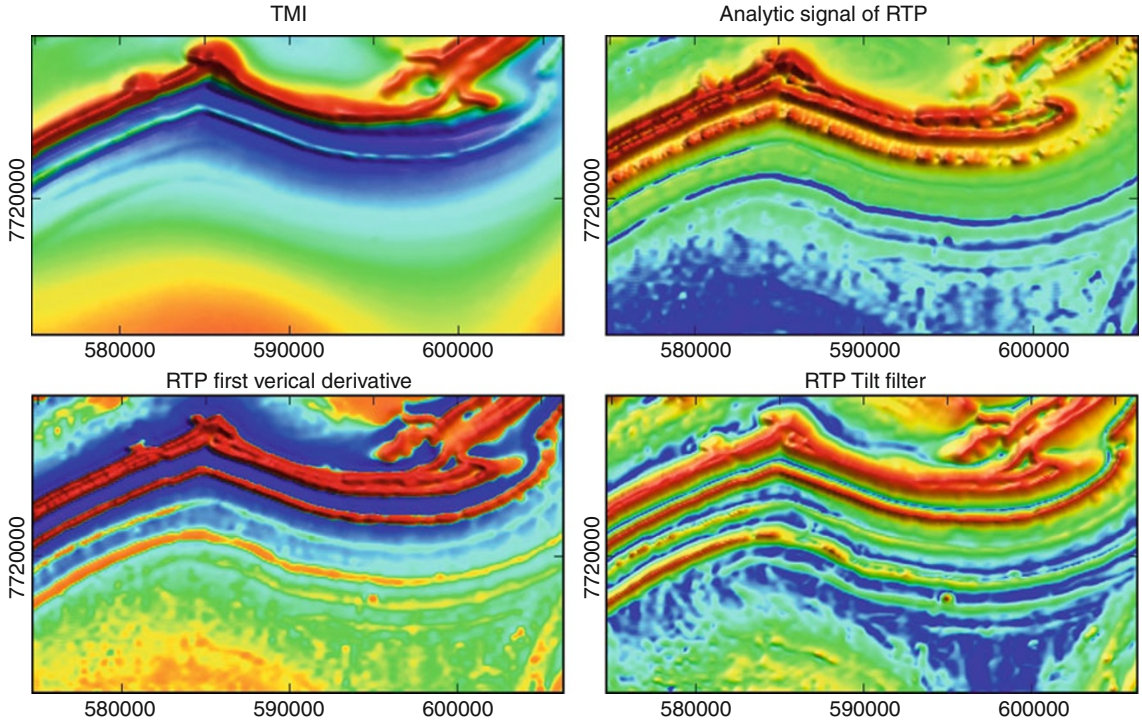
Vertical gradients of the field

Vertical gradient (or alternatively named "vertical derivative") filters preferentially amplify short-wavelength components of the field at the expense of longer wavelengths (Figure 3). Vertical gradient filters are generally applied to gridded data using FFT filters, and are therefore influenced by the characteristics of the gridding operator. Vertical gradient filters can also be applied to profile data on the assumption that the horizontal gradient measured along the profile is the only horizontal gradient. In some cases, this is a poor approximation, but for highly elongate anomalies, and with profiles approximately perpendicular to those anomalies, the approximation may be acceptable. If, furthermore, the magnetic sources are shallow compared to the line spacing of the survey, then there may be considerably more short-wavelength detail in the profile data than can be preserved through the gridding process. Vertical gradient profile filtering can be used to retain and enhance the shortest measured wavelengths during modelling and inversion by matching the outputs of the filter applied identically to the measured and computed fields (Figure 1). The second vertical derivative is a more extreme enhancement of a magnetic field. As such, it provides a more distinctive amplification of fields from shallow sources, but simultaneously is more sensitive to any short-wavelength artifacts in the data. In the event that a derivative of the field unacceptably accentuates noise or shallow magnetizations, the data can be preconditioned with a mild upward continuation before applying the derivative filter.

Bandpass and matched filters

The general relationship between depth of a source and the wavelength of its magnetic field can also be exploited to enhance the expression of sources from particular depths through bandpass filtering. Short-wavelength field variations can only be generated by shallow sources, so high pass filtering attenuates the expression of deep sources, and conversely, low pass filtering tends to attenuate the expression of shallow sources. However, bandpass filters rarely, if ever, provide a comprehensive separation of magnetic fields from different sources, so their outputs are not suitable for modelling or inversion studies (unless the identical filter is applied to the forward computed model fields).

The power spectrum display of a magnetic field commonly consists of segments which approximate to straight lines, the slope of which may be a function of the depth to the assemblage of sources whose field contributions dominate that part of the spectrum (Spector and Grant, 1970). The inspection of a power spectrum can support selection of cutoff wavelengths for bandpass filtering, with the taper of the filter also very important. Filters should not truncate abruptly nor should they target too narrow a band of wavelengths. Size, shape, and depth extent of bodies also strongly influence their magnetic field power spectrum, and the analysis by Spector and Grant is valid only statistically and for an assemblage of bodies. With some bandpass filtering, the transform back into the space domain has been claimed to localize fields from sources at particular depths, with terms such as "depth slice" and "pseudo-depth slice." These filters may produce images which are useful for interpretation, but any claims that they effectively invert the magnetic field data to supply a 3D mapping of magnetization are untenable. Nor does fitting a straight line to a segment of a power spectrum prove the effective wave-number separation of field contributions from sources at any particular depth, as implied by classifying these filters as matched filters. The combination of spectral characteristics and spatial localization supplied by wavelet processing (see *Wavelet Analysis*) promises considerable advantages over FFT methods,



Magnetic Data Enhancements and Depth Estimation, Figure 3 (All images are histogram equalized). The *top left* image shows TMI over the Frew River area of the Northern Territory of Australia (data courtesy of the Northern Territory Geological Survey and Geoscience Australia). The *bottom left* image is of the vertical gradient of TMI. This enhancement reveals weak anomalies across the center of the image which are poorly imaged in the TMI data. The *top right* image is the modulus of the analytic signal, which also enhances the expression of shallow magnetizations. The tilt filter image in the lower right is the most extreme enhancement and provides the best expression of weak, near-surface magnetizations across the lower gradient southern half of the area.

and some wavelet processing methods are now being used, particularly for localizing edges of magnetization in multi-scale edge detection.

Regional-residual separation

Potential fields are additive, and the magnetic field near the surface of the earth generally contains contributions from multiple crustal magnetizations. Separation of field variations from different sources is interpretive and nonunique, with no general analytic solution. One class of field separations is to remove local features of sharp curvature in the field ("anomalies" or "residuals") from broader, lower curvature variations due to deeper or more distant sources (the "regional"), a process termed "regional-residual" separation. Regional-residual separation is most reliably performed in cases for which the fields to be separated have quite different characteristics; if the characteristics of the fields are similar, their separation becomes more interpretive and less certain. Regional fields are also due to geological structures, but which are too deep or distant to be of immediate interest in a specific project. For modelling studies, the regional separation should be adjustable during the modelling, as a better realization of this separation emerges as the model is developed.

The modulus of the analytic signal

The modulus of the analytic signal (Nabighian, 1972, 1984; Roest et al., 1992; Hsu et al., 1996), commonly referred to simply as the analytic signal, is a popular gradient enhancement, given by:

$$\text{analytic signal of TMI} = \sqrt{\left((\partial \text{TMI}/\partial x)^2 + (\partial \text{TMI}/\partial y)^2 + (\partial \text{TMI}/\partial z)^2\right)} \quad (1)$$

For compact or thin bodies, the modulus of the analytic signal peaks approximately over the center of the body (Figure 3), for wider bodies over their edges. It is only weakly sensitive to magnetization direction, and indeed the two-dimensional case with only one horizontal gradient is independent of magnetization direction. The modulus of the analytic signal of TMI has much lower sensitivity to the inclination of the geomagnetic field than the original TMI data, and provides a means to analyze low latitude magnetic fields without the concerns of the RTP operator (Macleod et al., 1993; Rajagoplan, 2003). Source parameters such as depth can be estimated from statistics of the filter output, but these more advanced applications are more appropriately classified as interpretation, rather than enhancement processing.

The tilt filter

The tilt filter (Miller and Singh, 1994; Dentith et al., 2000; Verduzco et al., 2004; Salem et al., 2008) has also become a popular enhancement with an output given by:

$$\text{Tilt of TMI} = \tan^{-1}\left(\frac{\frac{\partial \text{TMI}}{\partial z}}{\frac{\partial \text{TMI}}{\partial h}}\right) \quad (2)$$

where

$$\partial \text{TMI}/\partial h = \sqrt{\left((\partial \text{TMI}/\partial x)^2 + (\partial \text{TMI}/\partial y)^2\right)} \quad (3)$$

As a ratio of gradients, the filter output retains no amplitude information, and the degree of enhancement of this filter is variable according to the local amplitude of the field gradients. In areas of low gradient, there is no signal to enhance, but the output of the filter is defined at all places; so it is important that tilt images be interpreted with reference to the original data (Figure 3). The tilt filter and analytic signal are both expressions of horizontal and vertical gradients, and as such, there are clearly linkages between them in their tendency to amplify field variations from shallow sources, highlight the location of subvertical edges (after RTP transform), and to provide estimates of depth to magnetization contrasts according to their sharpness of variation.

Multi-scale edge detection ("worming")

Local maxima in the horizontal gradients of a magnetic field can be analyzed using wavelets as a function of the depth to magnetization, with the advantage that wavelet functions can be selected for direct compatibility with potential field characteristics (Hornby et al., 1999; Ridsdill-Smith and Dentith, 1999; Vallée et al., 2004). Multi-scale edges defined by the wavelets have some affinities to different wavelength components separated in Fourier analysis, with the detail from shallow sources mostly carried in the small-scale edges, and broader, large-scale edges providing more information about deeper sources. Multi-scale edges can also be derived from mapping the migration of maxima in the horizontal gradients of the field at different levels of upward continuation (Holden et al., 2000; Lahti and Karinen, 2010). Wavelets are intrinsically linked to the analytic signal and tilt filter gradient functions, Euler deconvolution, and to other curvature functions previously used to map source edges (Blakely and Simpson, 1986; Grauch and Cordell, 1987).

Analytic derivation of magnetic source parameters

Nonuniqueness in interpreting a distribution of magnetization from study of a measured magnetic field precludes any definitive analytic solution of this problem. However, the interface between analytical processing operations and interpretation is blurred by methods that impose either implicit or explicit conditions on the distribution of magnetization and reduce the interpretation problem to one with much more accessible and restricted solutions. The most common of the imposed conditions is that magnetized bodies have depth and strike extents that can be summarized by a structural index. This factor is linked to the curvature of the magnetic field and its attenuation with increasing source depth and therefore must be supplied or derived for estimation of source parameters.

Automatic mapping of source magnetization distributions has been one of the most prolific subjects of geophysical publication over many years, but those methods that have achieved widespread acceptance and use can mostly be classified into a small number of groupings, namely; Wener deconvolution, Euler deconvolution, the Naudy method, and the source parameter imaging ("SPI") method. The fundamental assumptions of the various methods are necessarily similar, and in many cases identical, but nevertheless there are some practical differences that might lead to advantages in applying certain methods in certain circumstances. In all cases, the key factor for successful application of automated source detection and parameter estimation is the suitability of the geology. If the geology provides discrete, homogeneous, well-separated, moderate to strong magnetizations, in simple and predictable geometries, and with an absence of any additional magnetizations that would produce overlapping field variations, then any of the automated source estimation methods should work well. If however the geology results in irregular and indistinct distributions of magnetization, then there is no basis for the successful application of any of the methods, and they can only lead to fictitious and misleading results. This leads to a common dilemma in the interpretation of any data; whether to set strict quality conditions and accept just a few solutions of apparently high validity, or whether to relax the quality conditions to generate a more comprehensive set of solutions. In practice, validity of solutions is commonly evaluated from their quality coefficients, by clustering of solutions and their internal consistency, or simply by visual inspection and judgement as to whether the magnetic field anomaly the solutions are derived from is reliable and suitable. The ability to rapidly generate source parameter estimates from automatic or semiautomatic methods is always in part offset by the time and discretion required to evaluate and interpret the reliability and usefulness of those results. All automated source depth estimators also have in common that they are dependent on the quality and sufficiency of the magnetic field data. A concise review of magnetic depth estimation methods and practical advice on their application is given by Li (2003).

The Werner method

The Werner method of automatically generating magnetic source solutions from profile magnetic field measurements (Werner, 1953; Hartman et al., 1971) is often referred to as Werner deconvolution. It is not a true deconvolution, but solves for simple geological structures by solution of linear equations applied to profile data

sampled within a moving window. The equations are specified for bodies of nominated structural index. Ku and Sharp (1983) developed a two-stage approach in which the Werner solutions were subsequently submitted to a Marquardt inversion. Hansen and Simmons (1993) extended the Werner method by adapting it to run on an analytic signal transform of the data, and to detect and solve for multiple sources, and Hansen (2005) extended the Werner method to a three-dimensional multiple source analysis.

The Naudy method

The Naudy method (Naudy, 1971) is an elegant approach to finding and solving for magnetic sources from analysis of profiles of magnetic field data. As a window is moved along a magnetic field profile, the data is split into symmetric and asymmetric components about the center of the window. Firstly, the reference anomaly is computed for a target body of the nominal type and specified properties (primarily depth) and a correlation coefficient is derived between that target curve and the processed field data curves. A coefficient value beyond a set threshold flags a possible solution, which is then fed to a second stage inversion process that searches for the optimum source parameters. This operation can be repeated to scan the data across a range of target depths. A key advantage of the Naudy method is that the second stage inversion supplies bodies that can then be delivered into a modelling package for more detailed testing against the data they were derived from, and as a starting model for more interpretive studies. The Naudy method has been extended by Shi (1991) to run on the vertical derivative, and by Almond and Fitzgerald (1998) to adjust for the trend of source bodies automatically estimated from accompanying grid data.

Euler deconvolution

Euler deconvolution solves for Euler's homogeneity equation within a moving window assuming that the field variation within the window is due only to a single source body with a defined value of structural index. Thompson (1982) used a one-dimensional window along a profile, and Reid et al. (1990) extended the analysis to a two-dimensional window of grid data. Sources of error are that the structural index value used may be inappropriate and that the field variation within the window may be due to more than one source, and, as a consequence, the solutions themselves require interpretation. There have been numerous adaptations of the method to try to estimate the appropriate structural index for each solution (e.g., Barbosa et al., 1999) and also to analyze clusters of solutions (e.g., Silva and Barbosa, 2003; Ugalde and Morris, 2010). Euler deconvolution has also been extended to application with the modulus of the analytic signal (Salem and Ravat, 2003; Keating and Pilkington, 2004; Florio et al., 2006) and to solution for multiple sources (Hansen and Suciu, 2002). Other implementations of the Euler homogeneity equation to derivatives of the magnetic field that derive structural index as an integral part of the analysis are termed "extended" Euler methods (Mushayandebvu et al., 2001; Davis et al., 2010). Nabigian and Hansen (2001) have shown that the extended Euler method is a unification of Euler deconvolution and the Werner method. Thurston and Smith (1997) applied Euler's homogeneity equation to expressions of the complex analytic signal in a method termed source parameter imaging (SPI) which recovers source estimates for dipping 2-D contacts and dipping 2-D thin sheets from gridded magnetic field data. This method was further extended to include horizontal cylinders and automatic discrimination between different source models by Smith et al. (1998). These hybrid methods benefit from relationships discussed for simple analytic signal and tilt filter transforms as discussed in the section above on enhancement of magnetic field data, but these methods also require recovery of higher derivatives of the magnetic field, and ultimately it is the validity of those higher derivatives in the presence of noise and processing artifacts that determines the usefulness of these sophisticated methods.

Summary

A substantial toolbox of methods exists to process and enhance magnetic field data, and these capabilities are continuously being extended as computing power and speed increase. Most of the methods require careful inspection of the data to select optimum parameters and to detect any undesirable characteristic in the outputs. Enhancement operations, in particular, are selected and designed for a specific interpretational objective, and different enhancements can be applied to the same data for different objectives. Operations such as gridding and visualization are an integral part of processing and enhancement, and also require careful attention to avoid misrepresentation of data. There is an arbitrary delineation between methods such as multi-scale edge detection, which is described here as an advanced enhancement, and automatic depth estimators, such as Euler deconvolution, which attempt a marginally greater level of discrimination and represent a transition into interpretation. The current trend of automatic source depth estimation is towards ever greater sophistication using high-order derivatives of the field, but these methods place ever greater stress on the validity of the data, such that the derived parameters require careful evaluation.

Bibliography

Almond, R., and Fitzgerald, D. J., 1998. Naudy based automodelling with trend enhancements. *Exploration Geophysics*, **29**, 372–377.

Baranov, V., 1957. A new method for interpretation of aeromagnetic maps: pseudo-gravimetric anomalies. *Geophysics*, **22**, 359–383.

Barbosa, V. C. F., Silva, J. B. C., and Medeiros, W. E., 1999. Stability analysis and improvement of structural index estimation in Euler deconvolution. *Geophysics*, **64**, 48–60.

Blakely, R. J., 1996. *Potential theory in gravity and magnetic applications*. Cambridge: Cambridge University Press.

Blakely, R. J., and Simpson, R. W., 1986. Locating edges of source bodies from magnetic or gravity anomalies. *Geophysics*, **51**, 1494–1498.

Briggs, I. C., 1974. Machine contouring using minimum curvature. *Geophysics*, **39**, 39–48.

Dampney, C. N. G., 1969. The equivalent source technique. *Geophysics*, **34**, 39–53.

Davis, K., Li, Y., and Nabighian, M., 2010. Automatic detection of UXO magnetic anomalies using Euler deconvolution. *Geophysics*, **75**, G13–G20.

Dentith, M., Cowan, D. R., and Tomkins, L. A., 2000. Enhancement of subtle features in aeromagnetic data. *Exploration Geophysics*, **31**, 104–108.

Emilia, D. A., 1973. Equivalent sources used as an analytic base for processing total magnetic field profiles. *Geophysics*, **38**, 339–348.

Florio, G., Fedi, M., and Pasteka, R., 2006. On the application of Euler deconvolution to the analytic signal. *Geophysics*, **71**, L87–L93.

Grauch, V. J. S., and Cordell, L., 1987. Limitations of determining density or magnetic boundaries from the horizontal gradient of gravity or pseudogravity data. *Geophysics*, **52**, 118–121.

Guspi, F., and Novara, I., 2009. Reduction to the pole and transformations of scattered magnetic data using Newtonian equivalent sources. *Geophysics*, **74**, L67–L73.

Hansen, R. O., 2005. 3D multiple-source Werner deconvolution for magnetic data. *Geophysics*, **70**, L45–L51.

Hansen, R. O., and Miyazaki, Y., 1984. Continuation of potential fields between arbitrary surfaces. *Geophysics*, **49**, 787–797.

Hansen, R. O., and Simmons, M., 1993. Multiple-source Werner deconvolution. *Geophysics*, **58**, 1792–1800.

Hansen, R. O., and Suciu, L., 2002. Multiple-source Euler deconvolution. *Geophysics*, **67**, 525–535.

Hartman, R. R., Tesky, D. J., and Friedberg, J. L., 1971. A system for rapid digital aeromagnetic interpretation. *Geophysics*, **36**, 891–918.

Helbig, K., 1963. Some integrals of magnetic anomalies and their relation to the parameters of the disturbing body. *Zeitschrift fur Geophysik*, **29**, 83–96.

Holden, D. J., Archibald, N. J., Boschetti, F., and Jessell, M. W., 2000. Inferring geological structures using wavelt-based multiscale edge analysis and forward models. *Exploration Geophysics*, **31**, 617–621.

Hornby, P., Boschetti, F., and Horowitz, F. G., 1999. Analysis of potential field data in the wavelt domain. *Geophysical Journal International*, **137**, 175–196.

Hsu, S. K., Sibuet, J. C., and Shyu, C. T., 1996. High resolution detection of geologic boundaries from potential-field anomalies: an enhanced analytic signal technique. *Geophysics*, **61**, 373–386.

Keating, P., and Pilkington, M., 2004. Euler deconvolution of the analytic signal and its application to magnetic interpretation. *Geophysical Prospecting*, **50**, 15–25.

Ku, C. C., and Sharp, J. A., 1983. Werner deconvolution for automated magnetic interpretation and its refinement using Marquardt's inverse modelling. *Geophysics*, **48**, 754–774.

Lahti, I., and Karinen, T., 2010. Tilt derivative multiscale edges of magnetic data. *The Leading Edge*, **29**, 24–29.

Li, X., 2003. On the use of different methods for estimating magnetic depth. *The Leading Edge*, **22**, 1090–1099.

Li, X., 2008. Magnetic reduction-to-the-pole at low latitudes: observations and considerations. *The Leading Edge*, **27**, 990–1002.

MacLeod, I. N., Jones, K., and Dai, T. F., 1993. 3-D analytic signal in the interpretation of total magnetic field data at low magnetic latitudes. *Exploration Geophysics*, **24**, 679–688.

Miller, H. G., and Singh, V. J., 1994. Potential field tilt – a new concept for location of potential field sources. *Applied Geophysics*, **32**, 213–217.

Milligan, P. R., and Gunn, P. J., 1997. Enhancement and presentation of airborne geophysical data. *AGSO Journal of Australian Geology and Geophysics*, **17**, 63–75.

Mushayandebvu, M. F., van Driel, P., Reid, A. B., and Fairhead, J. D., 2001. Magnetic source parameters of two-dimensional structures using extended Euler deconvolution. *Geophysics*, **66**, 814–823.

Nabighian, M. N., 1972. The analytic signal of two-dimensional magnetic bodies with polygonal cross-section: its properties and use for automated anomaly interpretation. *Geophysics*, **37**, 505–517.

Nabighian, M. N., 1984. Toward a three-dimensional automatic interpretation of potential field data. *Geophysics*, **49**, 780–786.

Nabigian, M. N., and Hansen, R. O., 2001. Unification of Euler and Werner deconvolution in three dimensions via the generalized Hilbert transform. *Geophysics*, **66**, 1805–1810.

Naudy, H., 1971. Automatic determination of depth on aeromagnetic profiles. *Geophysics*, **36**, 717–722.

O'Connell, M. D., and Owers, M., 2008. A line spacing compression method and an improved minimum curvature operator for grid interpolation of airborne magnetic surveys. *Exploration Geophysics*, **39**, 148–154.

O'Connell, M. D., Smith, R. A., and Vallee, M. A., 2005. Gridding aeromagnetic data using longitudinal and transverse horizontal gradients with the minimum curvature operator. *The Leading Edge*, **24**, 142–145.

Phillips, J. D., 2005. Can we estimate magnetization directions from aeromagnetic data using Helbig's integrals? *Earth, Planets and Space*, **57**, 681–689.

Pilkington, M., and Urquhart, W. E. S., 1990. Reduction of potential field data to a horizontal plane. *Geophysics*, **55**, 549–555.

Rajagoplan, S., 2003. Analytic signal vs. reduction to pole: solutions for low magnetic latitudes. *Exploration Geophysics*, **34**, 257–262.

Reid, A. B., Allsop, J. M., Granser, H., Millett, A. J., and Somerton, I. W., 1990. Magnetic interpretation in three dimensions using Euler deconvolution. *Geophysics*, **55**, 80–91.

Ridsdill-Smith, T. A., and Dentith, M. C., 1999. The wavelet transform in aeromagnetic processing. *Geophysics*, **64**, 1003–1013.

Roest, W. R., Verhoef, J., and Pilkington, M., 1992. Magnetic interpretation using the 3-D analytic signal. *Geophysics*, **57**, 116–125.

Salem, A., and Ravat, D., 2003. A combined analytic signal and Euler method (AN-EUL) for automatic interpretation of magnetic data. *Geophysics*, **68**, 1952–1961.

Salem, A., Williams, S., Fairhead, D., Smith, R., and Ravat, D., 2008. Interpretation of magnetic data using tilt-angle derivatives. *Geophysics*, **73**, L1–L10.

Schmidt, P. W., and Clark, D. A., 1998. The calculation of magnetic components and moments from TMI: a case study from the Tuckers igneous complex, Queensland. *Exploration Geophysics*, **29**, 609–614.

Shi, Z., 1991. An improved Naudy-based technique for estimating depth from magnetic profiles. *Exploration Geophysics*, **22**, 357–362.

Silva, J. B. C., and Barbosa, V. C. F., 2003. 3D Euler deconvolution: theoretical basis for automatically selecting good solutions. *Geophysics*, **68**, 1962–1968.

Smith, R. S., and O'Connell, M. D., 2005. Interpolation and gridding of aliased geophysical data using constrained anisotropic diffusion to enhance trends. *Geophysics*, **70**, V121–V127.

Smith, R. S., Thurston, J. B., Dai, T. F., and MacLeod, I. N., 1998. iSPI^{TM} – the improved source parameter imaging method. *Geophysical Prospecting*, **46**, 141–151.

Spector, A., and Grant, F. S., 1970. Statistical models for interpreting aeromagnetic data. *Geophysics*, **35**, 293–302.
Thompson, D. T., 1982. EULDPH: a new technique for making computer-assisted depth estimates from magnetic data. *Geophysics*, **47**, 31–37.
Thurston, J. B., and Smith, R. S., 1997. Automatic conversion of magnetic data to depth, dip, and susceptibility contrast using the SPI™ method. *Geophysics*, **62**, 807–813.
Ugalde, H., and Morris, W. A., 2010. Cluster analysis of Euler deconvolution solutions: new filtering techniques and geologic strike determination. *Geophysics*, **75**, L61–L70.
Vallée, M. A., Keating, P., Smith, R. S., and St-Hilaire, C., 2004. Estimating depth and model type using the continuous wavelet transform of magnetic data. *Geophysics*, **69**, 191–199.
Verduzco, B., Fairhead, J. D., Green, C. M., and MacKenzie, C., 2004. New insights into magnetic derivatives for structural mapping. *The Leading Edge*, **23**, 116–119.
Werner, S., 1953. Interpretation of magnetic anomalies at sheet-like bodies. Sveriges Geologiska Undersok, Ser. C. Arsbok, 43, no. 6.

Cross-references

Geomagnetic Field, Global Pattern
Geomagnetic Field, Secular Variation
Gravity, Data to Anomalies
Magnetic Anomalies, Interpretation
Magnetic Gradiometry
Magnetic Methods, Airborne
Magnetic Methods, Principles
Magnetic Methods, Surface
Magnetometers
Remanent Magnetism
Wavelet Analysis

MAGNETIC DOMAINS

Susan L. Halgedahl
Department of Geology and Geophysics, University of Utah, Salt Lake City, UT, USA

Definition

Magnetic domain. A magnetic domain is a region in which the magnetic spins of atoms or molecules are aligned. Many magnetic particles consist of two or more magnetic domains, each of uniform magnetization separated by narrow zones called domain walls, in which the spins change orientation from one domain to the next.
NRM (natural remanent magnetization). The magnetic moment (per unit volume of per unit mass) acquired by a rock in nature.
TRM (thermal remanent magnetization). The magnetic moment acquired by a rock as it cools from the Curie temperature in the Earth's field.

Introduction

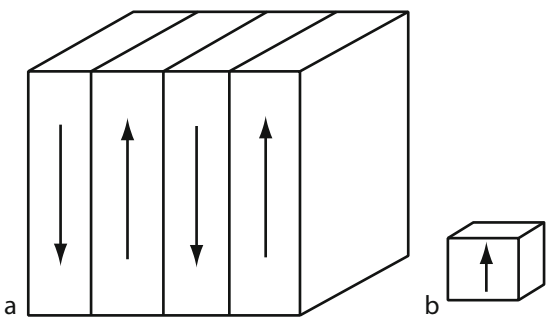
The existence of a paleomagnetic record attests to the ability of magnetic minerals in rocks to preserve their natural remanent magnetizations over geologic time. The building block of *Remanent Magnetism* is the magnetic domain, a region where the magnetic spins are aligned in the same direction. In the early days of paleomagnetism, it was thought that the stable components of natural remanent magnetization (NRM) mainly resided in extremely small magnetic mineral grains that occupied the single-domain (SD) state. However, it is now recognized that, due to their very small size and scarcity, SD particles may not be the major carriers of NRM in many rocks. Instead, it is more likely that much of NRM is carried by grains that, by virtue of their larger sizes, are subdivided into several domains (Figure 1). Adjacent domains are separated by domain walls, the transition regions where magnetic spins change direction from one domain to the next.

Remanences carried by SD assemblages can be highly stable, because destroying the remanence requires rotation of the particles' magnetic moments across strong energy barriers. In contrast, remanences carried by particles with two or more domains usually owe their stability to pinning of walls at crystal defects; these energy barriers are overcome more easily than those controlling SD grains. But another necessary criterion for paleomagnetic stability over geologic time is that a grain's overall domain state resists drastic reorganization caused by the changing physical/chemical conditions affecting the rock, such as moderately elevated temperature, moderate pressure, and some chemical alteration. For this reason, we discuss both theories and experiments to investigate how domain structure depends on grain size, applied field, and temperature in natural magnetic minerals and their synthetic analogues. (See detailed treatments of these subjects in standard texts such as Chikazumi (1964), Cullity (1972), Stacey and Banerjee (1974), and Dunlop and Ozdemir (1997)).

Classical Models of Magnetic Domain Structure

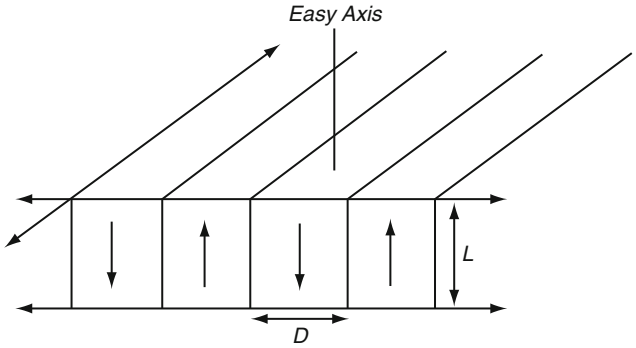
Exchange Energy

Models of magnetic domain structure are based on three energies: exchange energy, anisotropy energy, and magnetostatic energy. Magnetostatic energy is the fundamental reason why particles subdivide into two or more domains because, in so doing, this energy is reduced. However, this



Magnetic Domains, Figure 1 Illustrations of (**a**) a multidomain cube containing four domains and (**b**) a single-domain cube of uniform magnetization.

subdivision carries an energy "price": the energy of domain walls. In most classical theories, the first two energies above make the largest contributions to the surface energy of the wall. From seminal calculations, Heisenberg (1928) predicted the existence of the exchange interaction, which determines whether adjacent magnetic spins are aligned parallel or antiparallel with respect to each other in a specific substance. In a relatively few materials, this interaction leads to a configuration of lowest energy when neighboring spin vectors are aligned. This gives rise to the rare phenomenon of spontaneous magnetization (M_s) and magnetic remanence. Such materials possess a Curie temperature, above which spontaneous magnetism vanishes, because (a) thermal agitation disrupts the alignment of adjacent spins, and (b) thermal expansion of the lattice increases spin–spin distances and thus weakens the exchange interaction.



Magnetic Domains, Figure 2 Illustration of the semi-infinite, magnetized plate of thickness *L*, whose magnetostatic energy was calculated by Kittel (1949). The plate contains domains of identical width *D*, with magnetizations perpendicular to the plate's surface. Walls are assumed to be infinitely thin.

Magnetic anisotropy energy

Magnetic anisotropy causes magnetic behavior to depend on the direction in which it is measured. In magnetite, for example, magnetocrystalline anisotropy results in a lower external field being required to saturate a crystal along the <111> directions than along <100>. Magnetocrystalline anisotropy arises from the coupling between spins and orbits of the electrons, so that work is required to rotate spins out of the directions of lowest energy ("easy" directions). "Easy" and "hard" directions depend on the specific material and its crystal structure. In cubic materials such as magnetite, the easy directions are along <111>, whereas in iron <100> are easy directions (e.g., for more comprehensive discussions, see Chikazumi (1964); Cullity (1972)).

Also due to spin-orbit coupling, the lattice of a ferromagnetic material will spontaneously strain below the Curie temperature. The inverse effect is that external stress and/or variable internal strain can rotate the spontaneous magnetization away from the easy direction given by crystalline anisotropy. Depending on the material and the stress direction, applied stress may cause M_s to rotate either toward the stress axis or toward the direction perpendicular to the axis of stress.

Magnetostatic energy

Magnetostatic energy is the work, per unit volume, required to assemble a population of magnetic "free poles" into a particular geometric configuration. A simple but highly relevant case was addressed by Kittel (1949), who analyzed a semi-infinite plate of thickness L containing lamellar domains of uniform width D, with spontaneous magnetizations normal to the plate surface (Figure 2). Kittel obtained $E_m = 1.705\ M_s^2 D$, where E_m is the magnetostatic energy, per unit area of plate surface (erg/cm^2, in cgs). Kittel's calculations were refined by Rhodes and Rowlands (1954), who addressed finite, rectangular grains. Their results produced "Rhodes and Rowlands" functions, which yield the total magnetostatic energy of rectangular grains of specified axial ratios.

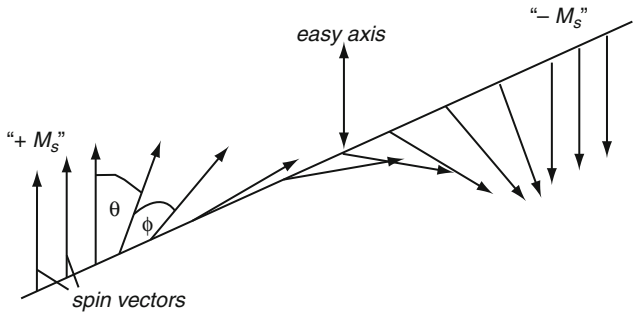
Energy and width of the domain wall

A domain wall is the transition region where the spontaneous magnetization changes direction from one domain to the next, so that each domain is distinct from a wall. (Micromagnetic theories relax this assumption and will be discussed in later.) The two most important energies contributing to the wall's energy and width are exchange energy and anisotropy energy, the latter due either to magnetocrystalline anisotropy or stress. The magnetostatic energy of the wall itself also plays a role; this energy was first analyzed by Amar (1958) and grows especially important when the wall width approaches that of the particle.

In the simplest models, spins in a wall reach an equilibrium configuration when magnetic anisotropy and exchange energies are balanced. For the 180° "Bloch" wall shown in Figure 3 in which spins rotate through 180°, the energy per unit area of wall surface is $E_w = 2\pi (AK)^{1/2}$, where K is the magnetic anisotropy constant due to magnetocrystalline anisotropy and/or stress and A is the material's exchange constant. Note that a submicroscopic change in either A or K can change the local wall energy. Such local changes constitute "wall pinning" sites, where wall motion is impeded by either an energy barrier or an energy trough. The width of a 180° wall is $\delta_w = \pi (A/K)^{1/2}$. In magnetite, values predicted for E_w and δ_w are approximately 0.9 erg/cm^2 and 0.3 μm, respectively (e.g., see Dunlop and Ozdemir, 1997).

Domain width versus grain size

For the magnetized plate illustrated in Figure 2, domain width D can be calculated for the lowest energy state by minimizing the sum of magnetostatic and wall energies per unit volume of material; this gives the half-power law $D = (1/M_s)\ (E_w L/1.705)^{1/2}$ (Kittel, 1949).



Magnetic Domains, Figure 3 Illustration of spins in a 180° Bloch wall. θ is the angle between a spin and the easy axis of magnetization; ϕ is the angle between adjacent spins.

Magnetic minerals in rocks, however, rarely grow as thin platelets. Moskowitz and Halgedahl (1987) calculated the number of domains versus grain size in rectangular particles of $x = 0.6$ titanomagnetite ("TM60": $Fe_{2.4}Ti_6O_4$) containing planar domains separated by 180° walls. Magnetostatic energy, including that from the walls' moments, was accounted for (Rhodes and Rowlands, 1954; Amar, 1958), and it was assumed that grains occupied states of absolute minimum energy. Their calculations yielded two principal results: (1) particles encompassing a wide range of grain sizes could contain the same number of domains, and (2) a plot of N (number of domains) versus L (grain thickness) could be fitted well by a power law $N \propto L^{1/2}$. It follows that domain width D also follows a half-power law in L. Thus, the general type of functional dependence of D on L is the same for both finite, rectangular grains and semi-infinite plates.

Single-domain/two-domain transition size d_o

Grains containing only two or three domains can rival the remanences and coercivities of SD grains, and such particles are often referred to as being "pseudosingle-domain" (PSD). A particle may favor a two-domain over a SD state because the magnetostatic energy associated with two domains is much lower than that for saturation. However, below d_o the energy price of adding a wall is too great to result in a state of minimum energy. At d_o, the total energy of the two-domain state and the SD state are equal. PSD grains can be common in many rocks and, in terms of interpreting rock magnetic behavior, it is important to know the grain sizes corresponding to the onset of PSD behavior in various magnetic minerals.

This transition size depends on the material's magnetic properties, its state of stress, the particle shape, and temperature. Moskowitz and Banerjee (1979) calculated total energies of SD, two-domain, and three-domain cubes of stress-free magnetite containing lamellar domains at room temperature. Magnetostatic energy of the walls was included in their calculations (Amar, 1958). They obtained $d_o \approx 0.08$ μm. Similar calculations for TM60 by Moskowitz and Halgedahl (1987) yielded $d_o \approx 0.5$ μm for unstressed particles at room temperature. Raising the stress level to 100 MPa shifted d_o upward to 1 μm.

Domains and domain walls near crystal surfaces

In relatively thick crystals, domains and domain walls may change their geometric styles near and at crystal surfaces, in order to lower the total magnetostatic energy with respect to that of the "open" structure shown in Figure 2. The particular style depends largely on the dominant kind of anisotropy, as well as on the relative strengths of magnetostatic and anisotropy energies. When $2\pi M_s^2/K >> 1$ in a uniaxial material, prism-shaped closure domains bounded by 90° walls may completely close off magnetic flux at the crystal surface (Figure 4a). Because closure domains greatly reduce magnetostatic energy, the "body" domains can be several times broader than predicted for the "open" structure.

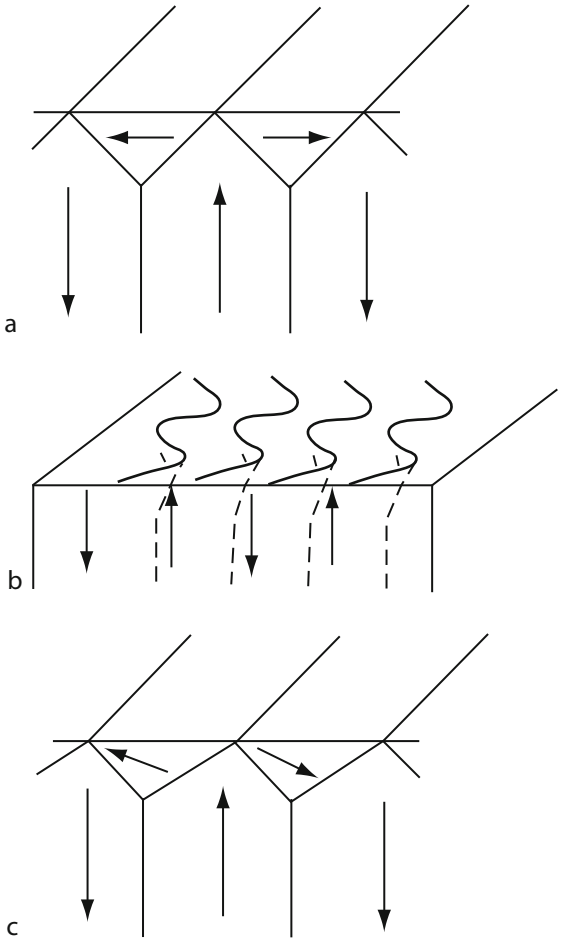
When $2\pi M_s^2/K << 1$ in a uniaxial substance, a large amount of anisotropy energy results when M_s is perpendicular to the "easy" axis, making the style of surface closure shown in Figure 4a energetically unfavorable. Instead, walls that are planar within the body of the crystal can become wavy at the surface (Figure 4b). In extremely thick crystals, wavy walls can alternate with rows of reverse spikes. These elaborate surface domain structures lower magnetostatic energy by achieving a closer mixture of "positive" and "negative" free magnetic poles (e.g., see Szymczak, 1968).

Large crystals governed by cubic magnetocrystalline anisotropy also reduce surface flux through prism-shaped closure domains at the surface. When <100> are easy directions, as in iron, closure domains are bounded by 90° walls (Figure 4a). When <111> are easy directions, as in magnetite, closure domains are bounded by 71° and 109° walls (Figure 4c).

Temperature dependence of domain structure

Understanding how domain structure evolves during both heating to, and cooling from, the Curie point is crucial to understanding the acquisition and thermal stability of thermal remanent magnetization (TRM). If the number of domains changes significantly during cooling from the Curie point in a weak field, it is reasonable to hypothesize that TRM will not become "frozen in," or blocked, until the overall domain structure reaches a stable configuration.

According to Kittel's original model (Figure 2), grains will nucleate (add) domain walls and domains with heating in zero field, if the wall energy drops more rapidly with increasing temperature than does the magnetostatic term. Energy-wise, in this first case, a particle can "afford" to add domains with heating. Conversely, during cooling from the Curie point, a grain will denucleate (lose) domains and domain walls if wall energy rises more quickly than does the magnetostatic energy with decreasing temperature. If wall energy drops less rapidly with



Magnetic Domains, Figure 4 Three styles in which domains and domain walls may terminate at a crystal surface. (**a**) Prism-shaped surface closure domains at the surface of a material which is either uniaxial, with $2\pi M_s^2/K >> 1.0$, or cubic, such as iron, whose easy axes are along <100>. Here, 90° walls separate closure domains from the principal "body" domains that fill most of the crystal. Arrows indicate the sense of spontaneous magnetization within the domains. (**b**) Wavy walls at the surface of a uniaxial material with $2\pi M_s^2/K < 1.0$. Waviness dies out with increasing distance from the surface. (**c**) Prism-shaped closure domains at the surface of a cubic material, such as magnetite, whose easy directions of magnetization are along <111>. Closure domains and body domains are separated by 71° and 109° walls.

increasing temperature than does the magnetostatic term, then the opposite scenarios apply. This predicted behavior relies on the assumptions that the particle is able to maintain a global energy minimum (GEM) domain state at all temperatures and that the total magnetostatic energy of the walls themselves can be ignored.

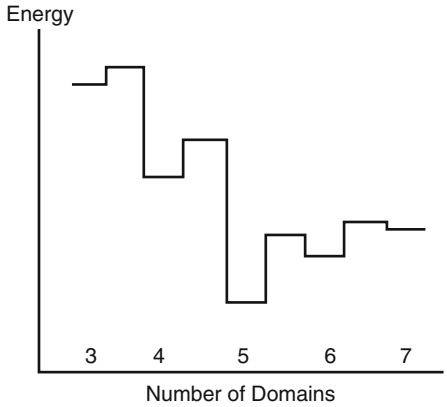
Using Amar's (1958) model, Moskowitz and Halgedahl (1987) calculated the number of domains between room temperature and the Curie point in parallelepipeds of TM60. As discussed earlier, they investigated two cases: dominant crystalline anisotropy (zero stress) and high stress ($\sigma = 100$ MPa). In TM60 grains larger than a few micrometers, most of their results gave an increase in the number of domains with heating. Exceptions to this overall pattern were cases in which walls broadened so dramatically with increasing temperature that they nearly filled the particle and rendered nucleation unfavorable. During cooling from the Curie point in zero field, the domain "blocking temperature" – the temperature below which the number of domains remained constant – increased both with decreasing grain size and with internal stress.

Micromagnetic models

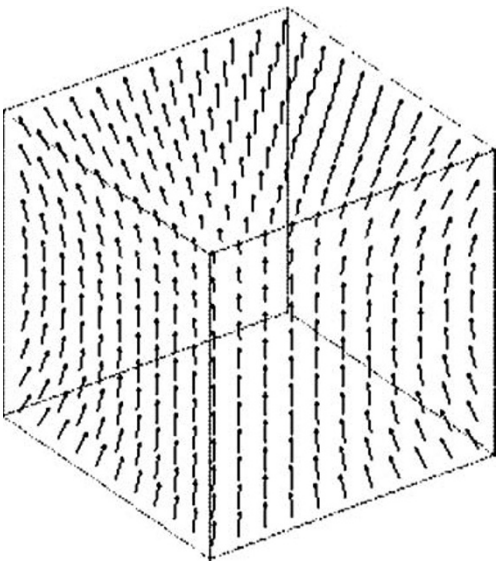
In contrast to classical models of magnetic domain structure, micromagnetic models do not assume the presence of discrete domains. Instead, they allow the orientations of M_s-vectors to vary among extremely small subvolumes, into which a grain is divided. A stable configuration is obtained numerically when the grand sum of exchange, anisotropy, and magnetostatic energies over all subvolumes is minimum. In rock magnetism, micromagnetic models have focused on magnetite, and they assume unstressed, defect-free crystals.

Moon and Merrill (1984, 1985) were the first in rock magnetism to construct one-dimensional (1D) micromagnetic models for defect-free magnetite cubes. Not only did they calculate the energies resulting from different numbers of domains within a grain of given size, but they also calculated the energy barriers associated with nucleation and denucleation of domains and domain walls. Their major breakthrough was the discovery that a particle can occupy a range of local energy minimum (LEM) domain states. Each LEM state is characterized by a unique number of domains and is separated from adjacent states by energy barriers. As illustrated in Figure 5, the GEM state is the configuration of lowest energy but, owing to the energy barriers between states, a LEM state can be quite stable as well.

Several authors have extended micromagnetic calculations for magnetite to two dimensions (2D) and three dimensions (3D) (e.g., Williams and Dunlop, 1989, 1990, 1995; Newell et al., 1993; Xu et al., 1994; Fabian et al., 1996; Fukuma and Dunlop, 1998; Williams and Wright, 1998). These models yield a variety of exotic, nonuniform configurations of magnetization, such as "flower" and "vortex" states. Analogous to Moon and Merrill's results, these models produce both LEM and GEM states, although with very different spin structures from those of 1D models. For example, Figure 6 illustrates a "flower" state in a cube magnetized parallel to the z-axis. The flower state is reminiscent of a classical SD state of uniform magnetization, except that the M_s-vectors are canted at and near the crystal surface. In relatively large cubes of magnetite – for example, 4 or 5 μm – both 2D and 3D models yield configurations closely approaching those of classical domain structures expected for magnetite, such as "body" domains and closure domains at



Magnetic Domains, Figure 5 Diagram illustrating the relative energies and energy barriers associated with local energy minimum (LEM) domain states. Each LEM state is characterized by a unique number of domains and is separated from adjacent states by nucleation and denucleation energy barriers. In this diagram, the global energy minimum, or GEM, domain state has five domains.



Magnetic Domains, Figure 6 Illustration of a "flower" state obtained through three-dimensional micromagnetic modeling of a cube largely magnetized along the cube's z-axis.

the surface (e.g., Xu et al. 1994; Williams and Wright, 1998). In submicron magnetite undergoing hysteresis, magnetization reversal can occur through LEM–LEM transitions (e.g., flower to vortex state) (e.g., Williams and Dunlop, 1995).

Dunlop et al. (1994) used 1D micromagnetic models to investigate LEM states in stress-free magnetite particles during TRM acquisition. Their study was motivated by experimental results reported earlier by Halgedahl (1991) (see discussions below). Dunlop et al. (1994) calculated the energy barriers for all combinations among SD/two-domain/three-domain transitions during cooling from the Curie point of magnetite in a weak external field. According to their results, after acquiring TRM under conditions of thermal equilibrium, populations would be overwhelmingly biased toward GEM domain states, and an individual particle should not exhibit a range of LEM states after several identical TRM runs.

Using renormalization group theory, Ye and Merrill (1995) arrived at a different conclusion. According to their calculations, short-range ordering of spins just below the Curie point could give rise to a variety of LEM states in the same particle after replicate coolings.

Experimental studies of magnetic domains

Methods of imaging domains and domain walls

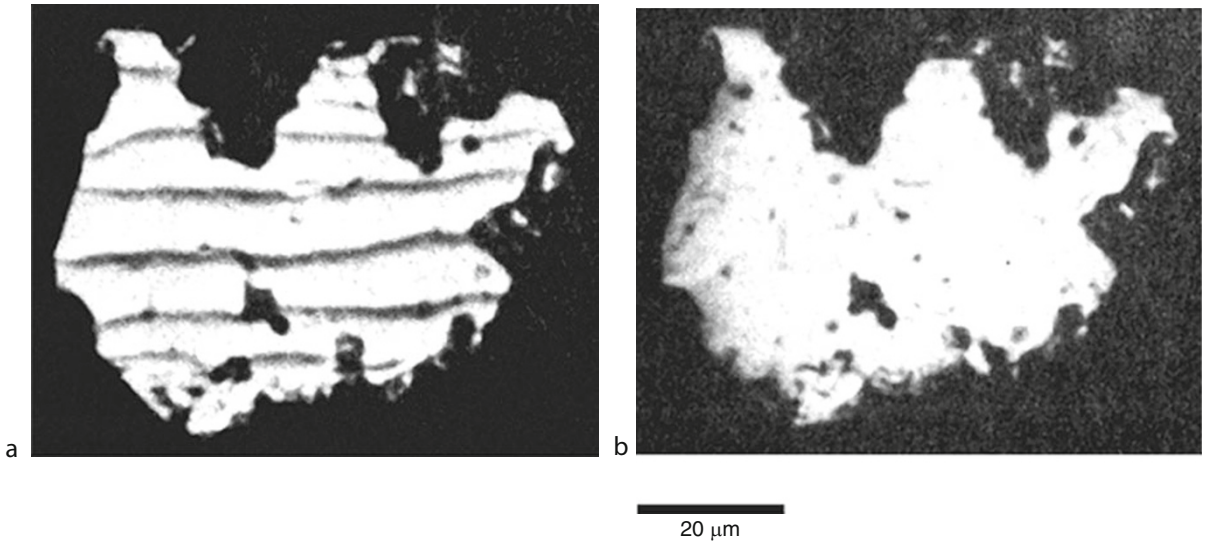
Rock magnetists have mainly used three methods to image domains and domain walls: the Bitter pattern method, which uses liquid magnetic colloid to image walls, with a resolution limit of about 1 μm (e.g., see details in Halgedahl and Fuller, 1983); the magneto-optical Kerr effect (MOKE), which images entire domains from the rotation of the polarization plane of light reflected by a magnetized specimen (e.g., Hoffmann et al., 1987; Worm et al., 1991; Heider and Hoffmann, 1992; Ambatiello et al., 1999); and magnetic force microscopy (MFM), which can image magnetic features as small as 0.01 μm through the voltage induced in a vibrating magnetic needle situated very close to a particle's surface (Williams et al., 1992; Proksch et al., 1994; Moloni et al., 1996; Pokhil and Moskowitz, 1996, 1997; Frandson et al., 2004). Two other methods – transmission electron microscopy (TEM) and off-axis electron holography – have been used in some studies of magnetite, most notably magnetotactic bacteria and magnetite intergrowths in ulvospinel (Dunin-Borowsky et al., 2001; McCartney et al., 2001; Harrison et al., 2002).

Styles of domains observed in magnetic minerals of paleomagnetic significance

In rock magnetism, the majority of domain observation studies have focused on four magnetic minerals, all important to paleomagnetism: pyrrhotite (Fe_7S_8), titanomagnetite of roughly intermediate composition (near $Fe_{2.4}Ti_6O_4$, or TM60), magnetite (Fe_3O_4), and hematite (Fe_2O_3).

Owing to its high magnetocrystalline anisotropy constant and relative insensitivity to stress, pyrrhotite behaves magnetically as a uniaxial material. When studied with the Bitter colloid method, pyrrhotite often exhibits fairly simple domain patterns which suggest lamellar domains separated by 180° walls (Figure 7a) (Soffel, 1977b; Halgedahl and Fuller, 1983).

Despite being cubic, intermediate titanomagnetites rarely, if ever, exhibit the arrays of closure domains, 71°, and 109° walls predicted from theories. Instead, these minerals often display very complex patterns of densely spaced, curved walls (Appel and Soffel, 1984, 1985).



Magnetic Domains, Figure 7 Bitter patterns on a particle of natural pyrrhotite (**a**) after demagnetization in an alternating field of 1,000 Oe and (**b**) in an apparently SD-like state after acquiring saturation remanence in 15 kOe.

However, some grains do exhibit a simple array of parallel walls, such as that shown in Figure 8a (e.g., Halgedahl and Fuller, 1980, 1981), but it is not unusual to observe wavy walls alternating with rows of reverse spikes (Halgedahl, 1987; Moskowitz et al., 1988). Both simple and wavy patterns suggest a dominant, internal stress that yields a uniaxial anisotropy, although the origin of this stress is still unclear.

Small magnetite grains grown in a synthetic rock-like matrix with the glass-ceramic method (Worm and Markert, 1987) generally display simple arrays of straight walls (e.g., Worm et al., 1991; Geiβ et al., 1996) (Figure 8b). In such samples, there appears to be a paucity of closure domains, perhaps the result of high internal stress generated by quenching during synthesis.

In contrast, domain images obtained from large crystals of natural magnetite and small magnetite grains in a variety of rocks reveal 180°, 71°, and 109° walls and closure domains, in accordance with theory (Bogdanov and Vlasov, 1965, 1966; Smith, 1980; Boyd et al., 1984; Ozdemir and Dunlop, 1993, 1997, 2006; Ozdemir et al., 1995; Ambatiello et al., 1999). Evidently, internal stress is low in these natural magnetites, so that magnetocrystalline anisotropy is dominant.

To date, domain studies on hematite have been limited to large (e.g., 100 μm to 1 mm) platelets. Even large crystals such as these contain very few walls and a fairly simple domain structure, owing to hematite's weak spontaneous magnetization (about 2 emu/cm^3) and low magnetostatic energy (Figure 8c) (Halgedahl, 1995, 1998).

Number of domains versus grain size

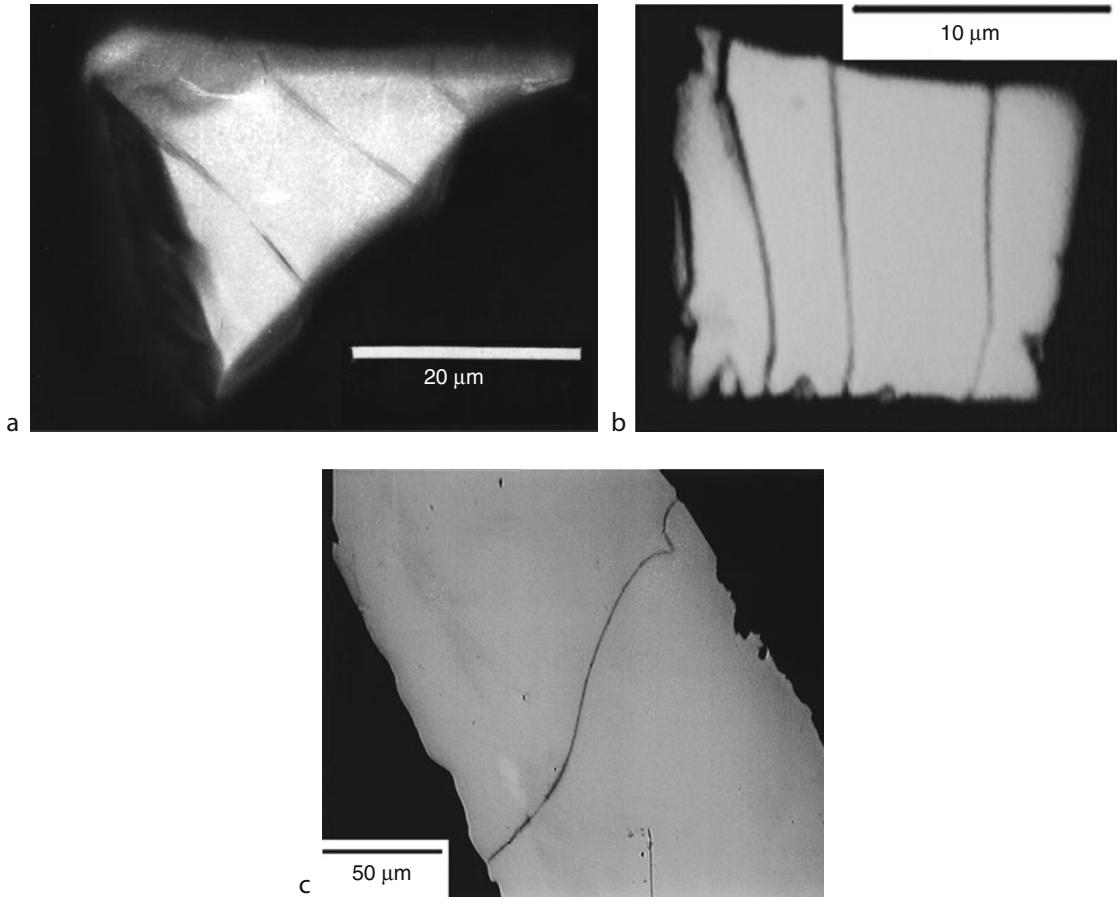
Both Kittel's original model of domains in a semi-infinite platelet and calculations for finite grains by (e.g., Moskowitz and Halgedahl, 1987) lead to the prediction that domain width $D \propto L^{1/2}$, where L is plate or particle thickness. These predictions are supported by domain studies of natural magnetic minerals which generally yield a power-law dependence of D on L, although the power may differ somewhat from 0.5.

Soffel (1971) was the first to study the grain-size dependence of the number of domains in a paleomagnetically important magnetic mineral. From Bitter patterns on grains of natural, intermediate ($x \approx 0.55$) titanomagnetite in a basalt, he determined that, on average, both D and N (number of domains = L/D) $\propto L^{1/2}$, where L was average particle size. His results yielded a single domain/two domain transition size of 0.6 μm and a wall energy density of about 1 erg/cm^2 for this composition.

Similarly, from Bitter patterns on natural pyrrhotite in a Bavarian diabase, both Soffel (1977b) and Halgedahl and Fuller (1983) found that $D \propto L^{0.40}$ to $D \propto L^{0.45}$, depending on the magnetization state. For this pyrrhotite sample, the SD boundary fell between 1.5 and 2 μm. Geiβ et al. (1996) obtained $D \propto L^{0.45}$ from Bitter patterns on synthetic magnetite particles grown in a glass-ceramic matrix (Worm and Markert, 1987). They obtained a SD/two-domain transition size for magnetite of approximately 0.25 μm.

Domain wall widths

When studied in the scanning electron microscope (SEM), patterns of dried magnetic colloid afford high-resolution views of domain walls. Moskowitz et al. (1988) applied this method to polycrystalline pellets of synthetic titanomagnetite substituted with aluminum and magnesium ($Fe_{2.2}Al_{0.1}Mg_{0.1}Ti_{0.6}O_4$: "AMTM60"). Dried Bitter patterns on unpolished surfaces were virtually identical



Magnetic Domains, Figure 8 (**a**) Bitter pattern on a grain of intermediate titanomagnetite ($x \approx 0.6$) in oceanic basalt drilled near the Mid-Atlantic Ridge. (**b**) Bitter pattern on a grain of magnetite synthesized with the glass-ceramic method. In this particular state of magnetization, the grain contains four walls, whose lengths are commensurate with the particle's length. Note that one wall is pinned very near the particle's extreme left-hand edge. The small triangular patterns at the lower edge of the particle represent walls which enclose small reverse spike domains. (**c**) Bitter pattern of a wall on a very large (approximately 150 μm width, 1 mm length) platelet of natural hematite from Elba, Italy. At most this platelet exhibits only 1–2 principal walls, although small edge domains are often observed. Apparently, the wall is bowing around a defect near the upper right-hand edge of the photograph. Only a small part of the crystal surface is shown.

in style to those common to materials controlled by strong uniaxial anisotropy. Patterns indicated closely spaced stripe domains, sinusoidally wavy walls, and nested arrays of reverse spikes. Measurements from SEM images yielded wall widths between 0.170 and 0.400 μm.

Very high resolution images of domains and domain walls in magnetite have been obtained with MFM. The first images were reported by Williams et al. (1992), who recorded a magnetic profile across a 180° wall on a {110} surface of a large natural magnetite crystal. Spins within the wall reversed their polarity of rotation along the length of the wall, demonstrating that walls in real materials can be much more complex than those portrayed by simple models. Using MFM, Pokhil and Moskowitz (1996, 1997) made a similar finding in glass-ceramic magnetite. Segments of opposite polarity were separated by Bloch lines, the transition regions where polarity changes sense. The number of Bloch lines within a wall varied among repeated alternating field (AF) demagnetization treatments. Thus walls, like particles, can occupy LEM states.

Owing to its high-resolution capabilities, the MFM can provide estimates of wall width. Proksch et al. (1994) obtained MFM profiles across a 180° wall on a {110} surface of natural magnetite, thus obtaining a wall width of about 0.21 μm.

Experimental evidence for local energy minimum (LEM) domain states

In their study of Bitter patterns on natural pyrrhotite, Halgedahl and Fuller (1983) noted that grains of virtually identical size could contain very different numbers of domains, despite these same particles having undergone

the same magnetic treatments. Moreover, they found that an individual particle could arrive in very different domain states – that is, with different numbers of walls – after different cycles of minor hysteresis.

These observations led Halgedahl and Fuller (1983) to the conclusion that a particle could occupy domain states other than the ground state. Subsequent domain observation studies by Halgedahl on AMTM60 and by Geiβ et al. (1996) on magnetite provided strong evidence for LEM states in other magnetic minerals.

LEM states and thermomagnetic treatments

A particularly unexpected type of LEM state in pyrrhotite and intermediate titanomagnetite is a SD-like state that certain particles can occupy after being saturated in a strong external field, even though these same particles readily accommodate walls in other states of magnetization. Bitter patterns on intermediate ($x \approx 0.6$) titanomagnetite in oceanic basalt and on natural pyrrhotite in diabase were studied during hysteresis by Halgedahl and Fuller (1980, 1983). In states of saturation remanence, most particles in a large population contained one or more walls, as expected on the basis of previous theories (Figure 7a). However, several tens of percent of the finer (5–15 μm) particles appeared saturated after the maximum field was shut off (Figure 7b). Halgedahl and Fuller (1980, 1983) proposed that grains that failed to nucleate walls could make a substantial contribution to saturation remanence. Particles in SD-like states remained saturated, until nucleation was accomplished by applying a back field of sufficient magnitude, and this nucleation field dropped off with increasing grain size L according to a power law in $L^{-1/2}$. Similarly, Boyd et al. (1984) reported Bitter patterns on natural magnetite grains carrying saturation remanence, and these patterns suggested SD-like states. These results were surprising, in view of magnetite's strong tendency to self-demagnetize.

Subsequent domain observations and hysteresis measurements of hematite platelets from Elba, Italy demonstrated that, after exposure to strong fields, even large crystals could arrive in states of near-saturation. Back fields were necessary to nucleate principal domains, and, similar to pyrrhotite, these nucleation fields also followed a power law in $L^{-1/2}$ (Halgedahl, 1995, 1998).

Results from titanomagnetite and pyrrhotite suggesting SD-like states were interpreted by Halgedahl and Fuller (1980, 1983) in light of previous theoretical and experimental work on high-anisotropy, industrial magnets. In such materials, nucleation of walls is an energetically difficult process, because the internal demagnetizing field, which triggers nucleation, may be too weak to overcome the anisotropy field, which hinders nucleation (Brown, 1963; Becker, 1969, 1971a, b, 1976). Such particles may require an external back field, which aids the demagnetizing field in order to nucleate walls and accompanying domains.

The style of domain structure and the range of LEM states that a particle can occupy can depend on thermomagnetic history. From Bitter patterns on natural, polycrystalline pyrrhotite, Halgedahl and Fuller (1981) discovered that crystallites displayed arrays of undulating walls on their surfaces after acquiring TRM in a weak external field. By contrast, after AF demagnetization in a strong peak field the same crystallites exhibited planar walls. Evidently, cooling locked in a high-temperature configuration of walls whose curved shapes promoted lower magnetostatic energy than would a planar geometry.

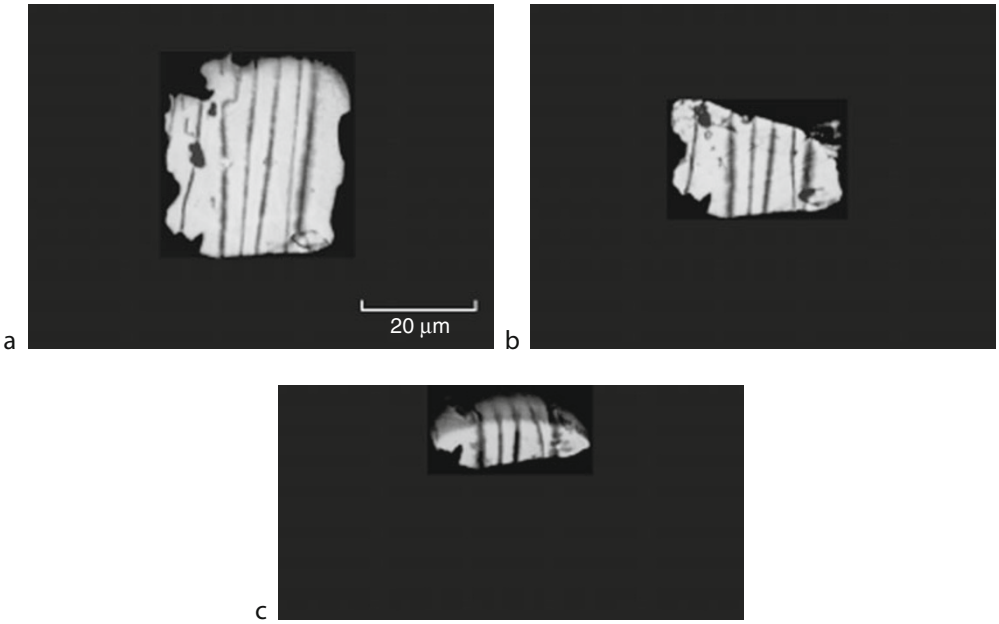
As indicated by Moon and Merrill's theoretical results for magnetite, a particular LEM state can be stable across a broad range of grain size. This prediction is born out by experiments by Halgedahl and Ye (2000), who investigated the effects of mechanical thinning on domain states in natural pyrrhotite. Individual pyrrhotite grains in a diabase were mechanically thinned and their Bitter patterns observed after each thinning step. Despite some grains being thinned to about one-fourth of their initial diameter, the widths of surviving domains and positions of surviving walls remained unaffected (Figure 9). Neither nucleations nor denucleations were observed, although calculations indicated that thinning would cause significant changes in GEM domain states.

Bitter patterns on crystallites of polycrystalline AMTM60 were studied after many replicate TRM acquisition and AF demagnetization experiments (Halgedahl 1991). In each particle, the number of domains varied from one experiment to the next, defining a distribution of LEM states. For states of weak-field TRM, this distribution could be broad and could include single-domain-like states (Figure 10). After replicate AF demagnetizations, however, typical distributions were narrow and clustered about a most probable state, possibly the GEM state.

Evolution of magnetic domain structures at elevated temperatures

The manner in which domain structure evolves with temperature carries clear implications for TRM. Cooling-induced nucleations and denucleations of walls/domains could trigger sudden changes of the internal demagnetizing field. As a result, preexisting walls which survive a domain transition could be dislodged from imperfections where they were pinned initially at higher temperatures.

The first Bitter patterns observed on magnetite above room temperature in the Earth's field were made by Heider et al. (1988), using hydrothermally recrystallized crystals. Depending on the particle, Bitter patterns could be followed to approximately 200°C, above which temperature the patterns grew too faint to discern. Surprisingly, heating to very moderate temperatures drove certain walls across much of a particle; in some cases, denucleation occurred. Upon cooling, walls reassembled in a similar, though not identical, arrangement to that observed initially at room temperature. In some cases, repeated thermal cycling between room temperature and about 200°C produced different numbers of domains in the same particle.



Magnetic Domains, Figure 9 Bitter patterns on a grain of natural pyrrhotite in a diabase before and after mechanical thinning. (**a**) Initial state, before thinning. (**b**) After thinning the particle to about one-half of its original length along a direction parallel to the trends of the Bitter lines. (**c**) Final state, after the particle has been thinned to about one-fourth of its original length.

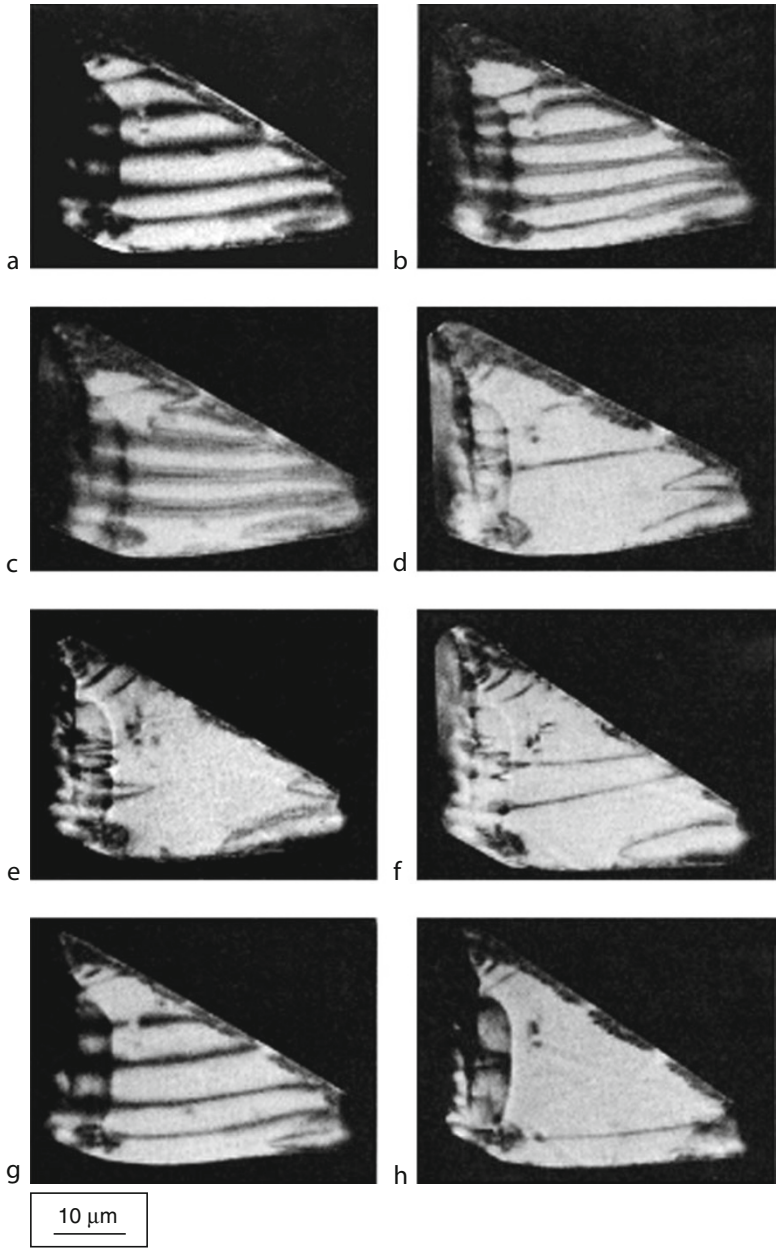
Ambatiello et al. (1999) used the MOKE to study domain widths versus temperature in several large (several mm) crystals of natural magnetite. At room temperature, {110} planes were dominated by broad lamellar domains that terminated in closure domains at crystal edges. On {111} planes, they found complex, nested arrays of very small closure domains, which finely subdivided the main closure structure. Domain widths generally increased with heating, and such changes were thermally reversible. To explain their results, these authors hypothesized that heating promoted an increasingly finer subdivision into small domains near the crystal surface; this would reduce magnetostatic energy with heating and thus cause broadening of body domains (also see Muxworthy and Williams, 2006).

Bitter patterns observed at elevated temperatures on natural, intermediate titanomagnetites were reported by Soffel (1977a), Metcalf and Fuller (1987a, b, 1988), and Halgedahl (1987). During heating, patterns gradually faded to obscurity, with few significant changes in domain structure. Remarkably, some titanomagnetite particles studied by Metcalf and Fuller (1987a, b, 1988) displayed no Bitter lines after cooling from the Curie point in the Earth's field. This observation provided evidence for Halgedahl and Fuller's (1983) earlier proposal that weak-field TRM acquired by populations of PSD grains could, in part, be attributable to particles which failed to nucleate walls during cooling.

The possible importance of LEM states to TRM acquisition was raised by work on synthetic, polycrystalline AMTM60 with a Curie point near 75°C (Halgedahl, 1991). As shown in Figure 10, replicate TRM experiments often produced different numbers of domains in the same particle. Some of these states suggested that the particle was entirely saturated, with no visible Bitter lines, or nearly saturated, with only small spike domains at grain boundaries. Patterns that evolved during cooling revealed that denucleation was one mechanism by which a particle arrived in a final LEM state. Denucleation occurred either through contraction of large, preexisting spike domains, or by straight walls moving together and coalescing into spikes. Often, spikes would collapse altogether. In some cases, denucleation left behind large volumes which, apparently, contained no walls, although walls were present elsewhere in the grain. In other cases, denucleation left behind a grain that appeared nearly saturated, but for small edge domains. These results led to the idea of trans-domain TRM (Halgedahl, 1991).

Magnetic domain structure in magnetite at low temperatures

Magnetite undergoes two types of transitions at low temperatures, which can profoundly affect domain structure and remanence (e.g., see detailed discussions in Stacey and Banerjee (1974); Dunlop and Ozdemir (1997)). At the isotropic point (approximately 130 K), the first magnetocrystalline anisotropy constant, K_1, passes through zero as it changes sign from negative at temperatures above the transition to positive below it. At the



Magnetic Domains, Figure 10 Bitter patterns on a crystallite of titanomagnetite ("AMTM60": see text) after each of eight replicate TRM acquisition runs in the Earth's field.

Verwey transition, T_v (approximately 120 K), magnetite undergoes a crystallographic transition from cubic to monoclinic.

Thus at 130 K domain walls should broaden dramatically, because wall width is proportional to $K_1^{-1/2}$. It follows that, by passage through 130 K, walls may break free of narrow defects which pinned them at other temperatures. At the Verwey transition, the easy axis of magnetization changes direction. In multidomain and pseudosingle-domain magnetite, these thermal passages may cause demagnetization and complete reorganization of domain structure. Low-temperature demagnetization has proved useful in removing certain spurious components of NRM which are surprisingly resistant to thermal demagnetization above room temperature. Thus it is important to determine how these transitions affect domains.

Using an MFM specially adapted to operate at low temperature, Moloni et al. (1996) were the first to image domains on {110} magnetite planes at temperatures near

the two transitions. As the crystal was warmed from 77 K to a few degrees just below T_v, the domain structure disappeared entirely below the instrument's noise level, evidently undergoing a complete reorganization near the crystallographic transition.

Bibliography

Ambatiello, A., Fabian, K., and Hoffmann, V., 1999. Magnetic domain structure of multidomain magnetite as a function of temperature: observation by Kerr microscopy. *Physics of the Earth and Planetary Interiors*, **112**, 55–80.

Amar, H., 1958. Magnetization mechanism and domain structure of multidomain particles. *Physical Review*, **111**, 149–153.

Appel, E., and Soffel, H. C., 1984. Model for the domain state of Ti-rich titanomagnetites. *Geophysical Research Letters*, **11**, 189–192.

Appel, E., and Soffel, H. C., 1985. Domain state of Ti-rich titanomagnetites deduced from domain structure observations and susceptibility measurements. *Journal of Geophysics*, **56**, 121–132.

Becker, J. J., 1969. Observations of magnetization reversal in cobalt-rare-earth particles. *IEEE Transactions on Magnetics*, **MAG-5**, 211–214.

Becker, J. J., 1971a. Magnetization discontinuities in cobalt-rare-earth particles. *Journal of Applied Physics*, **42**, 1537–1538.

Becker, J. J., 1971b. Interpretation of hysteresis loops of cobalt-rare-earths. *IEEE Transactions on Magnetics*, **MAG-7**, 644–647.

Becker, J. J., 1976. Reversal mechanism in copper-modified cobalt-rare-earths. *IEEE Transactions on Magnetics*, **MAG-12**, 965–967.

Bogdanov, A. K., and Vlasov, A. Ya., 1965. Domain structure in a single crystal of magnetite, (English trans.), Izv. *Earth Physics*, series no. 1, 28–32.

Bogdanov, A. A., and Vlasov, A. Ya., 1966. The domain structure of magnetite particles, (English trans.), Izv. *Physics of the Solid Earth*, **9**, 577–581.

Boyd, J. R., Fuller, M., and Halgedahl, S., 1984. Domain wall nucleation as a controlling factor in the behaviour of fine magnetic particles in rocks. *Geophysical Research Letters*, **11**, 193–196.

Brown, W. F., 1963. *Micromagnetics*. New York: John Wiley, 143 pp.

Chikazumi, S., 1964. *Physics of Magnetism*. New York: John Wiley, 664 pp.

Cullity, B. D., 1972. *Introduction to Magnetic Materials*. Reading, MA: Addison-Wesley, 666 pp.

Dunin-Borowsky, R. E., McCartney, M. R., Posfai, M., Frankel, R. B., Bazylinski, D. A., and Buseck, P. R., 2001. Off-axis electron holography of magnetotactic bacteria: magnetic microstructure of strains MV-1 and MS-1. *European Journal of Mineralogy*, **13**(4), 671–684.

Dunlop, D. J., Newell, A. J., and Enkin, R. J., 1994. Transdomain thermoremanent magnetization. *Journal of Geophysical Research*, **99**, 19741–19755.

Dunlop, D. J., and Ozdemir, O., 1997. *Rock Magnetism: Fundamentals and Frontiers*. Cambridge, UK: Cambridge University Press, 573 pp.

Fabian, K., Kirchner, A., Williams, W., Heider, F., and Leibl, T., 1996. Three-dimensional micromagnetic calculations for magnetite using FFT. *Geophysical Journal International*, **124**, 89–104.

Foss, S., Moskowitz, B. M., Proksch, R., and Dahlberg, E. D., 1998. Domain wall structures in single-crystal magnetite investigated by magnetic force microscopy. *Journal of Geophysical Research*, **103**, 30551–30560.

Foss, S., Moskowitz, B., and Walsh, B., 1996. Localized micromagnetic perturbation of domain walls in magnetite using a magnetic force microscope. *Applied Physics Letters*, **69**, 3426–3428.

Frandson, C., Stipp, S. L. S., McEnroe, S. A., Madsen, M. B., and Knudsen, J. M., 2004. Magnetic domain structures and stray fields of individual elongated magnetite grains revealed by magnetic force microscopy (MFM). *Physics of the Earth and Planetary Interiors*, **141**, 121–129.

Fukuma, K., and Dunlop, D. J., 1998. Grain-size dependence of two-dimensional micromagnetic structures for pseudo-single-domain magnetite (0.2–2.5 μm). *Geophysical Journal International*, **134**, 843–848.

Geiβ, C. E., Heider, F., and Soffel, H. C., 1996. Magnetic domain observations on magnetite and titanomaghemite grains (0.5–10 μm). *Geophysical Journal International*, **124**, 75–88.

Halgedahl, S. L., 1987. Domain pattern observations in rock magnetism: progress and problems. *Physics of the Earth and Planetary Interiors*, **46**, 127–163.

Halgedahl, S. L., 1991. Magnetic domain patterns observed on synthetic Ti-rich titanomagnetite as a function of temperature and in states of thermoremanent magnetization. *Journal of Geophysical Research*, **96**, 3943–3972.

Halgedahl, S. L., 1995. Bitter patterns versus hysteresis behavior in small single particles of hematite. *Journal of Geophysical Research*, **100**, 353–364.

Halgedahl, S. L., 1998. Barkhausen jumps in larger versus small platelets of natural hematite, *Journal of Geophysical Research*, **103**, 30575–30589.

Halgedahl, S., and Fuller, M., 1980. Magnetic domain observations of nucleation processes in fine particles of intermediate titanomagnetite. *Nature*, **288**, 70–72.

Halgedahl, S. L., and Fuller, M., 1981. The dependence of magnetic domain structure upon magnetization state in polycrystalline pyrrhotite. *Physics of the Earth and Planetary Interiors*, **26**, 93–97.

Halgedahl, S., and Fuller, M., 1983. The dependence of magnetic domain structure upon magnetization state with emphasis upon nucleation as a mechanism for pseudo-single domain behavior. *Journal of Geophysical Research*, **88**, 6505–6522.

Halgedahl, S. L., and Ye, J., 2000. Observed effects of mechanical grain-size reduction on the domain structure of pyrrhotite. *Earth and Planetary Science Letters*, **178**, 457–467.

Harrison, T. J., Dunin-Borkowski, R. E., and Putnis, A., 2002. Direct imaging of nanoscale magnetic interactions in minerals. *Proceedings of the National Academy of Sciences*, **99**, 16556–16561.

Heider, F., 1990. Temperature dependence of domain structure in natural magnetite and its significance for multi-domain TRM models. *Physics of the Earth and Planetary Interiors*, **65**, 54–61.

Heider, F., Halgedahl, S. L., and Dunlop, D. J., 1988. Temperature dependence of magnetic domains in magnetite crystals. *Geophysical Research Letters*, **15**, 499–502.

Heider, F., and Hoffmann, V., 1992. Magneto-optical Kerr effect on magnetite crystals with externally applied fields. *Earth and Planetary Science Letters*, **108**, 131–138.

Heisenberg, W., 1928. Zur theorie de Ferromagnetismus. *Zeitschrift für Physik*, **49**, 619–636.

Hoffmann, V., Schafer, R., Appel, E., Hubert, A., and Soffel, H., 1987. First domain observations with the magneto-optical Kerr effect on Ti-ferrites in rocks and their synthetic equivalents. *Journal of Magnetism and Magnetic Materials*, **71**, 90–94.

Kittel, C., 1949. Physical theory of ferromagnetic domains. *Reviews of Modern Physics*, **21**, 541–583.

McCartney, M. R., Lins, U., Farina, M., Buseck, P. R., and Frankel, R. B., 2001. Magnetic microstructure of bacterial magnetite by electron holography. *European Journal of Mineralogy*, **13**(4), 685–689.

Metcalf, M., and Fuller, M., 1987a. Magnetic remanence measurements of single particles and the nature of domain patterns in titanomagnetites. *Geophysical Research Letters*, **14**, 1207–1210.
Metcalf, M., and Fuller, M., 1987b. Domain observations of titanomagnetites during hysteresis at elevated temperatures and thermal cycling. *Physics of the Earth and Planetary Interiors*, **46**, 120–126.
Metcalf, M., and Fuller, M., 1988. A synthetic TRM induction curve for fine particles generated from domain observations. *Geophysical Research Letters*, **15**, 503–506.
Moloni, K., Moskowitz, B. M., and Dahlberg, E. D., 1996. Domain structures in single crystal magnetite below the Verwey transition as observed with a low-temperature magnetic force microscope. *Geophysical Research Letters*, **23**, 2851–2854.
Moon, T., and Merrill, R. T., 1984. The magnetic moments of non-uniformly magnetized grains. *Physics of the Earth and Planetary Interiors*, **34**, 186–194.
Moon, T. S., and Merrill, R. T., 1985. Nucleation theory and domain states in multidomain magnetic material. *Physics of the Earth and Planetary Interiors*, **37**, 214–222.
Moskowitz, B. M., and Banerjee, S. K., 1979. Grain size limits for pseudosingle domain behavior in magnetite: implications for paleomagnetism. *IEEE Transactions on Magnetics*, **MAG-15**, 1241–1246.
Moskowitz, B. M., and Halgedahl, S. L., 1987. Theoretical temperature and grain-size dependence of domain state in x = 0.6 titanomagnetite. *Journal of Geophysical Research*, **92**, 10667–10682.
Moskowitz, B. M., Halgedahl, S. L., and Lawson, C. A., 1988. Magnetic domains on unpolished and polished surfaces of titanium-rich titanomagnetite. *Journal of Geophysical Research*, **93**, 3372–3386.
Muxworthy, A. R., and Williams, W., 1999. Micromagnetic calculations of hysteresis as a function of temperature in pseudo-single domain magnetite. *Geophysical Research Letters*, **26**, 1065–1068.
Muxworthy, A. R., and Williams, W., 2006. Observations of viscous magnetization in multidomain magnetite. *Journal of Geophysical Research*, **111**, B01103, doi:10.1029/2005JB003902.
Newell, A. J., Dunlop, D. J., and Williams, W., 1993. A two-dimensional micromagnetic model of magnetization and fields in magnetite. *Journal of Geophysical Research*, **98**, 9533–9549.
Ozdemir, O., and Dunlop, D. J., 1993. Magnetic domain structures on a natural single crystal of magnetite. *Geophysical Research Letters*, **20**, 1835–1838.
Ozdemir, O., and Dunlop, D. J., 1997. Effect of crystal defects and internal stress on the domain structure and magnetic properties of magnetite. *Journal of Geophysical Research*, **102**, 20211–20224.
Ozdemir, O., and Dunlop, D. J., 2006. Magnetic domain observations on magnetite crystals in biotite and hornblend grains. *Journal of Geophysical Research*, **111**, B06103, doi:10.1029/2005JB004090.
Ozdemir, O., Xu, S., and Dunlop, D. J., 1995. Closure domains in magnetite. *Journal of Geophysical Research*, **100**, 2193–2209.
Pokhil, T. G., and Moskowitz, B. M., 1996. Magnetic force microscope study of domain wall structures in magnetite. *Journal of Applied Physics*, **79**, 6064–6066.
Pokhil, T. G., and Moskowitz, B. M., 1997. Magnetic domains and domain walls in pseudo-single-domain magnetite studied with magnetic force microscopy. *Journal of Geophysical Research*, **102**, 22681–22694.
Proksch, R. B., Foss, S., and Dahlberg, E. D., 1994. High resolution magnetic force microscopy of domain wall fine structures. *IEEE Transactions on Magnetics*, **30**, 4467–4472.
Rhodes, P., and Rowlands, G., 1954. Demagnetizing energies of uniformly magnetised rectangular blocks. *Proceedings of the Leeds Philosophical and Literary Society, Science Section*, **6**, 191–210.
Sahu, A., and Moskowitz, B. M., 1995. Thermal dependence of magnetocrystalline anisotropy and magnetostriction constants of single crystal $Fe_{2.4}$ $Ti_{0.6}$ O_4. *Geophysical Research Letters*, **22**, 449–452.
Smith, P. P. K., 1980. The application of Lorentz electron microscopy to the study of rock magnetism. *Institute of Physics Conference Series*, **52**, 125–128.
Soffel, H., 1971. The single-domain-multidomain transition in natural intermediate titanomagnetites. *Zeitschrift fur Geophysik*, **37**, 451–470.
Soffel, H. C., 1977a. Domain structure of titanomagnetites and its variation with temperature. *Journal of Geomagnetism and Geoelectricity*, **29**, 277–284.
Soffel, H., 1977b. Pseudo-single-domain effects and single-domain multidomain transition in natural pyrrhotite deduced from domain structure observations. *Journal of Geophysics*, **42**, 351–359.
Soffel, H. C., Aumuller, C., Hoffmann, V., and Appel, E., 1990. Three-dimensional domain observations of magnetite and titanomagnetites using the dried colloid SEM method. *Physics of the Earth and Planetary Interiors*, **65**, 43–53.
Stacey, F. D., and Banerjee, S. K., 1974. *The Physical Principles of Rock Magnetism*. Amsterdam: Elsevier, 195 pp.
Szymczak, R., 1968. The magnetic structure of ferromagnetic materials of uniaxial structure. *Electronics Technology*, **1**, 5–43.
Williams, W., and Dunlop, D. J., 1989. Three-dimensional micromagnetic modelling of ferromagnetic domain structure. *Nature*, **337**, 634–637.
Williams, W., and Dunlop, D. J., 1990. Some effects of grain shape and varying external magnetic fields on the magnetic structure of small grains of magnetite. *Physics of the Earth and Planetary Interiors*, **65**, 1–14.
Williams, W., and Dunlop, D. J., 1995. Simulation of magnetic hysteresis in pseudo-single-domain grains of magnetite. *Journal of Geophysical Research*, **100**, 3859–3871.
Williams, W., Hoffmann, V., Heider, F., Goddenhenreich, T., and Heiden, C., 1992. Magnetic force microscopy imaging of domain walls in magnetite. *Geophysical Journal International*, **111**, 417–423.
Williams, W., and Wright, T. M., 1998. High-resolution micromagnetic models of fine grains of magnetite. *Journal of Geophysical Research*, **103**, 30537–30550.
Worm, H.-U., and Markert, H., 1987. The preparation of dispersed titanomagnetite particles by the glass-ceramic method. *Physics of the Earth and Planetary Interiors*, **46**, 263–270.
Worm, H.-U., Ryan, P. J., and Banerjee, S. K., 1991. Domain size, closure domains, and the importance of magnetostriction in magnetite. *Earth and Planetary Science Letters*, **102**, 71–78.
Xu, S., Dunlop, D. J., and Newell, A. J., 1994. Micromagnetic modelling of two-dimensional domain structures in magnetite. *Journal of Geophysical Research*, **99**, 9035–9044.
Ye, J., and Halgedahl, S. L., 2000. Theoretical effects of mechanical grain-size reduction on GEM domain states in pyrrhotite. *Earth and Planetary Science Letters*, **178**, 73–85.
Ye, J., and Merrill, R. T., 1995. The use of renormalization group theory to explain the large variation of domain states observed in titanomagnetites and implications for paleomagnetism. *Journal of Geophysical Research*, **100**, 17899–17907.

Cross-references

Paleomagnetic Field Intensity
Paleomagnetism, Principles
Remanent Magnetism

MAGNETIC GRADIOMETRY

Harald von der Osten-Woldenburg
National Heritage Department, Regional Government of Baden-Wuerttemberg, Esslingen am Neckar, Germany

Definition

Magnetometer. Instrument with a single sensor that measures magnetic flux density.
Magnetic gradiometer. Pairs of magnetometers with sensors separated by a fixed distance.

The earth's magnetic field at any point on or near the earth's surface is the vector sum of the contributions from the primary field due to the dynamo in the earth's liquid core and the crustal field from the magnetic mineral content of local rocks. This vector has both an orientation and an amplitude. Slight changes in any direction influence the orientation and amplitude. Introduce a highly magnetic rock formation into an otherwise homogenous host, and the local magnetic vector will change. In a three-dimensional world, there are nine (3×3) spatial gradients forming a tensor which defines the anomalous field.

On the conceptual design of gradiometers

A gradiometer is a special measuring instrument that consists of more than one (usually two, rarely three or four) magnetometers. The differences between the readings of the sensors are seen as an approximation to the gradient of the magnetic field along their alignment. The (usually) two magnetometers of a gradiometer are arranged either vertically or horizontally. The distance of the two sensors can vary between several decimeters (for archaeological or environmental mapping) and several kilometers (for research of the deeper-lying geological structures of the Earth's crust, see *Magnetic Methods, Airborne*) depending on the field of application.

Magnetometers

Quantum magnetometers (see also *Magnetometers*) and saturable-core magnetometers are generally suitable for use in gradiometers. Alongside the fluxgate saturable-core magnetometers, cesium vapor magnetometers have become a standard in many areas (aeromagnetic survey, marine geophysics, environmental geophysics, and archaeological prospection) due to the improved noise signal levels against the Overhauser magnetometers or the helium-cooled sensors (Nabighan and Asten, 2002).

Fluxgate gradiometers measure a single component of the Earth's magnetic field (usually the vertical component), while gradiometers consisting of quantum magnetometers (i.e., proton precession and cesium vapor sensors) measure the component of an anomaly in the direction of the Earth's ambient magnetic field, as this is much larger than the anomaly itself (Blakely, 1996). Additionally, fluxgate magnetometers are directionally sensitive and subject to a temperature-related drift, and therefore often have to be recalibrated during measurements.

Instead of using a pair of different magnetic field sensors and differentiating their outputs in order to derive a magnetic gradient value, a single sensitive element can be used for measuring directly a magnetic gradient. This is done by the use of a stiff metallic string clamped at both ends and pumped with an AC current, the frequency of which is tuned to the second eigenmode of the string. The string is excited at that eigenmode in the presence of a quasi-static magnetic gradient. Then, the corresponding mechanical displacements of the string can be measured by the use of an inductive technique with an instrumental noise envelope of less than $\pm 10^{-12}$ m, per 1 s measurement interval (Veryaskin, 2001).

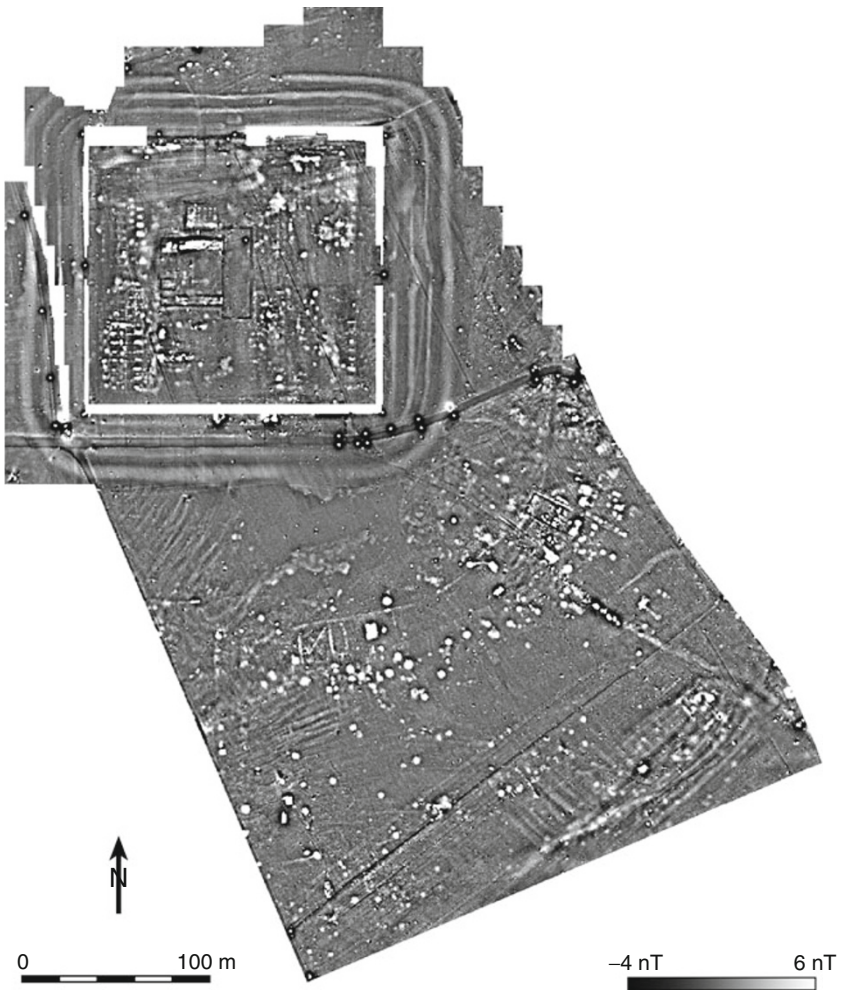
Advantages and limitations

In measurements with a gradiometer, the measured signal amplitude of the magnetic anomaly of an interfering body decreases by the power of three or four depending on the distance between the two sensors. A gradiometer is therefore somewhat less sensitive to deeper-lying features than a total-field magnetometer, where the signal amplitude of an anomaly decreases by the power of three depending on the distance of the sensor to the interfering body. However, gradiometers react very sensitively to interfering features close to the surface and possess a higher spatial resolution in small-scale structures than total-field

Magnetic Gradiometry, Figure 1 The four-channel fluxgate gradiometer of the Institute Dr. Foerster, Reutlingen, Germany. Measurements in a grid of 0.05 by 0.06 m are possible with this device. The two fluxgate sensors are 0.65 m apart.

magnetometers. The variability in time of the Earth's magnetic field and other sources of interference that are far enough away from the gradiometer, and whose influence on the sensors is equal or at least of a similar level during gradiometer measurements, can be eliminated – or at least significantly reduced – by subtracting the simultaneous readings of the two sensors from one another. An advantage of gradient measurements, arising from the ambiguity present in all potential field measurements, is the suppression of broad regional changes in the magnitude of the magnetic field – the long wavelength component. With gradiometer measurements, the local variations are enhanced, making small and weakly magnetic targets recognizable. Gradiometer surveys are particularly useful in areas that are geologically complex. Due to the resulting difference, the magnitude of the anomalies is, however, smaller than the amplitude that one would achieve with a total-field magnetometer. This means that a total-field magnetometer can detect anomalies of smaller and deeper-lying features which cannot be detected by a gradiometer. Fluxgate gradiometers are very sensitive to their position in relation to the magnetic vector field: an ever so slight change in the inclination of the instrument during measurements can lead to incorrect results (Schmidt and Clark, 2000).

The use of vertical gradiometers for magnetic mapping effectively removes the time variations in the magnetic field map without having to establish base stations, networks, or tie lines and is the only insurance against elimination of the high-frequency component that contains the small or weakly magnetic targets. Gradiometers have shown that they can offer a high degree of immunity from



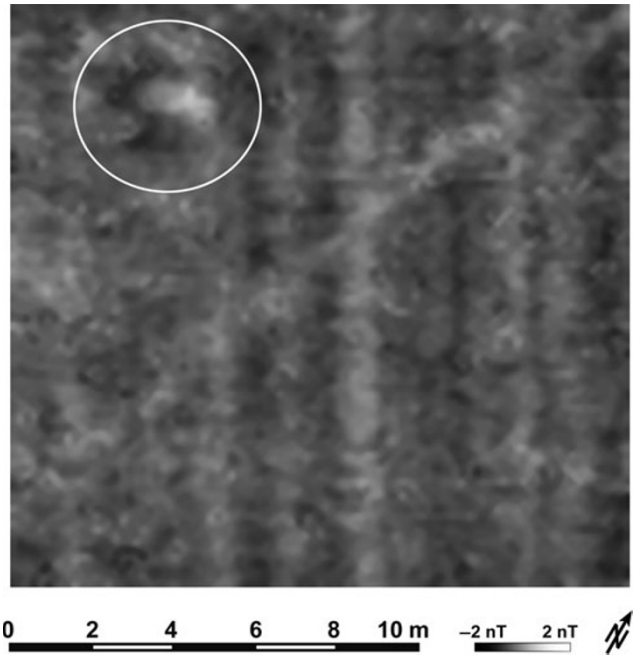
Magnetic Gradiometry, Figure 2 Archaeological survey: Magnetic gradiometer map of the Roman fort in Rainau-Buch, Baden-Württemberg, Germany. Four defensive trenches surround a fort measuring 140 by 150 m. In the center one can recognize the foundations of a stone building (principia), as well as several wooden barracks with their hearths. South of the fort is the adjacent settlement. Magnetic anomalies show the position of streets, stone buildings, cellars, and pits.

diurnal and minor magnetic storm activity in the ambient magnetic field; they can enhance near-surface, small, or weakly magnetic anomalies.

Areas of application

Vertical axial gradiometers are used for measuring the vertical differences (the vertical gradient) in the vertical component of the Earth's magnetic field. This configuration is commonly used for ground-based geophysical surveys in the field of environmental geophysics, in areas with complex geology and archaeological prospecting, as the influences of deeper-lying interferences can be suppressed by this setup of the sensors. For practical reasons, the distance between the sensors of a gradiometer used for ground-based surveys lies somewhere between 0.5 and 1 m. As a rule, the ground penetration and detection ranges to about one to two meters below the surface. Deeper detection can be achieved if the interfering body is accordingly large and strongly magnetized. Fluxgate gradiometers are commonly found in environmental geophysics and in archaeological prospecting (Figure 1). Surveys can be conducted at faster rates, and the highest requirements for spatial resolution of the smallest archaeologically relevant features can be met with measurement grids of 0.05 by 0.06 m without the fluxgate sensors influencing each other. Sometimes, these very small measuring intervals are necessary in archaeological prospecting in order to achieve the best possible spatial resolution of archaeological features (von der Osten-Woldenburg et al., 2006). The magnetic gradiometer map of a Roman fort and its adjacent settlement (Figure 2) that was obtained with a measurement grid of 0.25 by 0.25 m, shows a large number of diverse, often very characteristic anomalies which allow conclusions to be drawn on the position, shape, and size of defensive trenches, centrally located staff buildings with stone foundations, wooden barracks with their hearths, as well as buildings (stone-foundation buildings, street layouts, cellars, and pits) of the settlement adjacent to the fort.

Magnetic gradiometry is also employed in forensic geophysics, even if the radar method has been established as the standard procedure (Cheetham, 2005). However, certain general conditions such as a sufficient distance to the iron cores of reinforced concrete, fences, and other metal constructions must be provided, and it must be ensured that there is no metal refuse in the ground itself. The procedure is identical to that of archaeological prospecting: research is undertaken in small-scale, near-surface survey as the burials themselves are usually close to the surface. The human body only has a small measure of magnetic susceptibility, so a direct detection of the remains by magnetic gradiometry is not possible. But the excavation of a grave and the disruption of strata leads to disturbances of the layers of soil close to the surface that are usually magnetically homogenous, and the refilling of the grave with the excavated material – in which the magnetic particles are now statistically aligned – usually generates weak geomagnetic anomalies, often smaller than 2–3 nT (Figure 3). The displacement of soil with a significantly higher susceptibility by the human body can intensify this effect.



Magnetic Gradiometry, Figure 3 Forensic geophysics: Detection of a burial at the edge of a field (marked by a *circle*). The linear anomalies are produced by drainage pipes and geological structures in the substratum.

Axial gradiometers are also used in aeromagnetic surveys (see *Magnetic Methods, Airborne*), i.e., by helicopter for the detection of ordnance remains close to the surface (Doll et al., 2006), in the prospection of mineral deposits (see *Magnetic Methods, Surface*) and for research of anomalies in the Earth's magnetic field and deeper-lying geological structures of the Earth's crust (see *Magnetovariation Studies*) from stratospheric balloons with a base length of sensors as far as 6 km apart (Webers et al., 2009).

Transverse and longitudinal gradient magnetic measurements are applied in aeromagnetic surveys (Hogg, 2004), in marine geophysics, and space exploration (see *Magnetic Methods, Satellite*). The distance of the magnetometers from one to the other can range from 1.5 m (e.g., in cable and pipeline survey, or for detecting smaller metallic objects) to more than 500 m (for the detection of magnetic sediments close to the surface, shipwrecks, and the prospecting of mineral deposits), depending on the required area of application.

Summary

Magnetics belong to the most commonly used methods for prospecting in applied geophysics when attempting to

infer the position, geometry, and physical parameters of a body of interference or geological structure through the measurement of magnetic anomalies (Schmidt and Clark, 2006).

Gradiometers with two vertically arranged magnetometers that are positioned between 0.5 and 1 m apart, and which measure the vertical gradient and vertical component of the Earth's magnetic field have distinct advantages over total-field magnetometers, especially in ground-based magnetic measurements in environmental geophysics (surveying of contaminated sites, examination of building construction sites, archaeological research of former settlements): they offer a high resolution in smaller features which is, however, gained at the cost of a lesser penetration of the ground in comparison to measurements made with total-field magnetometers. Additionally, they are very sensitive to close-by sources of interference. Background interference (such as the variability in time of the Earth's magnetic field and other nearby sources of interference) can be eliminated or significantly reduced by subtracting the simultaneous readings of the two sensors in the gradiometer arrangement.

Bibliography

Blakely, R. J., 1996. *Potential Theory in Gravity and Magnetic Applications*. Cambridge: Cambridge University Press.

Cheetham, P., 2005. Forensic geophysical survey. In Hunter, J., and Cox, M. (eds.), *Forensic Archaeology: Advances in Theory and Practice*. London/New York: Routledge, pp. 62–92.

Doll, W. E., Gamey, T. J., Beard, L. P., and Bell, D. T., 2006. Airborne vertical magnetic gradient for near-surface applications. *The Leading Edge*, **25**, 50–53.

Hogg, S., 2004. Practicalities, pitfalls and new developments in airborne magnetic gradiometry. *First Break*, **22**, 59–66.

Nabighan, M. N., and Asten, M. W., 2002. Metalliferous mining geophysics – sate of the art in the last decade of the 20th century and the beginning of the new millennium. *Geophysics*, **67**, 964–978.

Schmidt, P. W., and Clark, D. A., 2000. Advantages of measuring the magnetic gradient tensor. *Preview*, **85**, 26–30.

Schmidt, P. W., and Clark, D. A., 2006. The magnetic gradient tensor: its properties and uses in source characterization. *The Leading Edge*, **25**, 75–78.

Veryaskin, A. V., 2001. Magnetic gradiometry: a new method for magnetic gradient measurements. *Sensors and Actuators A: Physical*, **91**(1–2), 233–235.

von der Osten-Woldenburg, H., Chaume, B., and Reinhard, W., 2006. New archaeological discoveries through magnetic gradiometry: the early Celtic settlement on Mont Lassois, France. *The Leading Edge*, **25**, 46–48.

Webers, W., Tsvetkov, Yu., Brekhov, O., Kraoivny, A., Nikolaev, A., Filippov, S., and Pchelkin, A., 2009. Complex satellite ("Swarm") and stratospheric balloons geomagnetic researches. In: *ESA's Second Swarm International Science Meeting*, Potsdam, 2009.

Cross-references

Magnetic Methods, Airborne
Magnetic Methods, Satellite
Magnetic Methods, Surface
Magnetometers
Magnetovariation Studies

MAGNETIC METHODS, AIRBORNE

Mike Dentith
School of Earth and Environment (M004), The University of Western Australia, Crawley, WA, Australia

Synonyms

Aeromagnetics

Introduction

Making measurements from the air of the Earth's magnetic field is a well-established geophysical survey method. The normal practice is to measure the scalar amplitude of the magnetic field in the survey area, this being referred to as the total magnetic intensity (TMI), the name reflecting the fact that the measurements represent the resultant of all the magnetic fields, of whatever source, in the vicinity of the measurement. The major contributions come from the geomagnetic field and the fields due to magnetic rocks in the upper crust, the latter being the fields of interest.

Initially used to infer the depth to magnetic rocks beneath nonmagnetic cover in sedimentary basin studies and for identification of major lineaments, the products from aeromagnetic surveys have evolved to now be a reliable means of creating detailed maps of the near-surface geology. Current developments in aeromagnetics are mostly related to the increasingly sophisticated inference of the 3D distribution of magnetism in the Earth's crust using inverse modeling methods.

Comprehensive accounts of aeromagnetic data, acquisition, processing, and interpretation are proved by Gunn (1997) and Reeves (2005).

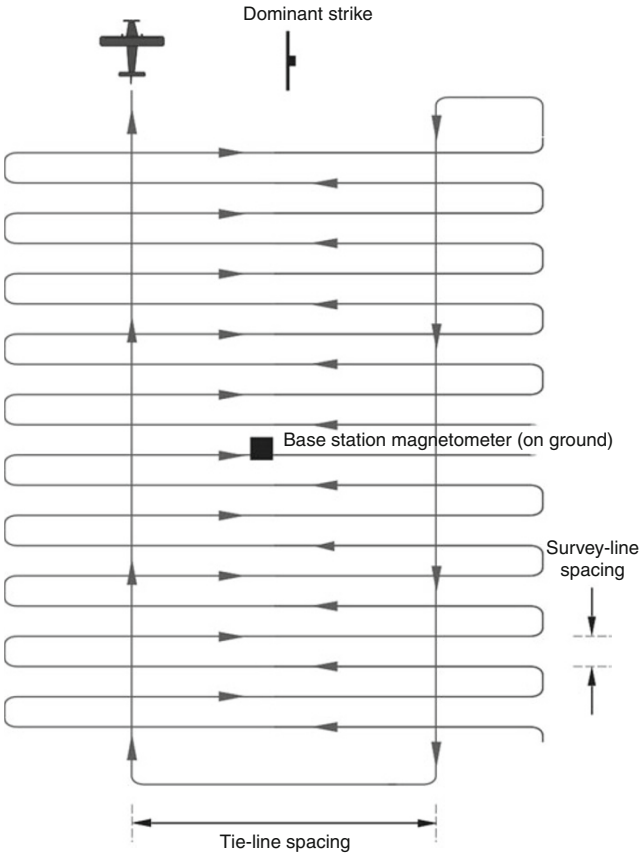
Making magnetic measurements from the air

Airborne magnetic data are acquired as a series of parallel "survey" or "flight" lines and more widely spaced orthogonal "tie" lines (Figure 1). The purpose of the tie lines is to create repeat measurements at the locations where they cross the flight lines, which are required during processing of the data (see below).

An advantage of making magnetic measurements from the air is that unless the survey area is small, typically less than a few square kilometers, it is much faster and cheaper to acquire the data than for a ground-based survey. The main disadvantage is that variations in topography, especially if the terrain is rugged, may seriously affect the ability to position the aircraft as would be preferred and also affect the individual measurements in a hard to predict way.

Equipment

Airborne measurements are routinely made using fixed-wing aircraft and helicopters. Surveys using fixed-wing aircraft are about a factor of four times cheaper than helicopter surveys, but helicopters have the advantage of



Magnetic Methods, Airborne, Figure 1 Schematic illustration of the flying pattern used in aeromagnetic surveys.

being able to fly lower and better maintain a consistent ground clearance in rugged terrain, but have the disadvantage of being less stable survey platforms.

To reduce effects of magnetic fields originating in the aircraft the magnetic sensor, usually a cesium vapor magnetometer (see *Magnetometers*), is located as far as possible from the main body of the aircraft and hence the magnetic fields associated with ferrous components and electric currents. For fixed-wing aircraft the magnetic sensor is mounted in a "stinger" located aft of the aircraft. In helicopters the sensor may be in a "boom" that positions the magnetometer in front of the aircraft or it may be within a towed "bird." The bird is a bomb-like device suspended on a cable beneath the aircraft. A "compensation" system within the aircraft is used to further reduce the effects of magnetic fields originating from the aircraft.

In addition to carrying one or more magnetometers, the survey aircraft also measures terrain clearance using an altimeter and GPS position and time. One or more "base station" magnetometers are also deployed on the ground to measure temporal changes in the magnetic field in the survey area to enable such variations to be removed from the data (see below).

The importance of source-sensor separation

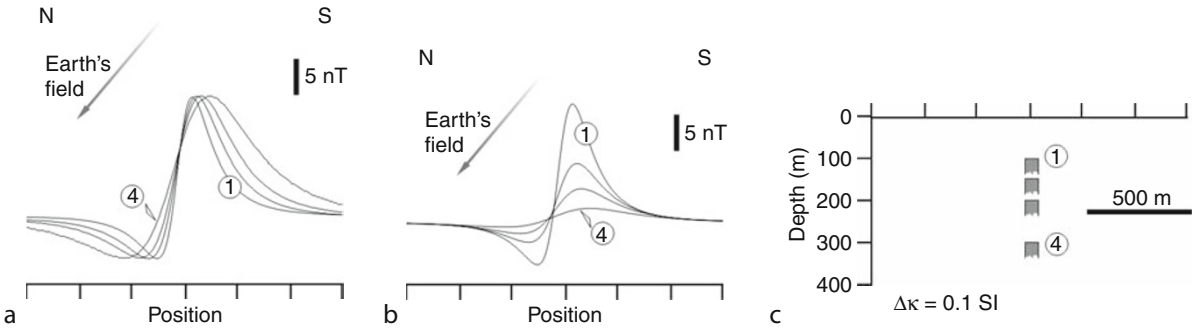
An important source of noise in aeromagnetic data is associated with changes in distance between the magnetic sensor and the magnetic materials. Any changes in this distance will create variations in TMI, which could easily be confused with those caused by changes in the local geology. As the distance increases the measured magnetic field variation decreases in amplitude and increases in wavelength (Figure 2b). The wavelength increase is most easily seen when the amplitude decrease is removed by rescaling the individual profiles so they have the same peak-to-peak amplitude (Figure 2a).

Ideally the survey aircraft maintains a constant ground clearance during the acquisition of the data, referred to as "drape" flying. In practice, aircraft are subject to the effects of wind and turbulence and the inaccuracies of human control. In addition, all aircraft have a limited ability to maneuver, especially climb. Figure 3 shows the flight path of a fixed-wing aircraft across a ridge and adjacent areas. When short wavelength variations in topography (B on the diagram) are encountered, the aircraft cannot respond quickly enough to maintain the separation and positive relief will bring the magnetic sources closer to the sensor, with a corresponding increase in the amplitude of the magnetic field (c.f. Figure 2), and vice versa. When large topographic features are encountered (A on the diagram) the aircraft will inevitably vary its ground clearance because of its limited ability to climb and to a lesser extent descend. If, as is the normal case, alternate flight lines are flown in opposite directions, adjacent parts of survey lines will have markedly different ground clearance adjacent to topographic features. This will introduce spurious variations into the data.

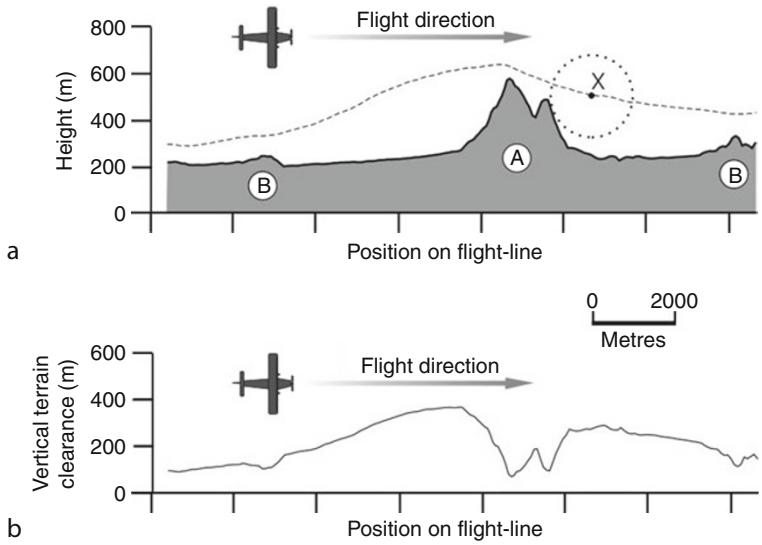
Current practice is to "program" a preplanned flight path for the aircraft, which the pilot follows using a real-time display. This program accounts for the terrain and the aircraft's performance. The aim is to achieve the best "loose drape," whereby terrain clearance effects are minimized as far as is practical. Note that even if a perfect drape was achieved, the effects of terrain would not be totally removed because even in the ideal scenario there are still the effects of variations in the distances to magnetic materials because it is not just the materials directly below the aircraft that are influencing the measurements. The measurement at X in Figure 3 will be affected more by the magnetic materials to the side than beneath because of the lesser distance to these materials, although this scenario only occurs in severe terrain.

Survey design

Sampling theory requires individual measurements of TMI to be spaced at a maximum of half the wavelength of the shortest wavelength of variation. Economic and safety factors mean this is rarely practical and some aliasing of responses is the norm. This is not necessarily a major problem since during qualitative interpretation of the data it is relative changes in amplitude and texture



Magnetic Methods, Airborne, Figure 2 Profile of total magnetic intensity across a dyke-like magnetic body. The four profiles correspond with the bodies at different depths in (**c**). The profiles in (**b**) represent variations in TMI. In part (**a**) the amplitudes have been normalized to the profile with maximum amplitude (shallowest source).



Magnetic Methods, Airborne, Figure 3 Example of the flight path of a fixed-wing survey aircraft across rugged topography. (**a**) Topographic and altimetry profile, (**b**) vertical terrain clearance. (Based on data in Flis and Cowan, 2000).

that are used, and in fact accurate definition of the variations is only required when specific anomalies need to be analyzed quantitatively (see *Magnetic Anomalies, Interpretation*) and often there is a follow-up, more detailed, survey of the area of interest to improve anomaly characterization.

Survey specifications

The survey geometry, illustrated in Figure 1, requires definition of three key parameters: the survey line spacing and orientation and the flight height. Typical values of two of these variables for different survey types are given in Table 1. Note that the tie-line spacing is a dependent variable, typically being set to ten times the survey line spacing, although this may be reduced to five times or less for high resolution surveys.

The survey line spacing controls the cost of the survey, which for fixed-wing aircraft is based on the total length of lines flown. The total line length of a survey in terms of the survey area and line spacing can be estimated from the equation below (Brodie, 2002).

$$\text{Total line length} = \frac{1000\left(1+\frac{\Delta_{\text{Survey lines}}}{\Delta_{\text{Tie lines}}}\right).\text{Survey area}}{\Delta_{\text{Survey lines}}}$$

Where $\Delta_{\text{Survey lines}}$ and $\Delta_{\text{Tie lines}}$ are the survey- and tie-line spacings in meters, respectively, the total line length is in kilometers, and the survey area is in square kilometers.

To this cost must be added nonproduction costs such as mobilization and "stand by" costs associated with factors outside the acquisition company's control, for example, bad weather, magnetic storms. With helicopter surveys the time spent in the air is also taken into account, and

Magnetic Methods, Airborne, Table 1 Typical survey specifications for aeromagnetic surveys

Survey line spacing (m)	Tie-line spacing (m)	Flight height (m)	Sampling ratio: Survey/Tie line[a]	Shortest along-strike wavelength sufficiently sampled
50	500	40	7.5	100
100	1,000	50	15	200
200	2,000	60	30	400
400	4,000	80	60	800

[a]assumed flight speed is 240 km/h (67 m/s) and 10 Hz sampling resulting in a 6.7 m along survey line sampling interval. These parameters are typical for a survey with a fixed-wing aircraft.

may be significant in mountainous areas where weather conditions severely restrict data acquisition both in terms of flying time and location (Mudge, 1996).

It is clear from Figure 2 that the lower the flight height the greater will be the amplitude of the targeted responses and also the shorter the wavelength of variation, and hence the greater the ability to resolve adjacent features. Figure 2 implies the lowest possible flight height should be used, but safety considerations mean flight heights are some tens of meters. "Crop-duster" fixed-wing aircraft or heliborne surveys may fly lower but ultra-low flying is not necessarily an advantage from a technical point of view since it means maintaining consistent terrain clearance is more difficult. Besides, there may be very shallow magnetic sources whose anomalies mask those of deeper, more significant variations (Doyle and Lindeman, 1985).

Reid (1980) showed that accurate definition of a TMI anomaly requires the sample spacing to be a maximum of half the source-sensor separation. The combination of the rate at which the aircraft's magnetometers make a measurement (typically 0.1 s) and the velocity of the aircraft (typically 220–280 km/h for a fixed-wing aircraft) result in an along flight line sample spacing of about 7 m for fixed-wing aircraft. Helicopter surveys with a 0.1 s sampling typically have a 4–5 m spacing, but the trend is to employ faster sampling magnetometers, reducing this spacing to even less. The important outcome from the above is that the along-line sampling exceeds the requirements of sampling theory for practical flight heights and the variations in TMI in the direction of the flight path can be considered as more than adequately characterized. The same is not true for variations in the directions perpendicular to the survey line direction because, as shown in Table 1, the spacing of these lines is very much greater than the along-line sampling interval. Note that since anomaly wavelength decreases with source-sensor separation, a low flight height combined with wide survey lines will lead to greater aliasing in the direction perpendicular to the survey lines. This is reflected in the parallel increase in flight height with line spacing in Table 1.

Conventionally airborne geophysical surveys are flown with the flight lines oriented perpendicular to the dominant geological strike direction. The rationale for this is that there will be shorter wavelength variations across strike than parallel to strike, and hence the greatest sampling density should be in the strike-perpendicular direction. For elongate anomalies the short wavelength variation parallel to strike is small, so significant aliasing does not occur. In practice economic considerations dictating the survey line spacing mean, there is inevitably some aliasing of TMI variations in the direction perpendicular to the survey lines.

Even if there is some variation in the strike across the survey area, and there is also a conflict caused by variations in stratigraphic and structural strike, it is best to use a single-line orientation. If this is not the case the data are very difficult to process into an easily interpreted form.

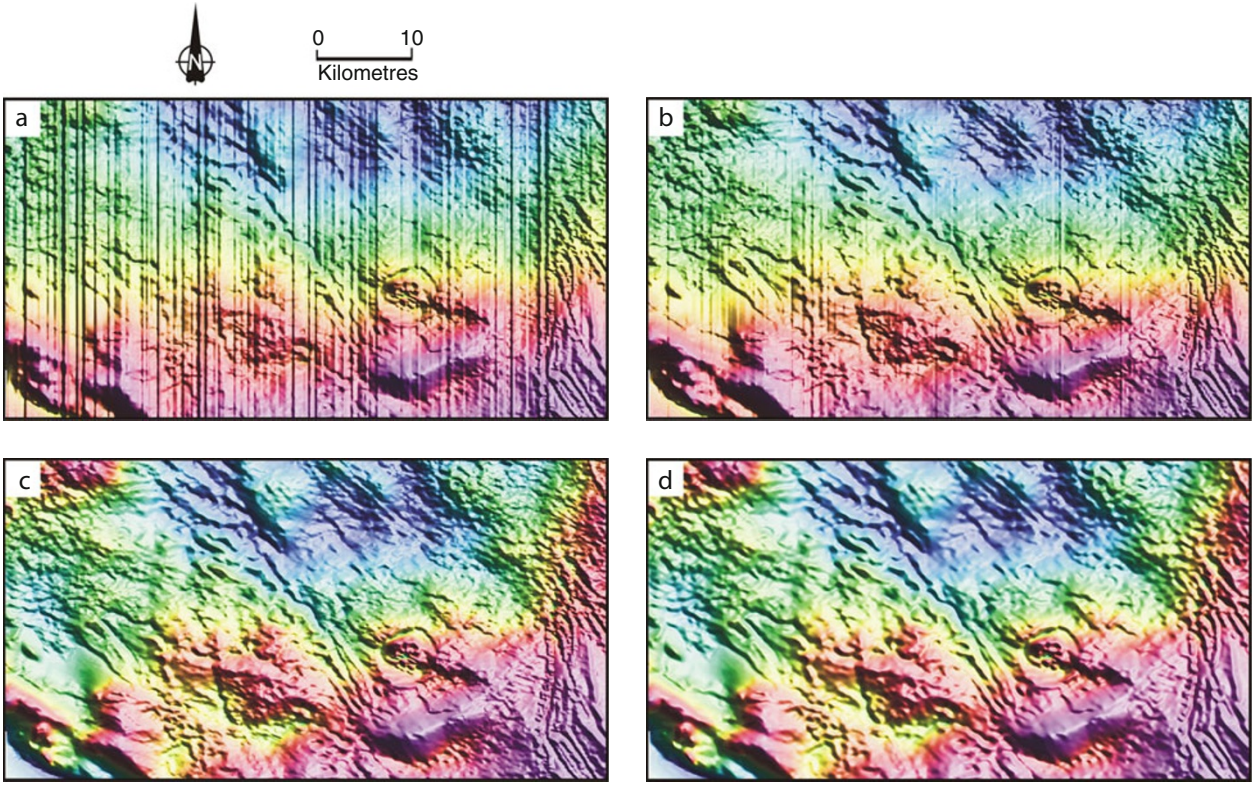
For surveys in equatorial regions the east–west elongation of magnetic responses associated with low declinations ($<30^\circ$) means it is always preferable to fly surveys with north–south survey lines.

Data processing

The term data processing refers here to the removal of noise prior to enhancement (see *Magnetic Data Enhancements and Depth Estimation*) and interpretation of the data (see *Magnetic Anomalies, Interpretation*). The aim of the processing is to create an instantaneous "snap shot" of TMI variations across the survey area with the contribution from the geomagnetic field removed. It is normal practice to complete at least preliminary processing in the field to ensure that the data are of acceptable quality.

The main source of noise in aeromagnetic data is temporal changes in the Earth's magnetic field, which occur during the course of the survey and hence from the mobile sensor's perspective are indistinguishable from lateral variations in TMI due to the geology of the survey area. It is to determine such variations that the repeat measurements inherent in the survey/tie-line geometry are required (Figure 1), along with the TMI time series recorded by each base station.

The removal of the noise is achieved in a series of steps. Figure 4 shows a raster display of an aeromagnetic dataset after each of the processing steps described below. Figure 4a shows the data after compensation for aircraft-related effects. Note the dominant characteristic is a north–south striping (corrugations) with a wavelength approximately equal to twice the line spacing; this being primarily the result of the temporal changes in the geomagnetic field during the survey and deviations of the aircraft from the ideal flight path.



Magnetic Methods, Airborne, Figure 4 Gridded TMI data from an aeromagnetic survey in Australia. The colors represent changes in amplitude (hot colors high, cold colors low) and the intensity represents a shaded relief presentation designed to highlight north–south trending features. (**a**) Data after compensation for magnetic fields originating in the aircraft. (**b**) As for (**a**) but with correction for temporal variations in the geomagnetic field as measured by a base station magnetometer. (**c**) As for (**b**) but with international geomagnetic reference field removed and tie-line leveling corrections applied. (**d**) As for (**c**) except for microleveling. Images created by Cowan Geodata Services.

Correction for temporal changes in the geomagnetic field

The Earth's magnetic field varies in intensity at a range of time scales; see *Geomagnetic Field, Theory*. Of significance to aeromagnetic surveys are magnetic storms (during which surveying must be discontinued) and micropulsation and diurnal variations. Micropulsations are variations in the geomagnetic field that last from a few seconds to a few minutes. They change in both phase and amplitude over distances comparable to survey dimensions. Diurnal variations have periods of 24 h and magnitudes of up to about 50 nT.

The principal means of removing micropulsation and diurnal variations is the base station magnetometer readings (the base station readings also allow the recognition of the onset of a magnetic storm). The readings in the aircraft and base station are time synchronized. In theory the variation at the (stationary) base station should represent geomagnetic field variations that occur during the course of the survey, and subtracting the base station readings from the time equivalent survey reading should leave a residual dataset free of the temporal variations. However, this is an imperfect process since the amplitude and phases of the changes vary with location. To counter this multiple base stations may be used for large surveys. The subtraction causes a base line shift in the data but this is unimportant since it is relative variations in TMI that are required for interpretation of the data (see *Magnetic Anomalies, Interpretation*). Often an average intensity of the magnetic field in the survey area is added to each measurement to return the data to closer to likely absolute values. Figure 4b shows the magnetic dataset after corrections for temporal changes in the geomagnetic field derived from a base station dataset. The degree of striping in the data is significantly reduced, although it is not eliminated.

Removal of geomagnetic field

It is normal practice to remove spatial variations in TMI caused by the geomagnetic field as defined by the International Geomagnetic Reference Field (IGRF), see *Geomagnetic Field, IGRF*. The wavelengths of variation of the field are very large compared to survey dimensions, and subtraction from the observed data will not usually significantly affect variations of interest. Nevertheless it is standard practice to remove such variations and it is computationally trivial to do so.

Tie-line leveling

Ideally the processing described above will mean the TMI measured along the survey- and tie-lines at their cross-over points will be identical. In practice, there will always be mismatches because of the imperfect aircraft compensation, spatial variations in diurnal and micropulsation related changes, but also because there are uncertainties in the aircraft's position, that is, where the cross-over occurs within the datasets. Differences in elevation between the survey- and tie-line measurements are the main source of error.

The minimization of the differences at the cross-over points is referred to as "leveling." Various approaches are used but normally there is an initial estimate of the extent to which location errors are responsible for the differences. Once these are accounted for the remaining differences at each line intersections are used to adjust the amplitudes along the survey lines. The most common method involves fitting a smooth function to the differences and using this to estimate leveling "corrections" at each measurement point along the survey lines. Figure 4c shows the magnetic data after removal of the IGRF (minimal effect) and tie-line based leveling. Corrugations in the data are now all but invisible.

Microleveling

Microleveling is a loosely defined term for making adjustments to the data when the processing stream outlined above is complete and artifacts appear to remain in the data. The details of these methods are mostly proprietary but are all various forms of low-pass filtering. The filter characteristics are often defined by the flight and survey line spacing, for example, Minty (1991). Microleveling is usually applied after the data have been gridded (see below), and the correction so derived used to adjust the individual survey measurements prior to re-gridding. Figure 4d shows the example magnetic data after application of a commercial microleveling process. There is some improvement in the lateral continuity of anomalies.

Gridding

By far the most common form of display and input to enhancement oriented processing operations are data projected on to a regular grid. Gridding the data is a compromise between honoring detailed sampling along the survey lines and producing smooth interpolation into areas between lines where no readings have been made. The grid cell size, that is, distance between interpolated locations, is conventionally between 1/3 and 1/5 of the survey line spacing. The loss of information in the survey line direction is balanced by an acceptable degree of reliability of interpolated points located between the survey lines. Nevertheless, especially with greater survey line spacing, aliasing frequently causes laterally continuous features to appear as a series of adjacent sub-circular features, referred to as "beads." This effect becomes worse as the trend of the anomalies approaches that of the survey lines.

Discussion of data processing

The processing described above is a combination of processes with a basis in field measurements and processes requiring the judgment of an expert processor, because it is not practical to collect the necessary data to process the data perfectly. There is no way to reliably judge the accuracy of the final measurements, but an accuracy of ± 1 nT is often quoted for individual points in the gridded datasets.

Summary

Mapping variations in TMI allows variations in rock magnetism in the upper crust to be mapped, from which geology may be inferred. The main sources of noise in magnetic data are temporal changes in the geomagnetic field and variations in the distance between magnetic sources and the magnetometer caused by changes in aircraft altitude. Temporal changes are accounted for by comparing survey data with readings contemporaneously recorded by a magnetometer on the ground. Discrepancies in repeat readings at the same location allow any remaining temporal change noise to be removed along with corrections for variations in survey altitude. The resulting data are projected on to a regular grid for display as raster images.

Bibliography

Brodie, R. C., 2002. Airborne and ground magnetics. In Papp, E., (ed.), *Geophysical and Remote Sensing for Regolith Exploration*, CRCLEME Open file report 144, 33–45.

Doyle, H. A., and Lindeman, F. W., 1985. The effect of deep weathering on geophysical exploration in Australia – a review. *Journal of the Geological Society of Australia*, **32**, 125–135.

Flis, M. F., and Cowan, D. R., 2000. Aeromagnetic drape corrections applied to the Turner Syncline, Hamersley Basin. *Exploration Geophysics*, **31**, 84–88.

Gunn, P., 1997. Airborne magnetic and radiometric surveys. *AGSO Journal of Australian Geology and Geophysics*, **17**(2), 216.

Minty, B. R. S., 1991. Simple micro-leveling for aeromagnetic data. *Exploration Geophysics*, **22**, 591–592.

Mudge, S., 1996. Helicopter magnetic surveys in rugged terrains: from planning to processing. *The Leading Edge*, 305–308.

Reeves, C., 2005. Aeromagnetic surveys: principles, practice and interpretation. *Geosoft*, 155.

Reid, A. B., 1980. Aeromagnetic survey design. *Geophysics*, **45**, 973–976.

Cross-references

Geomagnetic Field, IGRF
Geomagnetic Field, Theory
Magnetic Anomalies, Interpretation
Magnetic Data Enhancements and Depth Estimation
Magnetic Methods, Principles
Magnetometers

MAGNETIC METHODS, PRINCIPLES

Kusumita Arora
National Geophysical Research Institute, Council of Scientific & Industrial Research (CSIR), Hyderabad, India

Definition

The term "Magnetic Methods" in geophysics generally refer to the collection of techniques, which aim to interpret the structure of the Earth's crust based on the physical property of magnetism.

Most commonly this term refers to those applications by means of which investigations are carried out regarding the subsurface structure and processes, guided by the presence of magnetic material in the Earth's crust.

Introduction: nature of magnetism

This article attempts a simple language description of how the material property of magnetism is exploited in geophysical terms to derive information regarding various aspects of solid Earth structure and processes.

Intuitively, magnetism makes us think of an iron bar, which aligns itself in the north–south direction. A magnetized bar has its power concentrated at two ends, its poles; they are known as its north (N) and south (S) poles, because if the bar is hung by its middle from a string, its N end tends to point northwards and its S end southwards. The N end will repel the N end of another magnet, S will repel S, but N and S attract each other. The region where this is observed is called the magnetic field. Either pole can also attract iron objects such as pins and paper clips because under the influence of a nearby magnet, each pin or paper clip becomes itself a temporary magnet, with its poles arranged in a way appropriate to magnetic attraction. If a much stronger/bigger bar magnet is placed near the suspended one, the latter will get aligned in a direction oriented towards the former, deviating from the original N–S direction.

Magnetic attraction occurs even when iron magnets are not evident. The Earth itself behaves like a giant weak magnet, which exerts its influence on a very wide range of magnetic materials. Out in space there is no magnetic iron, yet magnetism is widespread. For instance, sunspots consist of glowing hot gas, which are all intensely magnetic. Study of magnetism over the centuries by scientists like William Gilbert, Charles Coulomb, Hans Christian Oerstad, Andre–Marie Ampere, Michael Faraday, Heinrich Lorentz, and James Clerk Maxwell has led to the knowledge that electricity and magnetism are different manifestations of the same fundamental force field and are related to each other at microscopic and macroscopic scales.

Magnetism is caused by the flow of electric current in a conducting material. In space, on the Sun and in the Earth's core, electric currents are the only source of magnetism. Within magnetic materials, bulk magnetism is the result of the motion of electrons in an atom, which sets up a magnetic field around itself.

Components of the Earth's magnetic field

The Earth's magnetic field probably originated about 3.45 billion years ago and played a vital role in protecting the planet from damaging solar radiation, which may have stripped it of its atmosphere and water, making it completely unsuitable for life as we know it (Tarduno et al., 2010). Though the exact cause of the magnetic force continues to elude researchers, it is well established that electric currents, be on microscopic or macroscopic scales are linked to the existence of a magnetic field. The Earth's magnetic field, which is the strongest one among the rocky planets, extend 36,000 miles into space.

In solid earth geophysics, we deal with the following components:

1. The weak all-pervading geomagnetic field, generated by complex processes in the Earth's core, having a dipole and a non-dipole component (mainly represented by secular variations).
2. The smaller but often locally stronger magnetic fields generated by magnetic materials present in the crust and upper mantle of the Earth.
3. Motional induction in the ocean waters.
4. The magnetic field of the Sun, very strong but far away from the Earth, acting on the magnetospheric field.
5. Ionospheric fields, which are influenced by Earth's main field and also magnetospheric effects.

All these magnetic fields interact with each other and what we can sense and measure is a summation of all of their influences. The Geomagnetic field has a low amplitude but a very large radius of influence from the Earth's core to the magnetosphere. Near the surface of the Earth, it is influenced locally by the presence of magnetic material. Beyond the Earth and its atmosphere, the magnetosphere and ionosphere are strongly influenced by the interaction of particles and radiation coming from the Sun.

"Magnetic Methods" are implemented with an intention to remove the effects of the ambient Geomagnetic and solar magnetic fields from the measurements of magnetic force on the Earth's surface, isolate and study the local magnetic fields and use them to trace the sources of crustal magnetic material. This leads to the knowledge of the internal structure and processes operating in the Earth's crust as well as concentrations of magnetic ores in the crust.

Magnetic materials/rocks

For this method to be applied, materials found in the Earth's crust and upper mantle need to be categorized by their response to externally applied magnetic fields, in this case the Earth's magnetic field. When a material is influenced by a magnetic field, the magnetic forces of the material's electrons will be affected. This effect is

known as Faraday's Law of Magnetic Induction. Materials can react quite differently to the presence of an external magnetic field, depending on the atomic and molecular structure of the material and the net magnetic field associated with the atoms. The magnetic moments associated with atoms have three origins: electron orbital motion, the change in orbital motion caused by an external magnetic field, and the spin of the electrons. An understanding of the magnetic property requires quantum mechanical analyses. A working understanding can be acquired from the simplified concept of atomic structure given by Ernest Rutherford. In most atoms, electrons occur in pairs, which spin in opposite directions. So, when electrons are paired together, their opposite spins cause their magnetic fields to cancel each other. Therefore, no net magnetic field exists. Alternately, materials with some unpaired electrons will have a net magnetic field and will react more to an external field. Most materials can be classified as diamagnetic, paramagnetic, or ferromagnetic.

The magnetization of a material is expressed in terms of density of net magnetic dipole moments μ in the material. We define a vector quantity called the magnetization M by

$$M = \mu_{\text{total}}/V.$$

Then the total magnetic field B in the material is given by

$$B = B_0 + \mu_0 M$$

where μ_0 is the magnetic permeability of space and B_0 is the externally applied magnetic field. When magnetic fields inside materials are calculated using Ampere's law or the Biot–Savart law, then the μ_0 in those equations is typically replaced by just μ with the definition

$$\mu = K_m \mu_0$$

where K_m is called the relative permeability. If the material does not respond to the external magnetic field by producing any magnetization, then $K_{\text{m}} = 1$. Another commonly used magnetic quantity is the magnetic susceptibility which specifies how much the relative permeability differs from one.

$$\text{Magnetic susceptibility } \chi_m = K_m - 1$$

For paramagnetic and diamagnetic materials, the relative permeability is very close to one and the magnetic susceptibility very close to zero. For ferromagnetic materials, these quantities may be very large.

Another way to deal with the magnetic fields which arise from magnetization of materials is to introduce a quantity called magnetic field strength H. It can be defined by the relationship

$$H = B_0/\mu_0 = B/\mu_0 - M$$

and has the value of unambiguously designating the driving magnetic influence from external currents in a material, independent of the material's magnetic response. The relationship for B above can be written in the equivalent form

$$B = \mu_0(H + M)$$

H and M will have the same units, A/m.

Diamagnetic metals have a very weak and negative susceptibility to magnetic fields. Diamagnetic materials are slightly repelled by a magnetic field and the material does not retain the magnetic properties when the external field is removed. These are solids with all paired electrons resulting in no permanent net magnetic moment per atom. Diamagnetic properties arise from the realignment of the electron orbits under the influence of an external magnetic field. Most elements in the periodic table, including copper, silver, and gold are diamagnetic.

Paramagnetic metals have a small and positive susceptibility to magnetic fields. These materials are slightly attracted by a magnetic field and the material does not retain the magnetic properties when the external field is removed. Paramagnetic properties are due to the presence of some unpaired electrons, and from the realignment of the electron orbits caused by the external magnetic field. Paramagnetic materials include magnesium, molybdenum, lithium, and tantalum.

Ferromagnetic materials have a large and positive susceptibility to an external magnetic field. They exhibit a strong attraction to magnetic fields and are able to retain their magnetic properties after the external field has been removed. Ferromagnetic materials have some unpaired electrons so their atoms have a net magnetic moment. They get their strong magnetic properties due to the presence of magnetic domains. In these domains, large numbers of atom's moments are aligned parallel so that the magnetic force within the domain is strong. When a ferromagnetic material is in the unmagnetized state, the domains are nearly randomly organized and the net magnetic field for the part as a whole is zero. When a magnetizing force is applied, the domains become aligned to produce a strong magnetic field within the part. Iron, nickel, and cobalt are examples of ferromagnetic materials. Components with these materials are commonly inspected using the magnetic method. Ferromagnetic materials will undergo a small mechanical change when magnetic fields are applied, either expanding or contracting slightly. This effect is called magnetostriction.

In oxide crystals, the large oxygen ions keep the metal ions far apart, and exchange of electrons among them can only occur through indirect exchange, resulting in antiparallel directions of adjacent atomic magnetic moments, giving rise to antiferromagnetism which is weak and positive. Ilmenite is such a mineral, which becomes paramagnetic at room temperatures. Parasitic ferromagnetism occurs in minerals having defects in their lattices, like hematite. In spinel crystal structures, the indirect exchange process involves anitiparallel and unequal magnetization

of the sublattices, resulting in ferrimagnetism, which changes to paramagnetism beyond Curie temperature. The most important such minerals are magnetite, pyrrhotite, goethite (Lowrie, 1997).

Thus, an external magnetic field will strongly attract ferromagnetic materials, weakly attract paramagnetic materials, and weakly repel diamagnetic materials. Further details are given in *Magnetic Domains*; *Remanent Magnetism* and *Magnetic Anisotropy*.

Measurement of magnetic field and magnetic anomaly

The Earth's magnetic field varies in strength and direction over time and space as a result of the varying degrees of interactions of the causal fields. Each of the components of the magnetic field, mentioned above, influence and is influenced by the presence of the other components. This makes the study of the Earth's magnetic field interesting and challenging. Sensors are designed to record the total effect of the Earth's magnetic field and also to track the variations of one or more of the components. Many different kinds of magnetometers have been developed over the history of study of magnetism, details may be checked up on in *Magnetometers*. Near-Earth and near-polar satellites measure the magnetic field from space.

The compilation and processing of all such measurements have allowed the magnetic field of the Earth to be mapped successfully (www.ngdc.noaa.gov/IAGA/vmod/igrf.html, http://smsc.cnes.fr/html-images/solid_earth.htm). Details of such techniques are to be found in *Magnetic Methods, Satellite*; *Magnetic Methods, Surface*; *Geomagnetic Field, Measurement Techniques*. The International Association of Geomagnetism and Aeronomy (IAGA), through many participating agencies of magnetic field modelers and the institutes involved in collecting and disseminating magnetic field data from satellites and from observatories and surveys around the world, generates the standard mathematical description of the Earth's main magnetic field, which contain the contributions of all the sources listed above.

A key concept in Magnetic Method is that of magnetic anomaly. An anomaly is a deviation from a normal. Hence, embedded in the word "anomaly" are the concepts of the *ideal* and the *real*, which is different from the ideal. In the case of the magnetic field of the Earth, the ideal field is visualized as that produced by currents within a homogeneous outer core, which is essentially contained in the International Geomagnetic Reference Field (IGRF), see *Geomagnetic Field, IGRF*; *Geomagnetic Field, Theory*; *Geomagnetic Field, Global Pattern*. This Main Field itself changes with time and the IGRF needs to be redefined periodically.

At any given point of time the IGRF is overlayed by the effects of secular variations and geomagnetic jerks, by variations in solar activity and crustal magnetization, which varies from place to place within the magnetic field (see *Geomagnetic Excursions*; *Geomagnetic Field, Secular Variation*). The standard technique of Spherical Harmonic Analyses (see *Spherical Harmonic Analysis Applied to Potential Fields*) enable the Main Field (dipole field, non-dipole field, and secular variations of each) to be quantified and separated from the total observed magnetic field. The higher order terms in the energy density spectrum represent the contribution to the total field by crustal magnetization. It is the aim of "magnetic methods" to isolate and study these contributions, which are generated by composition and structures of the Earth's crust under the influence of the ambient magnetic field.

Magnetic anomaly maps provide insight into the subsurface structure and composition of the Earth's crust. Anomalies trending parallel to the isochrons (lines of equal age) in the oceans reveal the temporal evolution of oceanic crust. Magnetic maps are widely used in the geological sciences and in resource exploration. Furthermore, the global magnetic map is useful in science education to illustrate various aspects of Earth evolution, such as plate tectonics and crustal interaction with the deep mantle. Distinct patterns and magnetic signatures can be attributed to the formation (seafloor spreading) and destruction (subduction zones) of oceanic crust, the formation of continental crust by accretion of various terranes to cratonic areas and large scale volcanism (both on continents and oceans). Anomaly maps depicting crustal magnetization represent the cumulative effects of the Earth's magnetic field from the time of formation of the rocks within the crust, through various stages of metamorphism and finally the present ambient field (see *Remanent Magnetism*). The Main Field of the Earth, generated from the core, changes its strength and direction over millions of years and these influences are superimposed in the magnetization patterns of the minerals in the rocks. Techniques of separating these (see *Remanent Magnetism*; *Paleomagnetism, Measurement Techniques and Instrumentation*) embedded signatures are available and yield information about the Earth's magnetic field of the past and the present.

The total magnetization of a rock is thus the summation of remanent and induced magnetizations, which have different directions and intensities. Hence, the resultant vector is not parallel to the present magnetic field. The Konigsberger ratio Q is defined as the ratio of the intensity of magnetization to that of induced magnetization. Commonly, values of Q are low for magnetic minerals on land because these minerals are dominantly multi-domain in structure, which are reoriented by changes of the ambient magnetic field. Also they frequently have high susceptibility and the Earth's present magnetic field can induce a strong magnetization. On the other hand, ocean basalts, which were formed by extrusion and rapid underwater cooling, have very high Q values because of the structure and size of the grains as a result of the quenching.

The component of induced magnetization can be neglected and interpretations made as if the magnetization is entirely remanent.

The variations of the magnetic field with altitude, latitude and longitude are dominated by the vertical and horizontal variations of the dipole field. The altitude correction is given by the vertical gradient of the magnetic field. The corrections for latitude and longitude are inherent in the reference field that is subtracted, or the latitude correction can be given by the north–south horizontal gradient of the magnetic field. In highly magnetic terrains, the disturbing effect of magnetized topography may require topographic corrections (see *Magnetic Data Enhancements and Depth Estimation*).

Interpretation of magnetic anomalies on land

By analyzing magnetic anomalies, geophysicists are able to learn about geologic structures, which may be buried beneath the Earth's surface. Geologic structures (such as faults or igneous intrusions) often produce small magnetic fields that distort the main magnetic field of the Earth. Volcanic rocks are the most prevalent magnetic lithology, and we expect high-amplitude, short-wavelength anomalies over volcanic terranes. Steeply dipping faults that offset subhorizontal units often produce magnetic anomalies that appear as linear trends on aeromagnetic maps. Latites, on the other hand, often are extruded from volcanic plugs and thus tend to produce intense, somewhat circular magnetic anomalies (often as magnetic lows, because the latites are generally reversely polarized). Metavolcanic rocks, gabbros, and some of the intrusive rocks can produce prominent magnetic anomalies.

The magnetic properties of sedimentary rocks, such as Paleozoic sequences of dolomite, limestone, and sandstone, are usually weak, such that the resulting magnetic anomalies are very small in amplitude and undetectable by airborne surveys. The Precambrian metasedimentary rocks are generally incapable of producing detectable magnetic anomalies, although there are some notable exceptions like the iron-rich metachert and metagraywackes.

Interpretation of magnetic anomalies in the oceans

The special characteristics of magnetisation of the oceanic crust was discovered in the 1950s, when large areas of oceanic crust were found to be characterized by long stripes of alternating positive and negative magnetic anomalies in the Pacific Ocean. On magnetic profiles perpendicular to the axis of a spreading ridge, these anomalies had a remarkably symmetric pattern on both sides of the ridge. Samples of pillow basalts dredged near to ridge crests have been found to possess moderate susceptibilities and large remanent magnetizations. The ocean basalts are able to account for most of the anomalies measured in the offshore. The oceanic crust formed at a spreading ridge acquires a thermoremanent magnetization in accordance with the geomagnetic field of the time of formation. While the polarity of the field remains constant, the crust formed carries the same signature. After a polarity reversal, freshly formed basalts acquire magnetization, which is parallel to the new field and opposite to the previous one.

In cases where oceanic magnetic anomalies have two sources, from the upper basalts and the lower gabbros, the two contributions will be a little out of phase spatially because of the curved depth profiles of cooling isotherms in the oceanic crust. The net effect will be an asymmetry of inclined magnetization directions on opposite sides of the ridge, which needs to be carefully resolved during interpretation. There are contrary view, which need to be established.

Locally anomalies of the continent-ocean margins are interpreted in terms of variations of the crustal composition.

Summary

The physical phenomenon of magnetism is one of the oldest subjects of study by humans. Simplified concepts, on the basis of which this property can be used to gather knowledge about the internal structure of the Earth forms the essence of this article. Details of each aspect have been dealt with in other articles of this volume.

Bibliography

Blakely, R. J., 1995. *Potential theory in gravity and magnetic applications*. Cambridge: Cambridge University Press.
http://geomag.org/models/wdmam.html
http://models.geomag.org/index.html
http://smsc.cnes.fr/html-images/solid_earth.htm
http://www.ngdc.noaa.gov/IAGA/vmod/igrf.html
Lillie, R. J., 1998. *Whole earth geophysics, an introductory textbook for geologists and geophysicists*. Prentice Hall.
Lliboutry, L., 1999. *Quantitative Physics and Geology*. Springer-Praxis Books in Geophysical Sciences.
Lowrie, W., 1997. *Fundamentals of Geophysics*. Cambridge University Press, pp. 229–306.
Tarduno, J. A., et al., 2010. Geodynamo, solar wind, and magnetopause 3.4 to 3.45 billion years ago. *Science*, **327**, 1238.
Telford, W. M., Geldart, L. P., Sheriff, R. E., and Keys, D. A., 1976. *Applied Geophysics*. London: Cambridge University Press.
Vine, F. J., and Wilson, J. T., 1965. Magnetic anomalies over a young oceanic ridge off vancouver Island. *Science*, **150**(3695), 485–489.

Cross-references

Geomagnetic Field, Global Pattern
Geomagnetic Field, IGRF
Geomagnetic Field, Measurement Techniques
Geomagnetic Field, Secular Variation
Geomagnetic Field, Theory
Magnetic Anisotropy
Magnetic Data Enhancements and Depth Estimation
Magnetic Domains
Magnetic Methods, Principles
Magnetic Methods, Satellite
Magnetic Methods, Surface
Magnetometers
Paleomagnetism, Measurement Techniques and Instrumentation
Paleomagnetism, Principles
Remanent Magnetism
Spherical Harmonic Analysis Applied to Potential Fields

MAGNETIC METHODS, SATELLITE

Dhananjay Ravat
Department of Earth & Environmental Sciences,
University of Kentucky, Lexington, KY, USA

Definition

Magnetic field. A vector force field produced by moving electrical charges (electric currents and magnetized material).
Satellite. A compact body orbiting another body in space, and, in this case, human-made container carrying equipment for measuring magnetic fields (magnetometers) and the direction of its sensors.

Introduction

Magnetometers (qv) onboard Earth-orbiting satellites measure precise near-Earth magnetic fields in space from sensors mounted on booms to reduce interference from the spacecraft's magnetic fields. Since the dawn of the age of the Earth-orbiting satellites, the POGO series (three missions during 1965–1971, periapsis (lowest altitude) of about 400 km, scalar only, Cain (2007)), Magsat (6 months during 1979–1980, alt. 325–550 km, Langel et al. (1982)), CHAMP (2000–2010, alt. 300–450 km, Lühr et al. (2009)), Ørsted (1999–2010+, alt. 650–850, Olsen (2007)) and Ørsted-2/SAC-C (2000–2004, alt. ~700 km, Olsen (2007)) satellites have made significant contributions to the field of geomagnetism. The Swarm mission (2011 onward) is a three-satellite constellation, at altitudes below 400 km and above 530 km, and will likely lead to an order of magnitude better resolution on the measured magnetic fields due to the possibility of simultaneously measuring certain external fields in different locations and the ability of measuring magnetic gradients (Friis-Christensen et al., 2006). The magnetic fields sensed in the near-Earth environment primarily originate from the Earth's core field dynamo, lithospheric magnetic fields from rocks' ferromagnetic components, and ionospheric fields and magnetospheric fields of different origin in addition to their electromagnetic-induced counterpart in the conducting mantle, crust, and the oceans. It is possible to model (or simulate) and separate all these components and study their characteristics, but this is difficult to do so precisely, for localized, rapidly time-varying fields due to sampling inadequacies. Slowly varying components like the core field and the nearly static components from the lithosphere can be isolated well with long period observations (i.e., core field models (e.g., POMME-6.0, Maus et al., 2009) and lithospheric field models (e.g., MF6, Maus et al., 2008) from CHAMP and Ørsted satellites). A continuous model of certain aspects of these magnetic fields from POGO to CHAMP/Ørsted satellite era, including the ground magnetic observatory data, has also been achieved for quiet external field conditions (Sabaka et al., 2004).

Two primary types of *Magnetometers* (qv) used aboard satellites are scalar and vector; the scalar magnetometers typically measure total magnetic field intensity (magnitude), whereas the vector magnetometers measure fields in a particular orientation (magnitude, typically, in three orthogonal directions) relative to a baseline. Vector measurements are essential for separating fields above and below the observation plane (Backus, 1970); vector measurements also reduce a particular non-uniqueness in the interpretation of scalar magnetic anomalies close to the magnetic equator (Backus, 1970; but see also, Maus and Haak, 2003).

Polar orbits (orbital inclinations $>80°$) of these satellites allow near-global coverage of the magnetic field observations. The relative position of the satellite with respect to Earth is always changing. For precision measurements (<1 nanotesla or nT) with vector magnetometers (fluxgate type), the knowledge of the attitude (direction) of the sensors is necessary within a few arcseconds and is derived from star cameras (modern satellites have multiple-head star cameras for cross-check and backup). Vector magnetometers are relative instruments and may also drift with time, meaning their baseline changes with time. Thus, their field is calibrated against scalar magnetometers (rubidium, cesium, helium, Overhauser proton precession magnetometers) onboard.

Sources of errors

The current magnetometer instrument accuracy is high (error of < 0.1 nT) (Nielsen et al., 1995) and the magnetic field measurement precision, which include errors in the sensor attitude determination, are also small (± 1 nT) (Olsen et al., 2003). The main difference in the processing of the satellite-borne magnetic data, from those of aeromagnetic/marine/ground surveys for crustal studies (qv *Magnetic Methods, Airborne*) is in the determination and treatment of external fields. External fields are dynamic and, since the satellite flies through the ionosphere, the ionospheric fields cannot be separately monitored for the purpose of their removal as in the case with the magnetic surveys close to the Earth's surface. Spatial and temporal models are necessary of different sources of fields, such as the core field, the lithospheric field, and the ionospheric and magnetospheric fields from a plethora of different phenomena that also involve Earth-Sun interactions. Many of these fields overlap in space and time and with only a few simultaneous observations there can be no perfect isolation of these different sources of fields. The imperfection matters most to the analyses and interpretation of crustal/lithospheric fields because these fields at satellite altitude are relatively small (a few nT and hence the signal-to-noise ratio is low) and contain coalesced effects of individual sources of geologic/tectonic features as a result of the high observation altitude. Also, the magnetic effects of all geologic features of small dimensions are attenuated below the noise level of the observations (Ravat et al., 2002) and information about these features

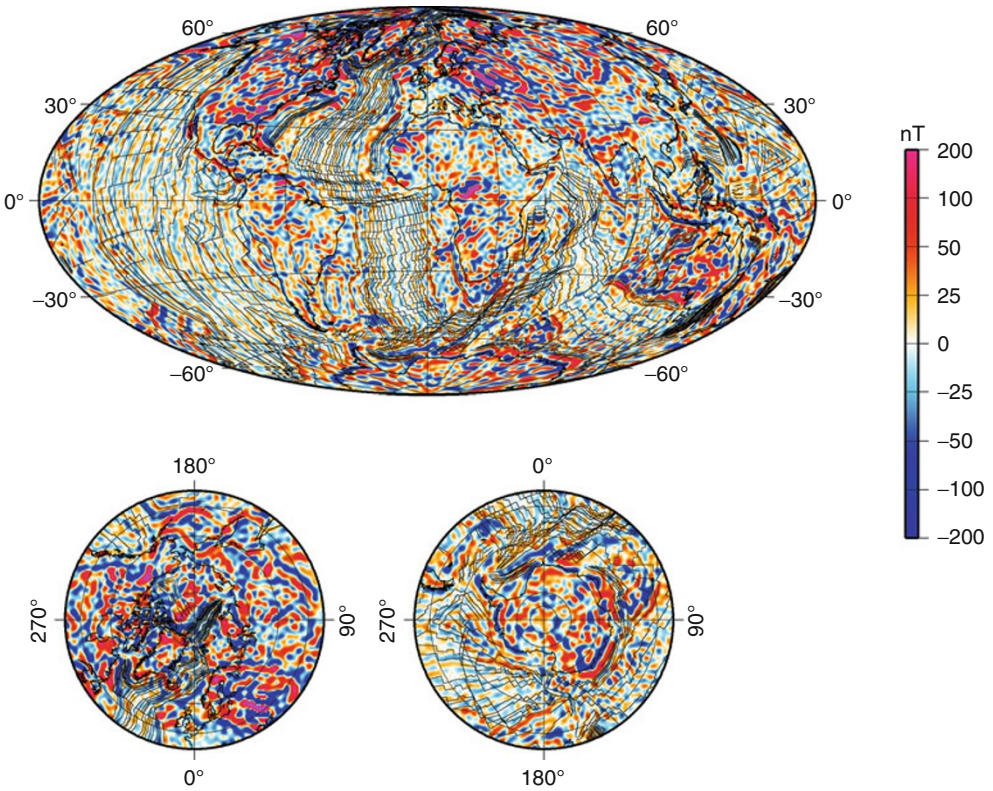
cannot be recovered. To minimize these difficulties for isolating the lithospheric fields, the data are selected from the quietest external field conditions; these quietest external fields are modeled to the best of one's capability, and the remaining external fields are filtered out (Ravat et al., 1995; Langel and Hinze, 1998; Maus et al., 2008). The ability of selecting a large amount of local night-time data (less ionospheric field contribution) at low altitude during geomagnetic quiet conditions (solar minimum) during the epoch of the CHAMP satellite has led to an order of magnitude advance in the determination of the global long-wavelength magnetic fields.

Overall characteristics of the geomagnetic field

Over 99% of the magnetic field measured at the satellite altitude originates from the Earth's core dynamo and only a fraction of a percent arises from the ferromagnetic components in the lithosphere (±20 nT at 350 km elevation, ±200 nT when downward continued to the Earth's surface). The core field varies in intensity from 24,000 nT at the geomagnetic equator to over 65,000 nT at the geomagnetic poles. It is convenient to represent the global field measured from satellites in terms of spherical harmonic coefficients. The core field dominates the longer wavelengths (>2,800 km) represented by the spherical harmonic coefficients of degree < 14, whereas the lithospheric field (i.e., magnetic anomalies) dominates in the complementary shorter wavelength part of the field. The features of the core field and its temporal changes (on the order of < ±150 nT/year) reflect the fluid flow processes in the outer core (Holme, 2007; Whaler, 2007). In addition, satellite magnetic-field data provide the spatial and temporal coverage necessary to study the conductivity variations within the Earth, and the phenomena of geomagnetic jerks (sudden change in the second derivative of the field with time) and the variations in the length of the day (i.e., changes in the Earth's rotation rate) (Olsen, 2007). Examples of the interpretation of the satellite-derived magnetic anomalies are given in Ravat (2007) and Purucker and Whaler (2007), and their features and the current use are described below.

Features of the lithospheric satellite-derived magnetic field

Figure 1 shows the vertical component of the magnetic field measured by CHAMP satellite, downward continued to the surface of the Earth. For the following discussion, features of this field should be compared to the CGMW, Geological Map of the World (http://ccgm.free.fr/index_gb.html). Even though there is no observable continent-ocean boundary-edge effect in the data, statistics is no longer needed to



Magnetic Methods, Satellite, Figure 1 Vertical component of the lithospheric magnetic field at the Earth's surface from CHAMP MF6 magnetic field model from Maus et al. (2008), superimposed with isochrons of the ocean-age model by Müller et al. (2008).

differentiate between continental and oceanic magnetic variations observed on this map. The correlation of the features of the magnetic field with geologic provinces is strong. The magnetic field over the oceans is vastly smaller in amplitude except over the edges of the magnetic polarity quiet zones (qv *Geomagnetic Field, Polarity Reversals*) and many oceanic islands and plateau regions. Over the continents, where not attenuated due to rise in the Curie isotherm (qv *Curie Temperature*) from recent tectonic/magmatic events, the magnetic field over the Archean and Precambrian regions stands out in comparison to Phanerozoic lithosphere. Many of the oceanic North–South-oriented stripes parallel to North–South trending oceanic ridges are real in this map, but some could be an artifact of the pass-by-pass filtering and misleveling of base levels in adjacent satellite passes (see, e.g., the NW-SE trending South West Indian Ridge between Australia and Antarctica also has north-south features perpendicular to the ridge – the expectation here is that stripes should be parallel to the ridge). Overall, however, the MF6 magnetic anomaly field (Maus et al., 2008) is a vast improvement over its predecessors and amenable to quantitative geologic modeling and new geologic interpretations.

Modern uses of satellite-altitude magnetic anomalies

Despite their long-wavelength nature compared to near-surface magnetic-anomaly data, there are several advantages and uses of satellite-altitude magnetic anomalies of lithospheric origin: (1) Availability of vector data make it possible to minimize some of the instabilities near the magnetic equator (Backus, 1970); (2) Several magnetic surveys since the inception of the aeromagnetic and marine magnetic surveying in the 1940s have been processed using ad-hoc processing of long-wavelengths fields and processing with different IGRF core-field models (qv *Geomagnetic Field, IGRF*) that are piecewise continuous, and hence cause artifacts in regional compilations. These problems have led to spurious intermediate- and long-wavelengths to the extent that misinterpretation of lower-crustal magnetic features can occur (e.g., zones of regional magnetization, depth to Curie isotherm/magnetic bottom). These problems can be overcome by replacing the intermediate- and long-wavelength field by downward-continued satellite-altitude field in regional compilations (Ravat et al., 2002; Bankey et al., 2002). The long-wavelength CHAMP field has also been used to provide a common datum in the preparation of the World Digital Magnetic Anomaly Map (http://projects.gtk.fi/WDMAM/project/perugia/index.html, Korhonen et al., 2007); and (3) Several regions of the world have little or no coverage of magnetic anomalies; in these regions, the downward-continued magnetic field gives some idea of the intermediate- and long-wavelength anomaly field that could be useful in regional geologic interpretation.

The main limitation of satellite-altitude magnetic fields is that the magnetic effects of geologic sources have coalesced at satellite altitudes. High precision of the MF6 field has partly circumvented this issue by downward continuation of the field to the surface; however, the part of the short-wavelength signal that is in the noise level at the satellite altitudes cannot be recovered. Thus, only features of >350 km wavelength can be reliably mapped with the present satellite-altitude data. Another milestone to be achieved is the separation of the lithospheric field from the field attributed to the core of the Earth (spherical harmonic degree <14). The latter may pose challenges in the recognition of the continental and oceanic boundaries and a number of large cratonic regions of the world (as discussed by Purucker et al., 1998; however, this inference has been challenged by the study of Hemant and Maus (2005)).

Summary

Over the last decade, the availability of continuous time series of high-precision magnetic measurements from space (Ørsted, CHAMP and Ørsted-2/SAC-C satellites) has significantly advanced the knowledge of magnetic fields observed in the near-Earth environment compared to the previous space-borne magnetic observations. These magnetic fields are produced by many sources, ranging from the core of the Earth to the regions deep in the Earth's magnetosphere, and include also effects caused by the Sun's direct influence on the Earth. The newest CHAMP-based lithospheric field models (e.g., MF6, Maus et al., 2008) have a wealth of geologic information that can be used to improve regional and local magnetic survey compilations and decipher regional geologic characteristics difficult to assess from local or regional magnetic anomaly compilations or areas devoid of near-surface magnetic measurements on the Earth.

Bibliography

Backus, G. E., 1970. Non-uniqueness of the external geomagnetic field determined by surface intensity measurements. *Journal of Geophysical Research*, **75**, 6339–6341.

Bankey, V., Cuevas, A., Daniels, D., Finn, C. A., Hernandez, I., Hill, P., Kucks, R., Miles, W., Pilkington, M., Roberts, C., Roest, W., Rystrom, V., Shearer, S., Snyder, S., Sweeney, R., Velez, J., Phillips, J. D., and Ravat, D., 2002. Digital data grids for the magnetic anomaly map of North America. *U.S. Geological Survey Open-File Report 02-414*. U.S. Geological Survey, Denver, CO.

Cain, J. C., 2007. POGO (OGO-2, -4 and -6 spacecraft). In Gubbins, D., and Herrero-Bervera, E. (eds.), *Encyclopedia of Geomagnetism and Paleomagnetism*. Dordrecht: Springer, pp. 828–829.

Friis-Christensen, E., Lühr, H., and Hulot, G., 2006. Swarm: A constellation to study the Earth's magnetic field. *Earth, Planets and Space*, **58**, 351–358.

Hemant, K., and Maus, S., 2005. Why no anomaly is visible over most of the continent-ocean boundary in the global crustal magnetic field. *Earth and Planetary Interiors.*, **149**, 321–333.

Holme, R., 2007. Large-scale flow in the core. In Olson, P. (ed.), *Treatise on Geophysics*. Amsterdam: Elsevier Ltd, Vol. 8, pp. 107–130.

Korhonen, J., Fairhead, J. D., Hamoudi, M., Hemant, K., Lesur, V., Mandea, M., Maus, S., Purucker, M., Ravat, D., Sazonova, T., and Thebault, E., 2007. *World Digital Magnetic Anomaly Map*. CCGM-CGMW/UNESCO. 1:50 000 000, 1 sheet and 1 DVD.
Langel, R. A., and Hinze, W. J., 1998. *The Magnetic Field of the Earth's Lithosphere—The Satellite Perspective*. Cambridge: Cambridge University Press.
Langel, R., Ousley, G., Berbert, J., Murphy, J., and Settle, M., 1982. The MAGSAT mission. *Geophysical Research Letters*, **9**, 243–245.
Lühr, H., Korte, M., and Mandea, M., 2009. The recent geomagnetic field and its variations. In Glassmeier, K.-H., Soffel, H., and Negendank, J. (eds.), *Geomagnetic Field Variations*. Berlin-Heidelberg: Springer-Verlag, pp. 25–63.
Maus, S., and Haak, V., 2003. Magnetic field annihilators: invisible magnetisation at the magnetic equator. *Geophysical Journal International*, **155**, 509–513.
Maus, S., Yin, F., Lühr, H., Manoj, C., Rother, M., Rauberg, J., Michaelis, I., Stolle, C., and Müller, R. D., 2008. Resolution of direction of oceanic magnetic lineations by the sixth-generation lithospheric magnetic field model from CHAMP satellite magnetic measurements. *Geochemistry Geophysics Geosystems*, **9**, Q07021, doi:10.1029/2008GC001949.
Maus, S., Manoj, C., Rauberg, J., Michaelis, I., Rother, M., and Lühr, H., 2009. NOAA/NGDC candidate models for the 11th generation International Geomagnetic Reference Field and the concurrent release of the 6th generation POMME magnetic model. http://www.geomag.us/models/pomme6.html.
Müller, R. D., Sdrolias, M., Gaina, C., and Roest, W. R., 2008. Age, spreading rates, and spreading asymmetry of the world's ocean crust. *Geochemistry Geophysics Geosystems*, **9**, Q04006, doi:10.1029/2007GC001743.
Nielsen, O. V., Petersen, J. R., Primdahl, F., Brauer, P., Hernando, B., Fernandez, A., Merayo, J. M. G., and Ripka, P., 1995. Development, Construction and Analysis of the "ØRSTED" Fluxgate Magnetometer. *Measurement Science and Technology*, **6**, 1099–1115.
Olsen, N. I., 2007. Ørsted. In Gubbins, D., and Herrero-Bervera, E. (eds.), *Encyclopedia of Geomagnetism and Paleomagnetism*. Dordrecht: Springer, pp. 743–746.
Olsen, N., Tøffner-Clausen, L., Sabaka, T. J., Brauer, P., Merayo, J. M. G., Jørgensen, J. L., Léger, J.-M., Nielsen, O. V., Primdahl, F., and Risbo, T., 2003. Calibration of the Ørsted vector magnetometer. *Earth Planets Space*, **55**, 11–18.
Purucker, M., and Whaler, K., 2007. Crustal magnetism. In Olson, P. (ed.), *Treatise on Geophysics*. Amsterdam: Elsevier Ltd, Vol. 5, pp. 195–235.
Purucker, M. E., Langel, R. A., Rajaram, M., and Raymond, C., 1998. Global magnetization models with a priori information. *Journal of Geophysical Research*, **103**, 2563–2584.
Ravat, D., 2007. Crustal magnetic field. In Gubbins, D., and Herrero-Bervera, E. (eds.), *Encyclopedia of Geomagnetism and Paleomagnetism*. Dordrecht: Springer, pp. 140–144.
Ravat, D., Langel, R. A., Purucker, M., Arkani-Hamed, J., and Alsdorf, D. E., 1995. Global vector and scalar Magsat magnetic anomaly maps. *Journal of Geophysical Research*, **100**, 20111–20136.
Ravat, D., Whaler, K. A., Pilkington, M., Sabaka, T., and Purucker, M., 2002. Compatibility of high-altitude aeromagnetic and satellite-altitude magnetic anomalies over Canada. *Geophysics*, **67**, 546–554.
Sabaka, T. J., Olsen, N., and Purucker, M. E., 2004. Extending comprehensive models of the Earth's magnetic field with Ørsted and CHAMP data. *Geophysical Journal International*, **159**, 521–547, doi:10.1111/j.1365-246X.2004.02421.x.
Whaler, K. A., 2007. Core motions. In Gubbins, D., and Herrero-Bervera, E. (eds.), *Encyclopedia of Geomagnetism and Paleomagnetism*. Dordrecht: Springer, pp. 84–89.

Cross-references

Curie Temperature
Geomagnetic Field, IGRF
Geomagnetic Field, Measurement Techniques
Geomagnetic Field, Polarity Reversals
Magnetic Anomalies, Interpretation
Magnetic Methods, Airborne
Magnetic Methods, Principles
Magnetic Methods, Surface
Magnetometers

MAGNETIC METHODS, SURFACE

Nandini Nagarajan
National Geophysical Research Institute,
Hyderabad, India

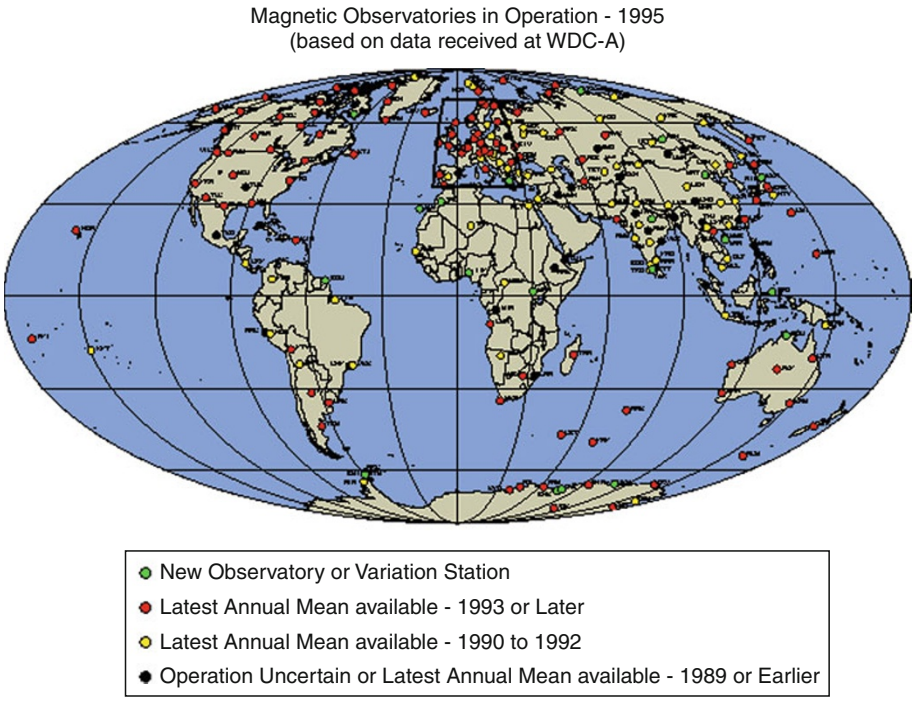
Definition

Magnetic observatory. Magnetic observatories continuously measure the Earth's magnetic field at fixed locations at the Earth's surface. They are designed to operate for several decades to stably record rapid and slow variations of the earth's magnetic field and report time-averaged values that estimate the form of the Earth's magnetic field. Since observatories started operating in the nineteenth century, the methods of measurement, reporting, and analysis have varied over region and time. After the International Geophysical Year (IGY – 1957–1958), the quality and organization of the data have been standardized, making spatial and temporal studies easier, more refined, and consequent models more accurate. The International Association for Geomagnetism and Aeronomy (IAGA) is the coordinating body for the magnetic observatories of the world, standardization of measurement, exchange of data, and presentation of models of the Earth's magnetic field. There are about 200 magnetic observatories reporting standardized data, and several more making continuous variation measurements (Figure 1).

INTERMAGNET. Classical torsion magnetometers, with analogue-photographic systems, were replaced in the 1980s by improved fluxgate magnetometers and digital data acquisition systems. More recently, a magnetic standard evolved by IAGA, INTERMAGNET (International Real-time Magnetic Observatory Network), has resulted in a network that provides one-minute values of the three components of the geomagnetic field, in near-real time. INTERMAGNET standards of measurement and digital reporting have been adopted by over 100 observatories, worldwide.

Observatory and semi-long-term measurements

Magnetic observatory data are used to report and study periodic and episodic variations that are classified by time scale and mechanism: Long term secular variation, 11-year solar cycle, annual, semi-annual, seasonal, and daily

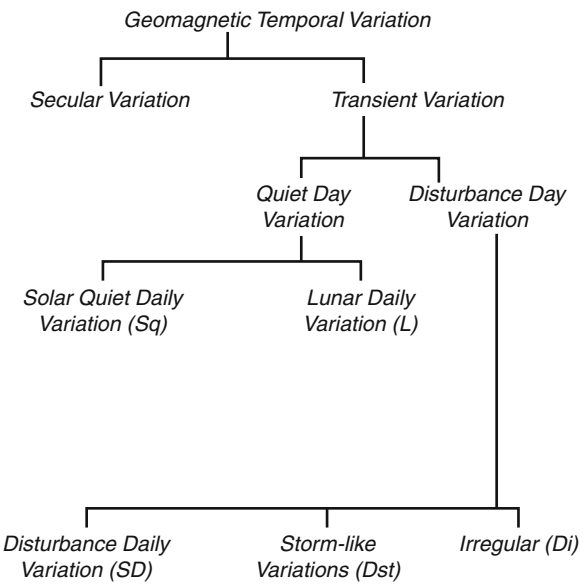


Magnetic Methods, Surface, Figure 1 Currently operating observatories, updated in 1996 (Jankowski and Sucksdorff, 1996).

variation and shorter transient variations classed as disturbances (Figure 2).

Observatory yearbooks often give the average solar daily variation for different months of the year, and even separately for quiet, disturbed and all days. Based on an analysis of magnetic activity by trained observers, and estimation of the activity indices (most commonly K), five quiet and disturbed days are selected for each month by the International Service of Geomagnetic Indices.

Magnetic variations have characteristics, varying over latitude and period attributed to different sources – solar activity from 11 years to 27 days; diurnal harmonics of ionospheric currents from 24 to 4 h; shorter events like sudden commencements, sudden impulses, substorms (bays) from −2 h to 1 min; magnetic pulsations, with small amplitude and short periods from 1 to 60 s. Hourly and daily variations are a measure of solar interaction with the Earth's magnetic field. This can be both solar radiation, interacting daily with the ionosphere or corpuscular streams interacting with the magnetosphere, manifested as disturbances. Regular solar daily variation is also a function of season, solar activity, and latitude. The variation of the magnetic field on magnetically quiet days is called the solar quiet day variation or Sq variation. Magnetic storms (∼day), substorms (∼hour), pulsations (∼min), sudden impulses, and solar flares are some of the transient disturbances. The longer, slower variations are obtained from monthly and annual means of observatory data. The trends seen in the data complied over the last 150 years have been used to produce spherical harmonic representation of the Earth's magnetic field that is internally



Magnetic Methods, Surface, Figure 2 Diagram showing the different types of geomagnetic variations (Okeke et al., http://www.saao.ac.za/∼wgssa/as4/fokeke.html).

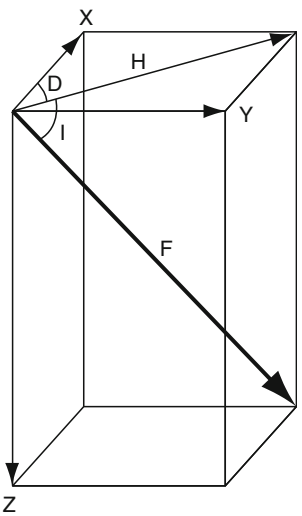
generated and its secular variation that are indicative of processes within the Earth's mantle and outer core. Definition of the main magnetic field and secular variation are currently supplemented by low-earth orbiting satellites (OERSTED,

CHAMP missions, for example). This is one of the fastest growing areas of study.

Magnetic measurements

Vector and scalar measurements: Magnetic measurements can be scalar (total field) or vector. To define the vector, three out of seven elements need to be measured simultaneously (Figure 3). Commonly, any three of these elements are measured in a magnetic observatory as variations, to fully define variations of the total magnetic field.

Absolute measurements: Most vector magnetometers used in magnetic observatories (earlier torsion, and recently fluxgates) usually drift in time due to several factors such as temperature variations, magnetic moment, and aging of the electronics. In order to periodically recalibrate the vector measurements, absolute measurements fully defining the magnetic field vector are performed, to complement continuous measurements, on a regular basis (typically once a week). Such measurements are made by a trained observer using a single-axis fluxgate magnetometer mounted onto a nonmagnetic theodolite and a scalar magnetometer. From which, the continuous measurements can be calibrated to a baseline defined by absolute measurements. Two types of data are available from observatories: preliminary data, which are made available in quasi-real time (less than 72 h), and definitive data, which are produced only once a year, typically a few months after the end of the calendar year. After reducing time variations to absolute baseline at a location, continuous observations of the vector magnetic field are produced at desired sampled intervals: minute, hour, day, month, year, etc. These data are available for all years going back to the start of observations at each observatory. The format for easy assimilation and distribution of this data are formulated by IAGA working groups – updated periodically since 1957–1958. This data is also available at World Data Centers for Geomagnetism, established in different countries, partly by their research institutions and partly funded by IAGA.



Magnetic Methods, Surface, Figure 3 Magnetic elements, defined in geographic and geomagnetic coordinates. (Jankowski and Sucksdorff, 1996).

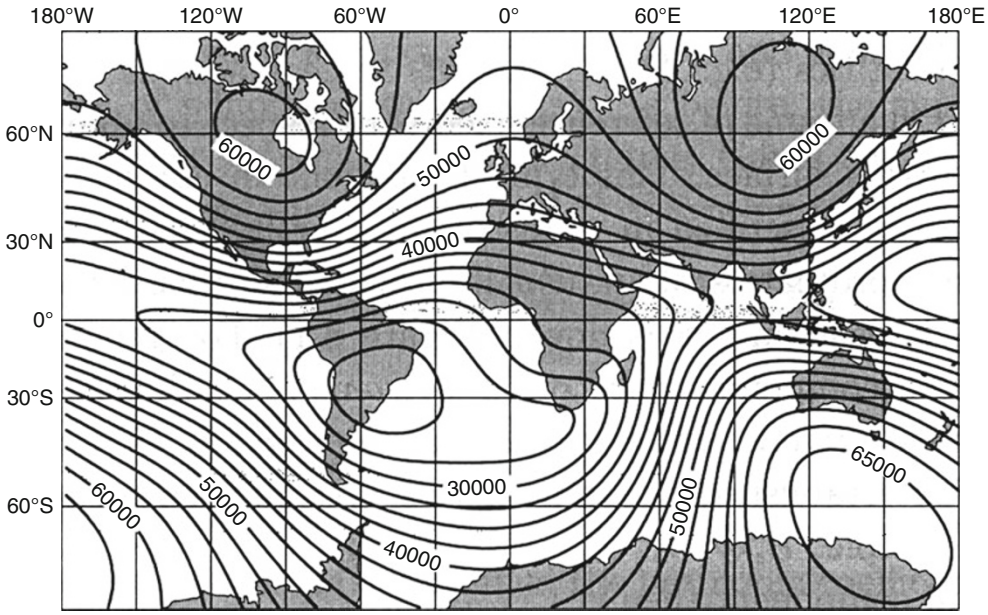
Products of observatory data

Magnetic field values at different points on the Earth, their variation on different timescales are obtained. The Earth's magnetic field has been approximated as a dipole and more accurately expressed as a spherical harmonic series, based on time-averaged means of observatory data, by Gauss, demonstrating that 95% of the field intensity is derived from inside the Earth and 5% due to external currents in the atmosphere (now known as magnetosphere). The magnitude of the magnetic field components varies from 0 to 60,000 nT. From variations of the magnetic field at regularly sampled intervals, much information can be extracted. The time-averaged values over months and years provide a data set approximating the potential field of the internally generated magnetic field. The absolute value of the field at a point on the Earth can therefore be estimated with accuracy of ~0.01%. These values define the shape of the Earth's magnetic potential field (Figure 4). This is termed the International Geomagnetic Reference Field, for a stated epoch (instant in time). Other data sets from magnetic surveys and satellites are added to refine the IGRF (see *Geomagnetic Field, IGRF*).

Secular variation: Variation in monthly and annual means defines the secular, slow change of the Earth's internal magnetic field. This has been mapped going back to about 300 years. Secular variation is, however, a small variation compared to absolute values and is determined as change over years of averaged variation. The change is as small as ~10 nT/yr. The pattern of secular variation places constraints on the theory of generation of the Earth's magnetic field and maps fluid flow in the outer core. It is also the only means of studying electromagnetic and thermal properties of the Earth's outer core (depth: 3,000 km) (see *Geomagnetic Field, Secular Variation*).

Solar radiation and quiet day variations

Daily variation at each observatory has contributed to the definition of solar-induced daily ionospheric currents (Sq). Variation with latitude and season has also been defined (Figure 5). These variations range between 10 and 100 nT, in each component. In this, diurnal variation is the most prominent. Monthly, seasonal, and annual variations are largely controlled by radiation (F10.7) from the sun. The effects of solar radiation on the ionosphere drive electric current loops in the ionosphere (height ~100 km). Study of these regular variations from both observatory data and in recent decades, from satellite data, has resulted in better models of the current loops and their variability. In addition, solar-induced disturbances affect the earth's magnetic field.



Magnetic Methods, Surface, Figure 4 Spatial variation of the Earth's magnetic field scalar (F), approximating a dipole field, over the Earth (After Jankowski and Sucksdorff, 1996).

Magnetic disturbances and solar activity

Study and classification of magnetic disturbances, the influence of sunspot activity, etc., were noted and a theory explaining these phenomena evolved, from a study of observatory data, during 1920–1940. This theory has since been corroborated from a direct observation of corpuscular radiation streams from the Sun and magnetospheric currents, by satellites. Magnetic disturbances are of varying time period from minutes to days. Magnetic storms are the most prominent among these with characteristics variations of ~100 nT. Short period events like bays, sudden commencement, and solar flares have amplitudes of ~10 nT [magnetic pulsations, storms].

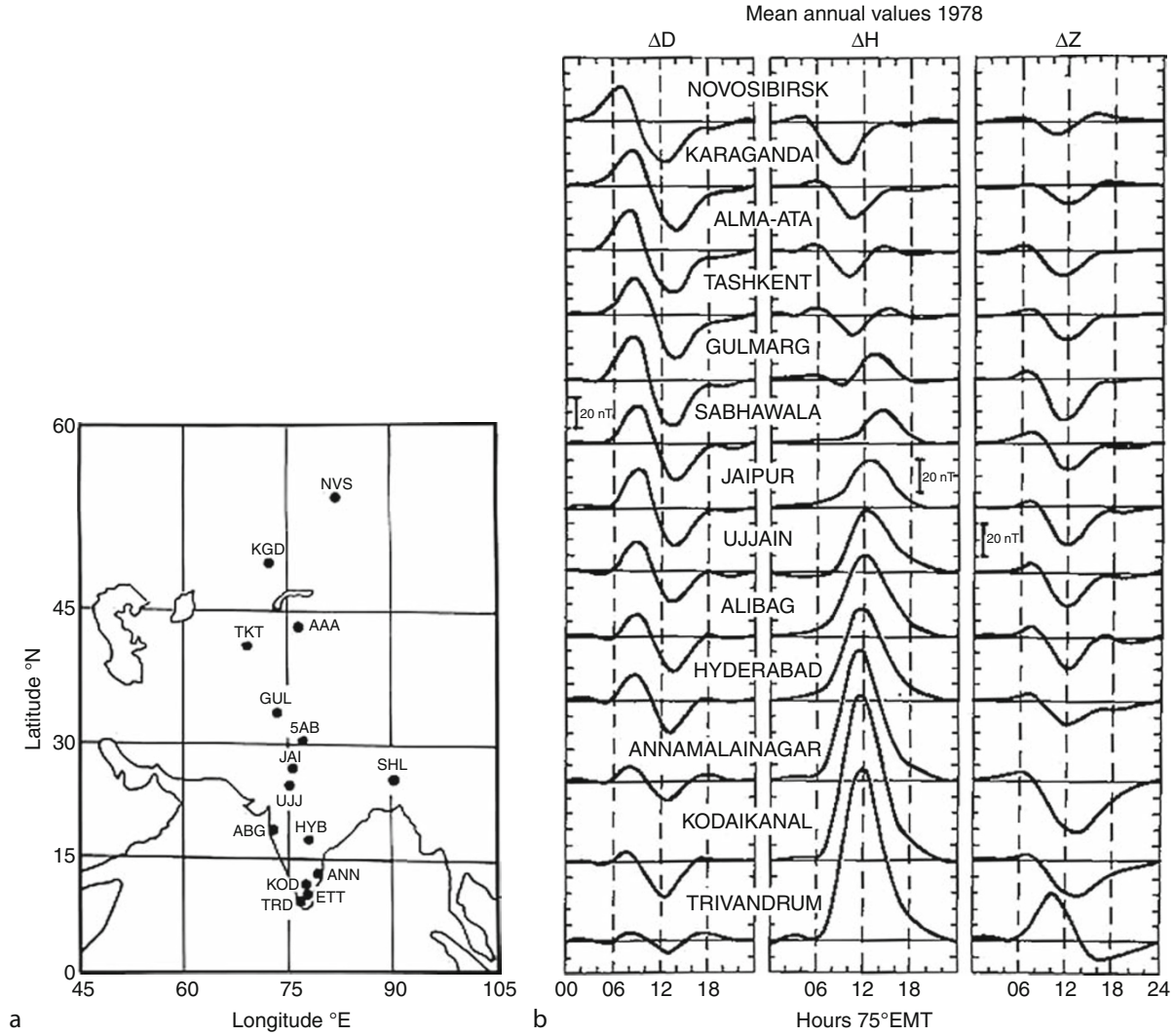
Each day's variations are analyzed and classified from quiet to disturbed, as a measure of magnetic activity in the earth's magnetosphere, in a manner that eliminates geographic difference. Geomagnetic indices were formulated as a composite of reported variations from observatories at different locations. These indices, and several new ones, continue to be reported and are updated, revised in accordance with discussion and resolutions passed by IAGA.

Uses of observatory data

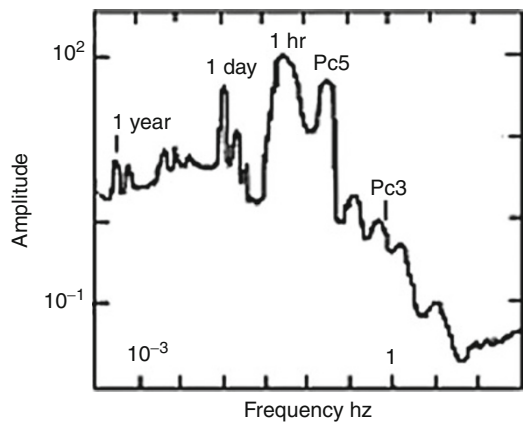
From the individual magnetic observatories established 150 years ago, and some earlier measurements, a pattern of global behavior of the magnetic field came to be defined. After the International Geophysical year (IGY), in 1957, studies of the earth's physical properties, time variations and mechanisms, intensified. Observatory data are useful in determining rapid, medium term as well as slow variations in earth's magnetic field. Each range of variations, studied for specific signals, contributes to the understanding of different processes (Figure 6).

Some of the phenomena being currently studied are:

1. Secular variation of the magnetic field at different places on the Earth. It is then expressed in spherical harmonic coefficients and included in the IGRF by adding time derivatives of coefficients. Knowledge of these variations is important in theoretical modeling of the geomagnetic dynamo. It is the only means of studying the motions of the electrically conducting fluid in the Earth's outer core, where the main geomagnetic field is generated. These data are also widely used to calculate time-varying core field models, often in combination with satellite magnetic data or historical data (see *Geomagnetic Field, Global Pattern*).
2. The regular variations of the magnetic field are related to rotation and/or orbital movements of the Earth, Sun, and Moon. The most prominent is the diurnal variation or solar daily variation, having amplitudes of the order of 10–100 nT(nanotesla). Solar radiation ionizes the higher atmosphere during the daylight hours, and the gravitational forces of the Sun and the Moon force the ionospheric layers in a tidal motion. So the ionized gas in the ionosphere moves in the magnetic field of the Earth, creating electric currents which are seen as daily variations in magnetic recordings. There are two well-known periodic variations, the solar daily variation and the lunar daily variation.
3. A global network is essential in monitoring the pattern of geomagnetic storms, local disturbances, mapping of



Magnetic Methods, Surface, Figure 5 Diurnal variation of the magnetic field, showing changes with latitude for solar quiet days (**a**) locations of observations (**b**) variations with latitude.



Magnetic Methods, Surface, Figure 6 An illustration of the wide spectrum geomagnetic variations (~1 s to 1 year) (Serson, 1983).

currents generated in the magnetosphere, and thereby, inferences about presence of charged particles, models of magnetic field lines.

4. Induced currents in the crust and mantle are caused by external variations of the earth's magnetic field. For each periodic and episodic event described above, a corresponding current is induced in the finitely conducting layers of the Earth. An estimation of these currents provides a measure of the electrical conductivity distribution within the Earth's crust and mantle. This has been applied as a tool to study the Earth's interior and a large volume of electromagnetic variation data has been collected solely to this purpose. In the past decade induced signals from satellite magnetic data have been used to model electrical conductivity distribution, over continents and oceans (see *Geoelectromagnetism*).

5. Ground observation networks provide calibration and control of magnetic survey satellites and complement detailed ground surveys of the magnetic field.
6. Magnetic pulsations –10 Hz to 600 s – are small amplitude variations generated at the boundary between magnetosphere and solar wind, under various dynamic conditions and transmitted through the magnetosphere and ionosphere, and carry signatures of the properties of these regions as well.

Land surveys

Magnetic surveys are conducted on land to determine changes in physical properties of subsurface rocks (viz., magnetic susceptibility). Magnetic field scalar (or vector) is measured using a suitable portable magnetometer at precisely determined points in the survey area. The measurements provide variation of the magnetic field strength over the area surveyed, at a desired sampling length (in m or km), depending on the extent or depth of anomaly expected. (*Magnetic Methods, Airborne; Magnetic Methods, Principles; Magnetometers; Magnetovariational Studies*) At the same time a record of temporal variations is made, at a central point in the survey area, or by tying up the survey to a nearby observatory. The field values, which could be scalar values of total field, or a component of the vector are then corrected for diurnal variation occurring during the survey. This is a crucial aspect of magnetic surveys. If the survey is over a sufficiently large region (∼100 km), IGRF values of the main magnetic field are also subtracted from the data points. The variation over the sampled points then indicates spatial changes, which are interpreted in terms of changes in magnetic properties of subsurface rocks. The variations found from this data are ∼100 nT and in the case of very large anomalies it maybe ∼1,000 nT. The accuracy of determining these anomalies depends on the accurate measurements, both at survey points and base stations. Most modern magnetometers can measure a spot value to 0.1 nT accuracy. But determining the anomalous value from this requires precise reduction of data. Errors of alignment, timing, and baseline values can affect subsequent analyses. Presence of man-made electromagnetic effects also affect surveys. Therefore, it is it is possible that errors of upto 2 nT are present. Fast surveys like marine and aeromagnetic are most effective in obtaining relative changes, rather than accurate estimation of magnitudes.

Most commonly, change in magnetic susceptibility indicates a change in composition of rock. In certain instances, in the study of crustal magnetization, it could indicate thermal changes (*Global Magnetic Anomaly Map*). Since the amount of ferromagnetic minerals varies considerably, among different crustal rocks, magnetic surveys are a quick and effective method for detecting subsurface variation in rock type. Magnetic surveys are a viable method to obtain information regarding the crystalline basement over areas of sedimentary cover. The method has also been used to provide information on upper crustal structure and thermal state, along major geotransects. The results of magnetic surveys are presented in the form of contour maps of intensity of total or any field component. The anomaly maps are prepared after subtraction of the IGRF, and do not contain any signatures of the dipole field of the earth, including averaged crustal effects. Such maps/data grids can be modeled as a simple potential field caused by the anomalous susceptibility distribution alone. This is the utility of magnetic surveys. Modeling of these potential fields is considerably advanced, with inversion over two and three dimensional bodies possible (see *Magnetic Anomalies, Interpretation*).

The most significant change in magnetic properties occurs at Curie point, the depth varies considerably in different tectonic settings. Therefore, large-scale magnetic surveys provide significant inputs to deep crustal studies. Mapping of spatial magnetic variation provides information about the thickness of the basement rock beneath sedimentary basins/supracrustal rocks and thermal state of the crust of that region.

Marine magnetic surveys

Methods of marine magnetic surveys are more elaborate, since instruments have to be towed behind a vessel and data collection is challenging. However, principles of analysis are similar. These methods map the variation in thickness of the oceanic crust – which consists of uniform magnetic material (basalt). In the case of marine magnetic surveys, the errors in measurement are greater, since the vessel tows the magnetic sensor, it is liable to more movement, making accurate measurement of position and magnetic value more error-prone. The magnetic anomalies arising from distribution of magnetic basalt on the ocean floor are at a distance of 1–5 km from the measuring platform (ocean surface). From a careful elimination of errors due to wave noise, position tracking, and instrument noise, long track of marine magnetic anomalies have been recorded. Over several decades, measurements have been made in the campaign mode, and latest measurements with better instruments have been tallied with older measurements, so that a maximum amount of data could be utilized. The amplitude of marine magnetic anomalies is ∼100–1,000 nT, after removal of IGRF values. The geometry of marine magnetic anomalies is striking, they trend in parallel lines and follow a long linear track across the oceans. This feature has been recognized as bands of basalt that have been added to the ocean floor in different ages, creating new ocean floor, for over 200 million years. Significant variations in marine magnetic measurements have been attributed to pulses of ocean-floor formation. These pulses belonging to different epochs have been identified by adjacent stripes of differing magnetization of the ocean floor – a significant discovery of the phenomenon of "seafloor spreading" (*Paleomagnetism, Magnetostratigraphy*).

Principles of survey measurement

The measurement of spatial variation of magnetic field depends on determining the value at a spot, obtaining the position, and time of measurement, and eliminating temporal variation, during the period of the survey, at each location. For this, data have to be reduced to a time epoch in contrast to observatory measurements, which provide time variations at a fixed point. The objectives of a survey govern the use of specific magnetic instruments, surveying parameters (for example, observation elevation and data spacing), and the coordinate system used for the analysis of the data. For geological interpretation either total field scalar or vector measurements are suitable. But mapping of internally induced currents – in the crust, for tectonic studies, or in the mantle and outer core, for studies of the deep interior – requires vector measurements. Instruments used vary from torsion magnetometer, to fluxgate, proton precession, SQUID, etc. The precision of most modern instruments is between 0.1 and 1.0 nT.

The residual, after removing temporal variation, is the variation in potential caused by deflection of field lines, through a volume of subsurface that has nonuniform magnetic properties.

Instrumentation technologies

The principles of magnetometers are:

1. Torsion magnetometers measure the opposing force to the magnetic field – as in torsion fiber suspended magnetometers and deflection measurements. This was the oldest and mostly accepted method for nearly 200 years. Examples: Quartz horizontal magnetometer (QHM), Balance magnetometer Zero (BMZ), Kew Magnetometer, La Cour/Eschenhagen variometers, and others.
2. Flux measurement, amplified and measured in a ferrite sensor, with appropriate hysteresis – fluxgate magnetometers. Accuracy of fluxgate magnetometers can be adapted to different ranges of measurement. When mounted in triaxial mode, they are give reliable values of the field. The Magnetometers are sensitive to temperature variation and internal noise, but these have been minmized by improved design. They are suitable for observatories, satellites, and unmanned vector measurements.
3. Magnetic Theodolite: Estimating opposing force to the magnetic field as applied to a sensor of finite magnetic moment – using an oscillating magnet and theodolite, an old, classical method for absolute values – several instruments are still in use as secondary standards. A modern version of this is the Declination Inclination magnetometer. Thus uses a singly fluxgate counted on a theodolite. This is adjusted to a null position to find the inclination of the magnetic field. In conjunction with total field measurement, this would result in vector components. The theodolite also gives azimuth and declination measurements. Since it is a null observation, accuracy is good. This has been adopted at standard for modern magnetic observatories and is increasingly used for surveys also.
4. Precession: Based on the frequency of precession of hydrogen atoms (protons) after a strong external field is applied. Precession frequency is proportional to the ambient field. This is the most stable accurate method at present. Counting of frequency is very accurate with improved digital time stamping. The proton precession magnetometer measures only the scalar of total field intensity. An improvement of the same principle is the Overhauser magnetometer, which stimulates precession almost continuously and thereby produces a steady signal that is a measure of ambient field. This magnetometer is now chosen for most remote, continuous, and surveying operations.
5. Optically pumped magnetometers: The principle is of optical excitation of cesium (or other) gas, with a focused beam of coherent light. Zeeman splitting effect produces splitting of the lines emitted by the sample gas. The frequency of split is proportional to the ambient magnetic field.
6. Induction coil magnetometers: These operate on the simple principle of measuring the output of a coil of many windings around a ferrite core. The signal is proportional to the derivative of the magnetic field. The induction coil has good response of a wide frequency range of measurement. This is used for rapid variation, e.g., pulsations, magnetotellurics, etc.

Summary

The shape, strength and variations of the Earth's magnetic field, have been studied over 500 years. Starting from measurements, repeated in time, at different places on earth, a model of the Earth's magnetic field, in space and time has been built up. Magnetic observatories are places where long time series of accurate measurements of the magnetic field are made, analysed and disseminated to institutions and persons seeking this information. Variations in different period ranges provide information about diverse processes, affecting the magnetic field. Short period variations (∼min, hr) are indicative of disturbances in the magnetosphere caused by the Sun's corpuscular streams and magnetic fields, medium-term variations (∼day, year) describe effects of the sun's radiation and long-term variations (∼year, decade, century) provide insights into the nature of the Earth's internal magnetic field, its origin, and modification in the Earth's outer core (secular variation). Closely spaced measurements on the Earth's surface are used in studies for resource exploration, tectonics and mapping the formation of the oceanic crust.

The form of the Earth's magnetic field is defined every five years, collating all available measurements as the IGRF. This also provides estimates of secular variation. The vast quantity of data collected about spatial and temporal variations of the Earth's magnetic field, has resulted in increasingly accurate models of processes on the Sun, in space and on and within the Earth.

Bibliography

Campbell, W. H., 1997. *Introduction to Geomagnetic fields*. Cambridge: Cambridge University Press.
Chapman, S., and Bartels, J., 1940. *Geomagnetism*. Oxford: Clarendon, Vol. I, II.
http://wdc.kugi.kyoto-u.ac.jp/-WDC
http://www.geomag.bgs.ac.uk/earthmag.html
http://www.icsu-fags.org/ps06isgi.htm
Jacobs, J. A. (ed.), 1989. *Geomagnetism*. London: Academic.
Jankowski, J., and Sucksdorf, C., 1996. *Guide for Magnetic Measurements and Observatory Practice*. Warsaw: IAGA, 238 pp.
Kearey, P., Brooks, M., and Hill, I., 2002. *An Introduction to Geophysical Exploration*, 3rd edn. New York: Wiley.
Matsushita, S., and Wallace, H. C. (eds.), 1967. *Physics of Geomagnetic Phenomena.* New York: Academic.
Maus, S., Macmillan, S., Chernova, T., Choi, S., Dater, D., Golovkov, V., Lowes, F., Lühr, H., Mai, W., McLean, S., Olsen, N., Rother, M., Sabaka, T., Thomson, A., and Zvereva, T., 2005. IGRF – The 10th generation international geomagnetic reference, International Association of Geomagnetism, Aeronomy (IAGA), Division V, Working Group VMOD, *Physics of the Earth and Planetary Interiors*, **151**, 320–322.
Merrill, R. T., McElhinny, M. W., and McFadden, P. L., 1998. *The Magnetic Field of the Earth*. London: Academic Press International Geophysics Series, Vol. 63.
Parkinson, W. D., 1983. *Introduction to Geomagnetism*. Edinburgh: Scottish Academic.
Serson, S. H., 1973. Instrumentation for induction studies on land. *Physics of the Earth and Planetary Interiors*, **7**, 313–322.
www.intermagnet.org

Cross-references

Geomagnetic Field, Measurement Techniques
Geomagnetic Field, Theory
Magnetometers

MAGNETIC MODELING, THEORY AND COMPUTATION

Mioara Mandea[1], Carmen Gaina[2,3,4], Vincent Lesur[5]
[1]Université Paris Diderot – Institut de Physique du Globe de Paris, Géophysique spatiale et planétaire, Paris Cedex 13, France
[2]Geodynamics Group, Geological Survey of Norway, Trondheim, Norway
[3]Physics of Geological Processes/School of Geosciences, University of Oslo, Oslo, Norway
[4]Center for Advanced Studies, Norwegian Academy of Sciences and Letters, Oslo, Norway
[5]Helmholtz-Zentrum Potsdam, Deutsches GeoForschungsZentrum – GFZ, Potsdam, Germany

Definition

Magnetic Modeling usually refers to the process of developing models of the Earth's magnetic field based upon measured magnetic data. A model may be as simple as recognizing that an anomaly is likely caused by a buried iron object, or it may involve sophisticated data processing and/or inversion to mathematically build a range of plausible models.

Introduction

When either a ground observatory or a satellite takes a geomagnetic field measurement, it represents the superposition of many sources. The greatest contribution is generated from the dynamo action within the fluid, iron-rich core of the Earth, known as the core field. Although depending upon the altitude and location of the measurement, the data may have sizeable contributions from the static lithospheric field being largely generated from rocks and remnant magnetism within the Earth's crust. Internal in origin are also small magnetic signals produced by the electrical currents associated to the water flow in oceans. On the other side, the external field sources originate in the ionosphere and magnetosphere. These sources include the daily solar quiet variations, the ring current variations, and the many other atmospheric current systems (e.g., the field-aligned currents and the auroral and equatorial electrojets). When modeling Earth's magnetic field one must also take into account the currents induced in the Earth from these external fields. Sporadic magnetic storms and pulses, which have short-term effects, also play a role in the measured magnetic field values.

An important feature of the geomagnetic field is its variability on many different timescales, from less than a second to millions of years (Mandea and Purucker, 2005). The knowledge of the Earth's magnetic field relies on various data sources, ranging from rock magnetization measurements to direct magnetic field measurements, from historic magnetic measurements to recent high-quality data provided by magnetic observatories and the magnetic satellites as MAGSAT, Ørsted, CHAMP, and SAC-C (Mandea et al., 2010a). A combination of ground and continuous satellite measurements allows the core magnetic field and its time variation to be described with a very high resolution in space and in time, on global scale (e.g., (Lesur et al., 2010; Olsen et al., 2010)). On the other side, recent dynamo simulations have also contributed to the understanding of the core field geometry and its dynamics, and temporal changes (Christensen and Wicht, 2007; Wicht et al., 2010).

The geomagnetic field changes in space and time, therefore magnetic observations must continually be acquired, on ground and near-Earth's space. The magnetic observations are performed in different ways and at various platforms. Geomagnetic observatories have provided the highest quality data, however with a very uneven spatial distribution. To increase the spatial density regionally, repeat station surveys are conducted by many teams, but at a national scale, only. Over the last decade, magnetic satellites have offered a uniform, global coverage, but face problems of altitude control and ambient current in the surrounding plasma.

This very brief description of the sources contributing to magnetic observations, suggests how challenging it is to obtain a reliable mathematical description of magnetic field for the modeling process. New mathematical tools need to be developed to accurately describe magnetic data. Depending on the available data and the base functions, those models can be global or regional, static or time dependent. The mathematical description of the different kinds of field models, their physical backgrounds, and applicability are vast subjects, and cannot be covered in detail here. The main aim of this contribution is to give to the reader a basic support and information for more detailed readings.

Global field modeling

Spatial and temporal description

The geomagnetic field is a global phenomenon, so the most obvious approach is to describe it on the whole Earth's surface. The magnetic field, following the Maxwell equations is source-free, so it can be written:

$$\nabla \times \vec{H} = \vec{J} + \frac{\partial \vec{D}}{\partial t} \tag{1}$$

$$\nabla \cdot \vec{B} = 0 \tag{2}$$

where $\vec{H}$ is the magnetic field, $\vec{B}$ is the magnetic induction, $\vec{J}$ is the current density, and $\partial\vec{D}/\partial t$ is the electric displacement current density. In a region without magnetic field sources, for example from the Earth's surface up to about 50 km, it is reasonable to assume that $\vec{J} = 0$ and $\partial\vec{D}/\partial t = 0$, so $\nabla \times \vec{H} = 0$, meaning that the vector field is conservative in the region of interest, and the magnetic field $\vec{H}$ can be expressed as $\vec{H} = -\nabla V$, where V is a scalar potential. Because $\vec{B} = \mu_0\vec{H}$ above the Earth's surface (where $\mu_0 = 4\pi \times 10^{-7}$ H m^{-1}), it follows that $\nabla \cdot \vec{H} = 0$ and that V has to satisfy Laplace's equation:

$$\nabla^2 V(\theta, \phi, r) = 0 \tag{3}$$

We underline that in geomagnetism the magnetic induction $\vec{B}$ is traditionally called "magnetic field." We use the same notation in the following. In spherical coordinates (θ, ϕ, r), which are the geocentric co-latitude, longitude, and radial distance, Laplace's equation takes the form:

$$\frac{1}{r}\frac{\partial^2}{\partial r^2}(rV) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial V}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 V}{\partial\phi^2} = 0 \tag{4}$$

Equation 4 solutions are products of three expressions (the first is to be only a function of r, the second, only a function of θ and the third, only a function of ϕ) and can be solved by separation of variables. A completely general solution is provided by spherical harmonics, and can be written in such a form to express two important contributions in the potential field: a part, V^{int} describing the internal (core and lithospheric) sources, and a part, V^{ext}, describing the external (mainly magnetospheric) sources:

$$\begin{aligned} V &= V^{\text{int}} + V^{\text{ext}} \\ &= a\sum_{n=1}^{N_{\text{int}}}\sum_{m=0}^{n}\left(g_n^m\cos m\phi + h_n^m\sin m\phi\right)\left(\frac{a}{r}\right)^{n+1}P_n^m(\cos\theta) \\ &\quad + a\sum_{n=1}^{N_{\text{ext}}}\sum_{m=0}^{n}\left(q_n^m\cos m\phi + s_n^m\sin m\phi\right)\left(\frac{r}{a}\right)^{n}P_n^m(\cos\theta), \end{aligned} \tag{5}$$

where a = 6371.2 km is a reference radius, (θ, ϕ, r) are geocentric coordinates, P_n^m are the associated Schmidt normalized Legendre functions, N_{int} is the maximum degree and order of the internal potential coefficients g_n^m, h_n^m, and N_{ext} is that of the external potential coefficients q_n^m, s_n^m.

Some of the base functions show special characteristics and are named individually:

- Zonal functions are characterized by $m = 0$ and are independent of longitude.
- Sectorial functions, are defined by $n = m$, so there are no zero crossings in latitude.
- Tesseral functions vary with both latitude and longitude.

The Gauss coefficients can be interpreted in terms of "sources." The first term, g_1^0, is associated with the geocentric dipole oriented along the vertical axis, that is, axis of the Earth's rotation, with dipole moment $g_1^0 4\pi a^3/\mu_0$. The next two terms, g_1^1, h_1^1, characterize the geocentric dipoles oriented along the two last orthogonal axis. The magnitude of the *geocentric dipole* is given by $M = (4\pi a^3/\mu_0)\left(\left(g_1^0\right)^2 + \left(g_1^1\right)^2 + \left(h_1^1\right)^2\right)^{1/2}$, and presently is tilted at roughly 11° from the rotation axis. The potential for a dipole falls off as r^{-2}, and the strength of the field components as r^{-3}. Similarly, for $n = 2$, the terms represent the *geocentric quadrupole* (potential falls off as r^{-3}, and the strength as r^{-4}), for $n = 3$ the terms represent the *geocentric octopole* (potential falls off as r^{-4}, and the strength as r^{-5}), and so forth.

The first three Gauss coefficients $\left(g_0^1, g_1^1, h_1^1\right)$ represent the *dipole* field, while the remaining terms represent the *non-dipole field*. The lowest degree terms correspond to the largest wavelength features of the field, and is dominated by zonal harmonics. The first Gauss coefficient has a cos θ dependence, so at distance r the "wavelength" of the associated magnetic feature is $2\pi r$. Generally speaking, the term g_n^m has a wavelength of $2\pi r/\left(n+\frac{1}{2}\right)$ (Backus et al., 1996).

To describe only the part of the geomagnetic field with sources beneath the Earth's surface, only the part V^{int} of Equation 5 is used. However, this first term of Equation 5 describes both the core and lithospheric fields. It is possible to separate sources originating in the core from those originating in the lithosphere (Lowes, 1966) by computing

the mean square value of the field over the Earth's surface produced by harmonics of a given n:

$$W_n = (n+1)\sum_m \left(\left(g_n^m\right)^2 + \left(h_n^m\right)^2 \right). \tag{6}$$

Variation of W_n depends on degree n, it is known as the power or Lowes–Mauersberger spectrum, and Figure 1 shows its behavior. The steep part of the spectrum ($n \leq 13$) is clear and indicates the signature of the long wavelengths of the core field, the transition degrees ($n = 14$–15) can be attributed to signals from both the core and the lithospheric sources, and the higher degrees ($n \geq 16$) are dominantly lithospheric in origin.

The above description in terms of spherical harmonics is used to characterize static or snapshot fields. Considering the definition of the components X, Y, and Z of the magnetic field (in the northward, eastward, and radially inward directions) and the definition of the spherical polar coordinates (Merrill et al., 1996), it is apparent that $X = -B_\theta$, $Y = B_\phi$ and $Z = -B_r$. For a given epoch, the components of the geomagnetic field are obtained from the model coefficients using the internal part of Equation 5, as derivatives of potential:

$$\begin{aligned} B_\theta &= -\frac{1}{r}\frac{\partial}{\partial\theta}(V^{\text{int}}) \\ &= -\sum_{n=1}^{N_{\text{int}}}\sum_{m=0}^{n}\left(g_n^m\cos m\phi + h_n^m \sin m\phi\right)\left(\frac{a}{r}\right)^{n+2} \\ &\quad \times \frac{\partial}{\partial\theta}P_n^m(\cos\,\theta) \end{aligned} \tag{7}$$

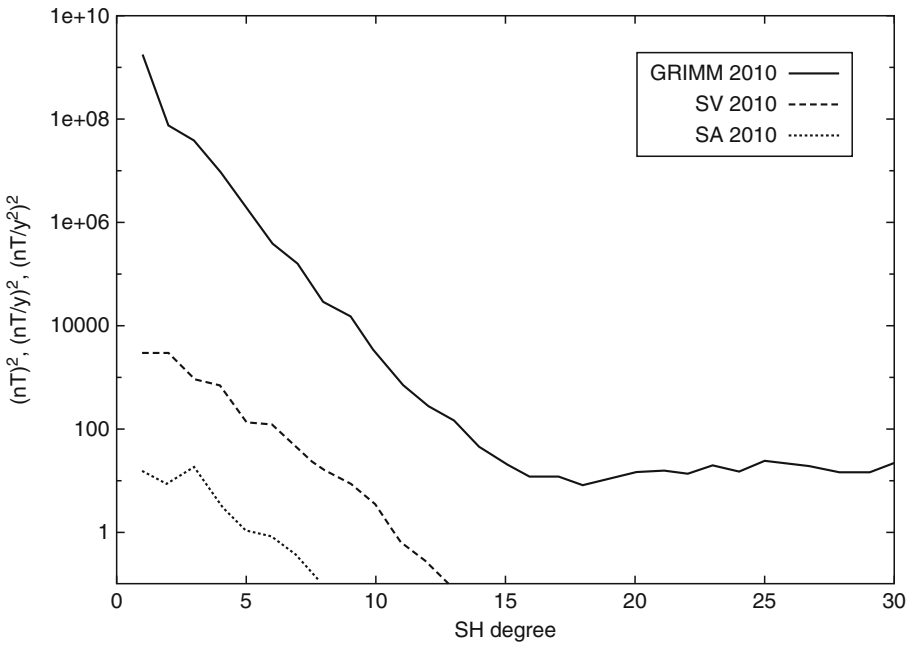
$$\begin{aligned} B_\phi &= -\frac{1}{r\,sin\,\theta}\frac{\partial}{\partial\phi}(V^{\text{int}}) \\ &= \frac{1}{sin\theta}\sum_{n=1}^{N_{\text{int}}}\sum_{m=0}^{n} m\left(g_n^m \sin m\phi - h_n^m\cos m\phi\right)\left(\frac{a}{r}\right)^{n+2} \\ &\quad \times P_n^m(\cos\,\theta) \end{aligned} \tag{8}$$

$$\begin{aligned} B_r &= -\frac{\partial}{\partial r}(V^{\text{int}}) \\ &= -\sum_{n=1}^{N_{\text{int}}}\sum_{m=0}^{n}(n+1)\left(g_n^m\cos m\phi + h_n^m\sin m\phi\right)\left(\frac{a}{r}\right)^{n+2} \\ &\quad \times P_n^m(\cos\,\theta) \end{aligned} \tag{9}$$

Traditionally, the system of coordinates used in geomagnetism consists of the north, east, and center components (sometimes denoted X, Y, Z).

Since the geomagnetic field changes in space and time, the Gauss coefficients are also time-dependent. In order to model time variations of internal origin, the *secular variation*, about 1 year of continuous observations is needed. This variation is often assumed to be constant over short timescales, and can be introduced by adding a secular-variation potential, V_{sv}, truncated to $N_{\text{sv}}^{\max}$:

$$\begin{aligned} V_{\text{sv}} = a\sum_{n=1}^{N_{\text{sv}}^{\max}}\left(\frac{a}{r}\right)^{n+1}\sum_{m=0}^{n}(t-T_0)\left(\dot{g}_n^m\cos(m\phi)\right. \\ \left. + \dot{h}_n^m\sin(m\phi)\right)P_n^m(\cos(\theta)) \end{aligned} \tag{10}$$



Magnetic Modeling, Theory and Computation, Figure 1 The power spectrum obtained from the GRIMM model: the steep part of the spectrum ($n \leq 13$) indicates the core field contribution, the transition degrees (n = 14–15) can be attributed to both core and lithospheric contributions, and the higher degrees ($n \geq 16$) describe the lithospheric contribution. Units are in nT^2.

where $\dot{g}_n^m$, $\dot{h}_n^m$ denote the time derivative of the internal Gauss coefficients, T_0 the reference time (i.e., the epoch of the core field model), and t is the considered time.

The time dependence of the core field coefficients $\left\{g_n^{m(t)}, h_n^m(t)\right\}$ up to a given degree n can also be described by temporally more complex functions, for example, B-splines. Generally, in the geomagnetic field modeling process, the cubic (i.e., order 4) B-splines (Bloxham and Gubbins, 1985; Jackson et al., 2000; Olsen et al., 2006; Wardinski et al., 2008), or higher order B-splines (Sabaka et al., 2004; Lesur et al., 2008; Olsen et al., 2009) are used. Splines of order 4 result in a piecewise linear representation of $\partial_t^2 \vec{B}$, which is not favorable for studying rapid core field changes and geomagnetic jerks. Higher order splines result in a smoother representation of $\partial_t^2 \vec{B}$ and are therefore preferable. In order to be able to describe rapid field temporal variations of the core field, a 6-month knot separation is recommended. Recently, due to a decade of high-resolution, high-precision satellite magnetic data, the secular acceleration can also be described, by coefficients $\ddot{g}_n^m$, $\ddot{h}_n^m$, denoting the second time derivative of the Gauss coefficients.

Inverse problem and implementation

To obtain a geomagnetic model and thereafter the field component description, an inverse problem needs to be solved, for which the standard model estimation technique is summarized below.

The aim is to describe a given dataset in terms of a series of parameters. The data can be assembled in a data vector $\mathbf{d}$ of length N

$$\mathbf{d}^{\mathbf{T}} = (\mathbf{d_1}, \mathbf{d_2}, \ldots, \mathbf{d_N}). \tag{11}$$

The model is defined by a set of M parameters assembled in a model vector of length M

$$\mathbf{m}^{\mathbf{T}} = (\mathbf{m_1}, \mathbf{m_2}, \ldots, \mathbf{m_M}). \tag{12}$$

The basic assumption is that the data are related to the model parameters through a function $\mathbf{F}(\mathbf{m})$, leading to the relation:

$$\mathbf{d} = \mathbf{F}(\mathbf{m}), \tag{13}$$

which, for linear (or linearized) problems, reduces in a matrix form:

$$\mathbf{d} = \mathbf{A}\mathbf{m} \tag{14}$$

where $\mathbf{A}$ is an $N \times M$ matrix of coefficients, independent of data values and model parameters. In the case of spherical harmonics analysis, the data often consist of three components of the magnetic field vector. The Gauss coefficients define the parameter vector $\mathbf{m}$. The $\mathbf{A}$ matrix elements contains M base functions evaluated at N data positions – that is, spherical harmonics (or their gradients) and their radial dependences.

Generally, more data are available than parameters to be resolved, so $N \geq M$. In this case, the above equation cannot be satisfied for every data, the inverse problem is overdetermined and has no exact solution. Hence, model parameters are sought such that the above equation is solved approximately. The least square method minimizes the sum of the squares of the errors in order to find a solution. The usual solution to this problem is given by:

$$\mathbf{m} = \left(\mathbf{A}^{\mathbf{T}}\mathbf{W}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathbf{T}}\mathbf{W}\mathbf{d} \tag{15}$$

as long as the inverse matrix $\mathbf{G}^{-1} = \left(\mathbf{A}^{\mathbf{T}}\mathbf{W}\mathbf{A}\right)^{-1}$ exists. To define the weight matrix $\mathbf{W}$, it is observed that the magnetic data from different platforms, and at different times, can be affected by different error sources and these error sources can even be correlated. The covariance matrix $\mathbf{C_d}$ associated with these errors must be estimated, inverted, and then used as weight matrix, that is, $\mathbf{W} = \mathbf{C_d^{-1}}$. For uncorrelated data, the covariance matrix is diagonal. In practice, this is rarely the case, but for computational simplicity diagonal covariance matrices are often used.

Inverse problems are sometimes ill-posed – that is, some parameters or combination of parameters are not resolved by the data. In that case, the inverse of $\mathbf{G} = \left(\mathbf{A}^{\mathbf{T}}\mathbf{W}\mathbf{A}\right)$ does not exist, or is numerically difficult to estimate. It results that small errors in data values can lead to considerable changes in the parameters. The error amplification must be controlled by appropriate measures in the solution process: inverse problems can be regularized by introducing regularization matrices in space and time. A solution in the least square sense writes now:

$$\mathbf{m} = \left(\mathbf{A}^{\mathbf{T}}\mathbf{W}\mathbf{A} + \lambda_t\mathbf{\Lambda_t} + \lambda_s\mathbf{\Lambda_s}\right)^{-1}\mathbf{A}^{\mathbf{T}}\mathbf{W}\mathbf{d}. \tag{16}$$

It corresponds to a minimum of the functional:

$$\Phi = (\mathbf{d} - \mathbf{A}\mathbf{m})^{\mathbf{T}}\mathbf{W}(\mathbf{d} - \mathbf{A}\mathbf{m}) + \lambda_t\mathbf{m}^{\mathbf{T}}\mathbf{\Lambda_t}\mathbf{m} + \lambda_s\mathbf{m}^{\mathbf{T}}\mathbf{\Lambda_s}\mathbf{m} \tag{17}$$

The choice of λ_t and λ_s damping parameter values is a difficult, and often computationally intensive task. If criteria exist to determine these parameters, they are often unpractical to establish, and therefore the parameters choice often depends on the modeler skills, interests, and a priori knowledge on the solution. The same formalism is used to impose constraints on the model solution.

The models developed are more and more complex due to the high quality of the recent satellite data, and to the overall progress in data acquisition techniques and quality. This combined with the developments of new mathematical tools, and better insight on the physical process underlying the generation of the magnetic field, has led to the developments of complex codes that use up-to-date numerical techniques.

From a computational view point, when large datasets are used, the most time-consuming step in building a model is to compute the matrix $\mathbf{G} = (\mathbf{A}^{\mathbf{T}}\mathbf{W}\mathbf{A})$. Typically, a parallel computing environment is necessary to compute this matrix in a reasonable amount of time, that is, less

than a few days. Often, a re-weighted least-square approach is needed to reach the best solution, requiring the evaluation of **G** numerous times. In comparison, inverting the matrix takes only few minutes for models with 10–20 thousands of parameters. Overall, several months are needed by a trained modeler to derive a robust model of the Earth's interior magnetic field.

Potpourri of recent global models

An important prerequisite for deriving more precise and higher resolution models of the geomagnetic field is the availability of high-quality measurements. Under the assumption of suitable data, the modeling task may be separated in a few crucial steps. The first one is linked to data selection, as in each single measurement all contributions exist, and the wish is to minimize the external ones. In the second, the inversion of magnetic measurements (as shown before) to obtain a mathematical description of the field distribution is foreseen, and finally a control of the final model is needed.

Here, some of the most recent models are listed, and the interested reader can find more information on the specified web sources. Firstly, models based on the last decade of satellite data are indicated, thereafter models covering a longer period of time.

The International Geomagnetic Reference Field (www.ngdc.noaa.gov/IAGA/vmod/IGRF-10) IGRF is an internationally agreed series of global spherical harmonic models of the Earth's core magnetic field, and it is provided by the IAGA Division V-MOD. The Gauss coefficients are considered linearly time dependent for 5-year intervals (Finlay et al., 2010).

The GRIMM series (GFZ Reference Internal Magnetic Model) (www.gfz-potsdam.de/portal/gfz/Struktur/Departments/Department+2/sec23/topics/models/GRIMM2) by Lesur et al. (2008, 2010) are based on CHAMP satellite data and observatory hourly means. GRIMM-2 covers the period 2001–2010 and aims at describing the very rapid variations of the core field. One of its special characteristic is the use of full vector satellite data at high latitudes at all local times, for a better separation into the contributions generated by the ionosphere and field aligned currents on one hand, and, contributions generated by the Earth's core and lithosphere, on the other hand.

The CHAOS models (www.space.dtu.dk/English/Research/Scientific_data_and_models/Magnetic_Field_Models.aspx) derived by Olsen et al. (2006, 2009) as well as Olsen and Mandea (2008) are based on satellite and observatory data, with an interest to the subdecadal timescales variations. As GRIMM series, the advantage of these series of models is the fact that they are able to describe the field itself, but also its secular variation and acceleration.

The MF series (www.geomag.us/models/MF6.html), with the last version published by Maus et al. (2008), is a model built to mainly describe the global lithospheric field. The last published version, MF6 estimates the lithospheric magnetic field to spherical harmonic degree 120, which corresponds to 333 km wavelength resolution. The very last version MF7 is now available.

Some geomagnetic field models cover longer period of time, from decades to centuries, as:

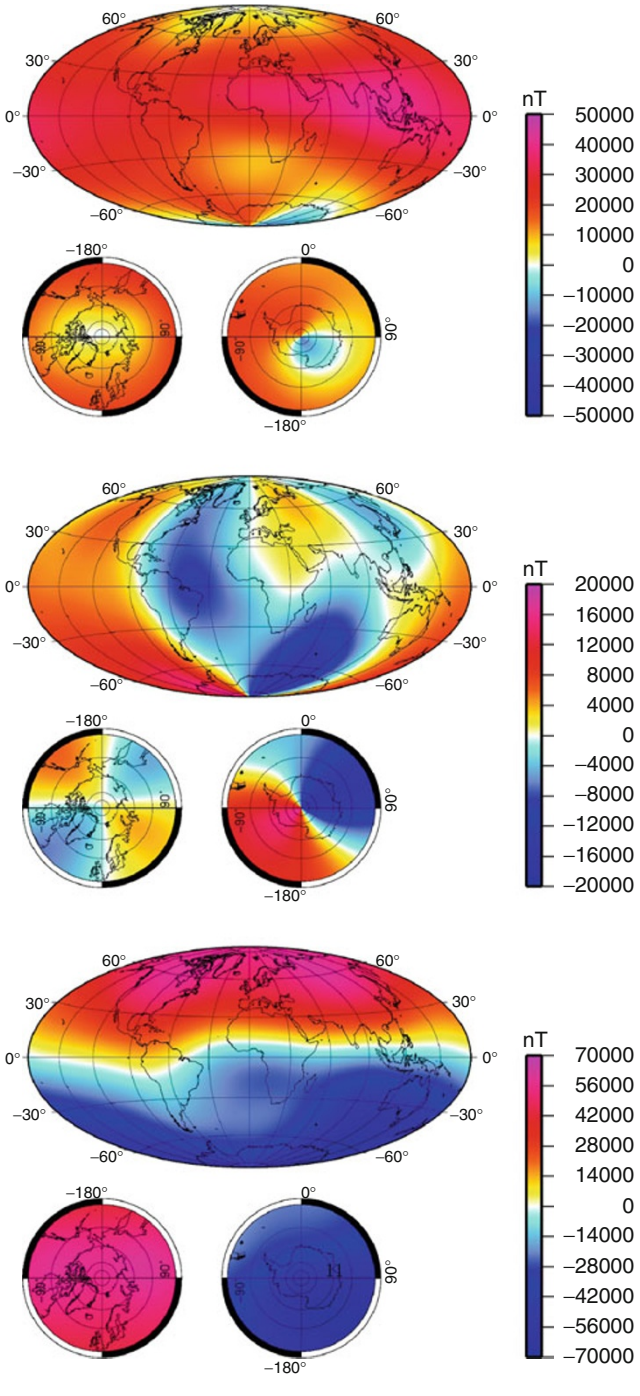
The CM4 model (core2.gsfc.nasa.gov/cm/), as the fourth version of the Comprehensive Model, has been developed by Sabaka et al. (2004). This model relies on POGO, MAGSAT, Ørstedt, and CHAMP satellite data and observatory hourly means and is continuous through the time span from 1960 to 2002. The spherical harmonic expansion goes to degree and order 65 with the secular variation represented by cubic B-splines with 2.5-years knot spacing through degree and order 13.

The C^3 FM (www.gfz-potsdam.de/portal/gfz/Struktur/Departments/Department+2/sec23/topics/models/C3FM), as Continuous Covariant Constrained-end-points Field Model Wardinski and Holme (2006), is a time-dependent model, between 1980 and 2000 with Gauss coefficients expanded in time on a basis of cubic B-splines. This model is constrained to fit field models from two satellites (MAGSAT in 1980 and Ørsted in 2000) and to fit both magnetic observatory and repeat station secular variation for the period in between. The model reveals a short-term secular variation on subdecadal timescale with a higher spatial resolution than previously resolved.

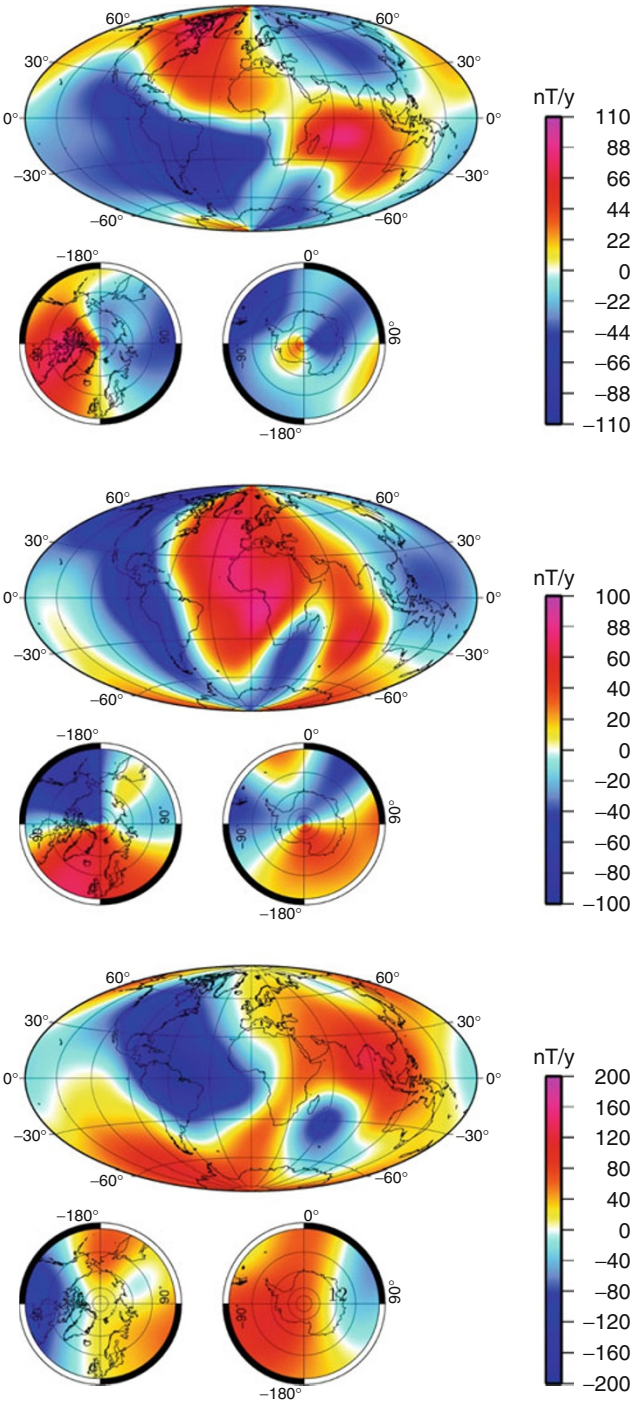
The gufm (www.epm.geophys.ethz.ch/cfinlay/gufm1.html) model of Jackson et al. (2000) provides a complete picture of the evolution of the core magnetic field, from 1590 to 1990. This time-dependent model is based on a massive compilation of historical observations of the magnetic field, originating from observations taken by mariners engaged in merchant and naval shipping, as well as survey data, observatory data (which runs back to the mid-nineteenth century), and data from the POGO and MAGSAT satellites.

The CALSK series (www.gfz-potsdam.de/portal/gfz/Struktur/Departments/Department+2/sec23/topics/models/CALS7K) are based on archeo- and paleomagnetic data and give a global description of field evolution further back in the past, although with lower spatial and temporal resolution. Continuous spherical harmonic models on millennial timescales give estimates of spatial field structure up to spherical harmonic degrees around 4 and temporal variations in the order of 100 years. The longest continuous and regularized spherical harmonic model to date goes back 7 kyrs (Korte and Constable, 2005).

Some models together with online calculators or source codes are available. The interested reader can find on the German Research Centre for Geosciences web site (www.gfz-potsdam.de/portal/gfz/Struktur/Departments/Department+2/sec23/topics/models), the British Geological Survey web site (www.geomag.bgs.ac.uk/gifs/igrf_form.shtml), or the National Oceanic and Atmospheric Administration web site (www.geomag.org) good overviews on a number of field models together with user-friendly calculators. Figures 2 and 3 show maps of the



Magnetic Modeling, Theory and Computation, Figure 2 Maps of the core field components (Units in nT): northward component (*top*), eastward component (*middle*), and radially inward direction (*bottom*) at the Earth's surface in 2010 from the field model GRIMM-2. Molleweide projection is used.



Magnetic Modeling, Theory and Computation, Figure 3 Maps of the core field secular variation (Units in nT/year): northward component (*top*), eastward component (*middle*), and radially inward direction (*bottom*) at the Earth's surface in 2010 from the field model GRIMM-2. Molleweide projection is used.

core field components and their secular variation for the northward, eastward, and radially inward direction at the Earth's surface in 2010 from the field model GRIMM-2 (Lesur et al., 2010).

Regional field modeling

To describe the magnetic field at regional scales, different methods have been proposed during the last decades,

starting with the Spherical Cap Harmonic Analysis proposed by Haines (1985) and the related translated origin spherical cap harmonic analysis (De Santis, 1991; Korte and Holme, 2003). The spherical cap harmonic analysis permits the use of data from only a region of the Earth while satisfying the constraints of potential field theory. Indeed, the region to be studied has to be defined by a spherical cap and satisfies the zero curl and divergence conditions. The method is claimed to be valid over any spherical cap at any altitude above the Earth's surface. On the basis of these assertions, the method has been widely used for constructing regional magnetic maps (De Santis et al., 1997; Korte and Haak, 2000). However, difficulties occur when using this approach. The main one appears as a failure to correctly model the radial dependence (i.e., that the models cannot be continued and data acquired at different altitudes cannot be simultaneously inverted). In order to solve this difficulty, Thébault et al. (2004) have re-posed this method as a boundary value problem within a cone extending above the reference surface, thereby allowing satellite data to be downward continued to the Earth's surface.

Some different basis functions, such as harmonic splines Shure et al. (1982), or with more local support, wavelet-like functions (Holschneider et al., 2003; Lesur, 2006), are better suited to describe datasets with variable resolution over the area of interest. Only these two approaches are summarized in the following. Schott and Thébault (2010) provide a detailed overview of various regional modeling approaches, and the interested reader is advised into this paper.

Harmonic splines

A detailed description of harmonic splines can be found in Shure et al. (1982), and a very recent revisited approach with an application for the Southern African continent magnetic data in Geese et al. (2010). The main interest of this representation is that it offers a smoothness property to the derived field model.

The method considers the magnetic field as expressed as the gradient of a scalar potential, as shown in the previous section. Using the spherical harmonics Y_n^m of degree n and order m, the internal part of the global magnetic field can be written as

$$B = -\nabla a \sum_{n,m}^{N} g_n^m \left(\frac{a}{r}\right)^{n+1} Y_n^m(\theta,\phi) \tag{18}$$

with $Y_n^m(\theta,\phi)$ being the Schmidt normalized spherical harmonics. An alternative representation of the field is given by:

$$B(\theta,\phi,r) = -\sum_i \alpha_i \nabla F_i^{Nr}(\theta,\phi,r) \tag{19}$$

$$F_i^{Nr} = a \sum_{n,m}^{N} f_n (n+1) \left(\frac{a}{r}\right)^{n+1} \left(\frac{a}{r_i}\right)^{n+2} \times Y_n^m(\theta_i,\phi_i) Y_n^m(\theta,\phi) \tag{20}$$

$$f_n = \frac{2n+1}{4\pi n^2 (n+1)^4} \tag{21}$$

With this choice of parameter f_n values, the modeled magnetic field minimizes the quantity:

$$\int_\Omega \left|\nabla_h^2 B_r\right|^2 \mathrm{d}\omega \tag{22}$$

In the framework of regional analysis, in order to insure a fit to the data, it is necessary to introduce for the representation of the magnetic field two further sets of functions similar to Equation 20. This gives:

$$B(\theta,\phi,r) = -\nabla\left(\sum_i \alpha_i^r F_i^{Nr}(\theta,\phi,r) + \sum_i \alpha_i^\theta F_i^{N\theta}(\theta,\phi,r) + \sum_i \alpha_i^\phi F_i^{N\phi}(\theta,\phi,r)\right) \tag{23}$$

$$F_i^{N\theta} = a \sum_{n,m}^{N} f_n \left(\frac{a}{r}\right)^{n+1} \left(\frac{a}{r_i}\right)^{n+2} \times \frac{\partial}{\partial\theta} Y_n^m(\theta_i,\phi_i) Y_n^m(\theta,\phi)$$

$$F_i^{N\phi} = a \sum_{n,m}^{N} f_n \left(\frac{a}{r}\right)^{n+1} \left(\frac{a}{r_i}\right)^{n+2} \frac{1}{\sin\theta} \times \frac{\partial}{\partial\phi} Y_n^m(\theta_i,\phi_i) Y_n^m(\theta,\phi). \tag{24}$$

This model is equivalent to the harmonic spline model presented in Geese et al. (2010) and Shure et al. (1982).

An example of a recent application of this approach is found in Geese et al. (2010). The investigated area is the Southern African continent where the geomagnetic field is weaker and changes more rapidly compared to other regions of the Earth. The regional dataset, consisting of repeat station surveys and observatory annual means, has been used in the described system of representation. The regional field model covers the time span from 1960 to 2002 and gives a detailed view of the changes of the geomagnetic field in this region.

Wavelet analysis

The wavelet frames and the recent work by Holschneider et al. (2003); Freeden and Michel (2004); Chambodut et al. (2005) have brought the use of wavelet frames for constructing regional potential field models. In particular on the sphere, the used functions should satisfy the following properties: (1) harmonic prolongation must be easily computable; (2) function itself has to be numerically easy to compute; (3) functions must be localized on the sphere.

The indicated requirements lead to constructions based on the Poisson kernel of spherical functions. A Poisson wavelet at a given point on the unit sphere is expressed as:

$$\psi_a^k(\vec{x}) = N_a^k \sum_n (an)^k e^{-an} Q_n(\cos\ \mu) \tag{25}$$

where N_a^k is a normalization constant and μ is the angle between the wavelet position and the data point position. The Q_n function is defined by:

$$Q_n(\cos\ \mu) = (2n+1)P_n(\cos\ \mu), \tag{26}$$

that can be expressed easily in terms of products of spherical harmonics.

The wavelet is defined by three parameters: scale a, position, and order k. The scale a is a measure of the wavelength of the wavelet without dimension, and order k characterizes the global shape of the wavelets (number of oscillations). The wavelets on the sphere may also be considered as a restriction on the sphere of a function that is harmonic outside and that has singularities inside it. The main characteristic of a wavelet is its relatively well-localized amplitude on the sphere, hence well adapted to regional modeling, while spanning a relatively limited range of wavelengths.

Recently, a few implementations of the wavelet representation for potential fields have been done. For example, Panet et al. (2006) have applied the wavelet method to improve a global, but low-resolution gravity field model, by using regional high-resolution data. Over the same area of interest, French Polynesia, a lithospheric magnetic field model based on CHAMP satellite data has been computed, applying the same approach. The regional field models allow authors to suggest that the origin of the volcanism in this region may be controlled by lithospheric weakness due to the regional stress field.

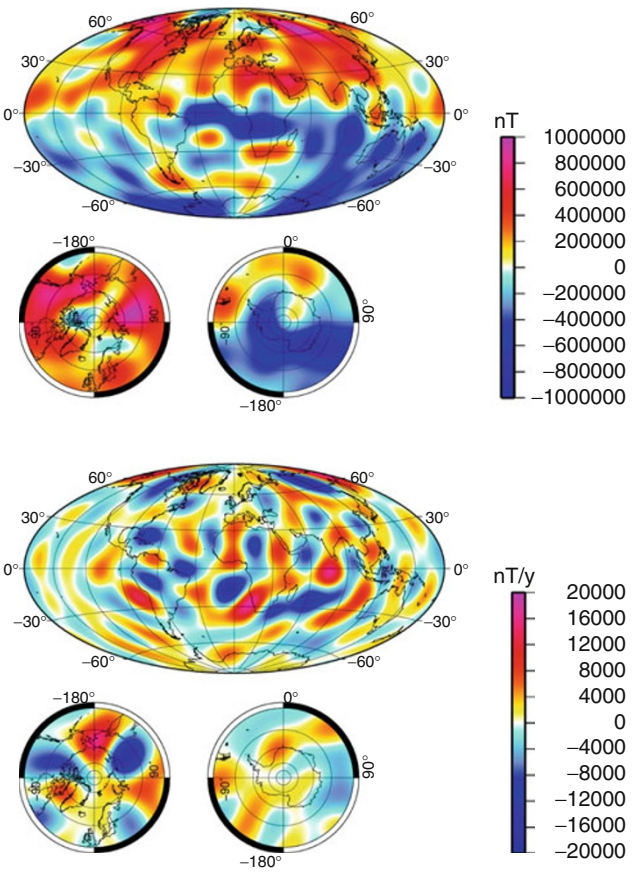
Interpretation

The geomagnetic field models, at global and regional scales, play an important role in our modern world. The Earth's magnetic fields can be used as a tool in screening our planet and understanding it deep interior structure and processes. Moreover, these models can be used for some specific applications, as in navigation, petroleum industry, in technological systems submitted to large geomagnetically induced currents, in satellite operations, or exploration geophysics. From this list, which is far to be complete, two examples are given, one linked to interpretation of the magnetic field and its secular variation at the core mantle boundary, and one linked to the lithospheric field.

Core magnetic field at core–mantle boundary

As described above, models of the magnetic field generated inside the Earth are dominated for the long wavelengths by the core field, and for the shorter wavelengths by the lithospheric field. The longest wavelengths of these models – roughly up to spherical harmonics of degree 14 – can therefore be used to try to understand the generation process of the core field. Results from seismological observations have shown that the core can be separated in two regions; a solid inner core of radius estimated at 1,220 km and a liquid outer core in the spherical shell extending up to 3,485 km. This liquid outer core is very conductive and made mainly of liquid iron in convection, which in the presence of a magnetic field, can regenerate this field, balancing its natural decay due to ohmic dissipation. This process is known as the geodynamo. The best way to retrieve information about the dynamics of the core, and the fluid motions at the top of the core is through magnetic data acquired at, or above, the Earth's surface. The magnetic field models are therefore downward continued, through the mantle, to the core mantle boundary (CMB) assuming that the mantle is an insulator. In principle, only the radial component of the magnetic field can be derived just below the CMB, because of the unknown jump of the electrical conductivity across the CMB, associated with a discontinuity of the magnetic field components tangential to the boundary.

An image of this vertical down component of the core magnetic field just below the CMB is shown in Figure 4, together with its expected time derivative. These images



Magnetic Modeling, Theory and Computation, Figure 4 Map of the core field vertical down component (Units in nT) and its secular variation (Units in nT/year) at the core–mantle boundary, for year 2010 and from the field model GRIMM-2. Molleweide projection is used.

correspond to a magnetic field model truncated to spherical harmonic degree 13. Contributions from shorter wavelengths may considerably modify these images, but these contributions are unknown, being hidden by the lithospheric field or the noise, at the Earth's surface. However, considerable information can be readily extracted from these maps. One can see that the core field has a much more complex structure at the CMB than on the Earth's surface (Figure 2). In particular, the presence in the southern hemisphere of a large patch of reverse orientation of the magnetic field – the field there is oriented toward the core interior, whereas in a simple dipole approximation one would expect the field to be oriented outward. This patch and its evolution has attracted scientists' attention over the last decades because it is associated with the weakness of the Earth's field in the South-Atlantic region, it contributes significantly to the decrease of the strength of the Earth's dipole field component, and it has sometimes been thought as an early stage of a reversal process (Hulot et al., 2002; Olsen and Mandea, 2008) regarding the secular variation, it is obviously concentrated at midlatitudes and high latitudes of the northern hemisphere. It is extremely weak under the Pacific ocean and southern part of the southern hemisphere.

The combination of the Maxwell equations leads to the so-called induction equation. Under the frozen-flux approximation, which neglects the diffusion in front of induction (Roberts and Scott, 1965), the radial component of this equation writes:

$$\frac{\partial}{\partial t} B_r = -\nabla_h \cdot (B_r U_h) \tag{27}$$

where B_r is the radial component of the magnetic field, U_h is the horizontal component of the flow under the CMB, and ∇_h is the tangential gradient on the sphere. This equation simply shows that the main source of secular variation is the advection of the field lines by the flow. Since radial component of both the field and its secular variation can be estimated at the CMB, this equation provides a way to get some information on the flow at the top of the liquid outer core, for large spatial wavelength and relatively short timescales. By combining this equations with constraints derived from the equations of fluid dynamics and/or magnetohydrodynamics, it is possible to derive physical properties of the outer core and shed light on unexplained phenomenon as geomagnetic jerks (Mandea et al., 2010b) or contributions to the variation of the length of the day.

Lithospheric magnetic field over the Arctic region

The magnetic signature of the lithosphere is related to all the processes that affected and shaped the figure of our planet. To first order, the remanent magnetic contributions are canceled out at large wavelength and the induced part dominates above continental domains. The exact contribution of the remanent magnetization is actually difficult to estimate, unless the physical properties of the magnetic sources are known through rock sampling and analysis. It is possible to estimate what is their relative contribution by using global forward approaches, mostly relying on the distribution of the vertically integrated magnetization. A simple estimate is based on the lateral magnetization contrast associated with the boundary between oceanic and continental domains. Indeed, the abrupt boundaries between the thin oceanic crust and the thicker continental crust translates into a large-scale magnetic signal which sometimes occur in the proximity of continental margins, and also above continental and oceanic basins. This does not suffice to explain all the observed magnetic anomalies. Those interested in the crustal magnetic field are referred to some recent reviews, like Langlais et al. (2009).

Generally speaking, the lithospheric magnetic field reflects the magnetization associated with minerals cooling down below their Curie temperature in the presence of a magnetic field. Above the Curie temperature, the magnetization vanishes, and minerals exhibit then paramagnetic susceptibility, with little effect compared to magnetization. Therefore, rocks are essentially nonmagnetic at temperatures greater than the Curie temperature characterizing their main magnetic minerals constituents. Taking into account a normal geothermal gradient, sources of the measured magnetic signal are then restricted to some maximum 30–40 km depth, excepting the old cratonic areas where the Curie depth – depth at which the temperature reach the Curie temperature – may be deeper. Methods to estimate depth to the bottom of magnetic sources from magnetic data are based on the spectral analysis of magnetic anomalies (Spector and Grant, 1970; Blakely, 1995; Bouligand et al., 2009).

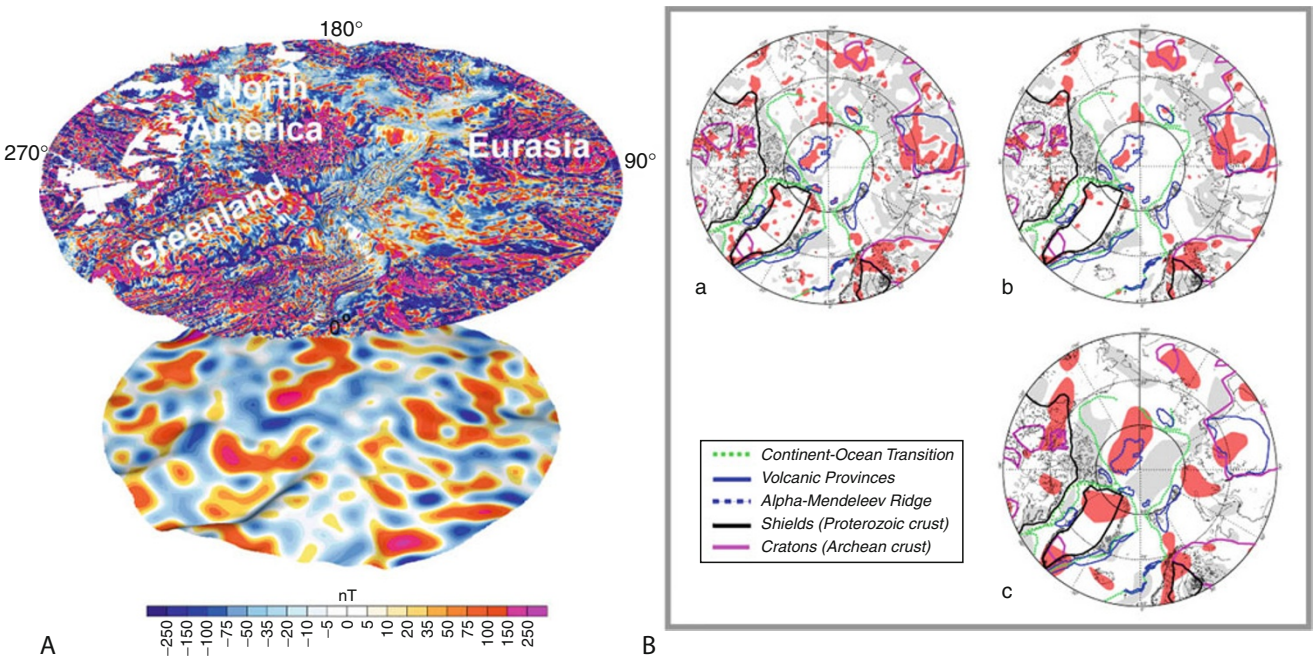
It is important to note here that another source for the deep crustal magnetism has been suggested in the last decade. This is the magnetism resulted from chemical interfaces, or lamellar magnetism. This concept shifted the paradigm of having magnetic rocks only above the Curie temperature (580) depth and can be the source for old, deep-seated crust on Earth and other planets (Robinson et al., 2002).

For a comprehensive example of how the most prominent magnetic anomalies observed in the global compilations of ground and satellite magnetic data (like the World Digital Magnetic Anomaly Map WDMAM, Korhonen et al., 2007) can be analyzed, see Mandea and Thébault (2007). In this contribution we present an example of regional magnetic anomaly modeling and the suggested interpretation, over the Arctic area.

In the last decade, the Arctic region became a frontier for scientific discoveries and exploration studies. All circum-Arctic nations embarked on a quest for collecting new data and deciphering the structure and evolution of the Arctic crust and lithosphere. In the same time, due to the inaccessibility and harsh climate of this area, satellite data plays a crucial role for the Arctic science. Along with bathymetric, gravimetric, and seismological studies, acquisition, processing, and interpretation of magnetic data are vital components for the understanding of

lithospheric architecture. The last couple of years have been favorable not only for data acquisition at high latitudes (due to low solar activity), but also for concerted efforts to gather and assemble regional datasets. Besides the World Digital Magnetic Anomaly Map (Korhonen et al., 2007) and similar global compilations (Hamoudi et al., 2007; Hemant et al., 2007; Maus et al., 2008), a higher resolution map for the Arctic region (upward 60° North) is currently available to the scientific community (Gaina and the CAMP-GM group, 2011) and, together with a satellite lithospheric model (Maus et al., 2008), is used here as a brief example into the Arctic magnetic anomaly interpretation.

Figure 5A shows the difference between magnetic anomalies recorded closer to Earth's surface, which display the magnetic signature of bodies in the Earth's crust, and magnetic anomalies measured from satellites (in this case at about 350 km), which image thicker than normal and/or deeply buried magnetic bodies. By upward continuing the magnetic data collected closer to the surface (either shipborne or airborne), a first-order estimation of shallow versus deep-magnetized bodies can be made. Figure 5B shows the higher and lower magnetic anomalies obtained by upward continuing the magnetic gridded data of the Arctic (CAMP-GM without using the satellite data) at 60 and 100 km, that roughly correspond to the lower continental crust and upper mantle. In addition, the MF6 lithospheric model (Maus et al., 2008) is shown at 350 km altitude. The magnetic signature of the boundary between the continental and oceanic crust cannot be detected anymore in the upward continued data or from satellite altitude, even for volcanic margins (note the volcanic areas outlined by blue lines along the eastern coast of Greenland and western coast of Eurasia). The oldest continental crust, usually found in the cratonic areas and as Proterozoic accreted crust, generates the largest positive magnetic anomalies. This crust contains large and deep volcanic bodies in the North American shield, Greenland, the Baltic shield in Eurasia, and the Siberian platform in NE Asia, which are imaged by the satellite data (Figure 5B). One of the most striking feature in the Arctic domain is the strong magnetic anomaly close to the north pole that correlates with a large, igneous oceanic plateau called the Alpha Mendeleev ridge. The age and exact origin of this Large Igneous Province is not known, although numerous studies speculate on an episode of Cretaceous



Magnetic Modeling, Theory and Computation, Figure 5 (**A**) 3D image of the Circum-Arctic grid of magnetic anomalies compiled from aeromagnetic datasets (Gaina and the CAMP-GM group, 2011) draped on bathymetry and topography of the Arctic (IBCAO, (Jakobsson et al., 2008)). Lower panel shows the lithospheric magnetic model based on satellite data (MF6, (Maus et al., 2008)) downward continued at the geoid height. The MF6 grid is draped on the Bouguer anomaly of the Arctic that has been upward continued in such a way that the major relief corresponds to areas of thick crust (mainly under continental, cratonic area or large volcanic provinces) or thin crust (under the oceanic basins or highly extended continental crust). (**B**) Minimum (gray) and maximum (red) amplitude of magnetic anomalies of the Arctic (CAMP-GM) upward continued at 60 and 100 km (**a** and **b**, respectively) and magnetic anomalies from satellite data at 350 km altitude. Lower map shows major boundaries between different lithospheric domains that may contribute to magnetic anomalies. Regions of thick, old Archean crust and adjacent accretionary crust are characterized by regional metamorphism and plutonic intrusions that have strong magnetic anomaly signatures.

volcanism that affected an old oceanic crust or a partially continental crust. The intensity and extent of the magnetic anomalies recorded by aircrafts or satellites point toward a very thick, volcanic crust, but, as in the case of other oceanic Large Igneous Provinces, only deep sea drilling will be able to reveal the true nature of the underlying crust at the core of the Arctic.

Summary

To model and interpret the Earth's magnetic field, two ingredients are crucial: global high accuracy, high-resolution data and reliable mathematical tools. Starting with the next years, a trio of satellites, Swarm mission, will be launched for a unique view inside the Earth. This mission will provide the most detailed data yet on the geomagnetic field of the Earth and its temporal evolution, giving new insights into improving our knowledge of the Earth's interior and climate. The Swarm constellation satellites will be placed in three different polar orbits, two flying side by side at an altitude of 450 km and a third at an altitude of 530 km. These measurements, together with those provided by magnetic observatories, shipborne or airborne instruments, are valuable data essential for modeling the geomagnetic field and its interaction with other physical aspects of the Earth's system.

Bibliography

Backus, G., Parker, R., and Constable, C., 1996. *Foundations of Geomagnetism*. New York: Cambridge University Press.

Blakely, R., 1995. *Potential Theory in Gravity and Magnetic Applications*. Cambridge: Cambridge University Press.

Bloxham, J., and Gubbins, D., 1985. The secular variation of Earth's magnetic field. *Nature*, **317**, 777–781.

Bouligand, C., Glen, J. M. G., and Blakel, R. J., 2009. Mapping Curie temperature depth in the western United States with a fractal model for crustal magnetization. *Journal of Geophysical Research*, **114**, B11104, doi:10.1029/2009JB006494.

Chambodut, A., Panet, I., Mandea, M., Diament, M., Holschneider, M., and Jamet, O., 2005. Wavelet frames: an alternative to spherical harmonic representation of potential fields. *Geophysical Journal International*, **163**, 875–899, doi:10.1111/j.1365-246X.2005.02754.x.

Christensen, U., and Wicht, J., 2007. Numerical dynamo simulations. In Olson, P. (ed.), *Treatise on Geophysics: Core Dynamics*. Oxford: Elsevier.

De Santis, A., 1991. Translated origin spherical cap harmonic analysis. *Geophysical Journal International*, **106**, 253–263.

De Santis, A., Falcone, C., and Torta, J. M., 1997. SHA vs SCHA for modelling secular variation in a small region such as Italy. *Journal of Geomagnetism and Geoelectricity*, **49**, 359–371.

Finlay, C. C., Maus, S., Beggan, C. D., Bondar, T. N., Chambodut, A., Chernova, T. A., Chulliat, A., Golovkov, V. P., Hamilton, B., Hamoudi, M., Holme, R., Hulot, G., Kuang, W., Langlais, B., Lesur, V., Lowes, F. J., Lühr, H., Macmillan, S., Mandea, M., McLean, S., Manoj, C., Menvielle, M., Michaelis, I., Olsen, N., Rauberg, J., Rother, M., Sabaka, T. J., Tangborn, A., Tøffner-Clausen, L., Thébault, E., Thomson, A. W. P., Wardinski, I., Wei, Z., and Zvereva, T. I., 2010. International geomagnetic reference field: the eleventh generation. *Geophysical Journal International*, **183**(3), 1216–1230.

Freeden, W., and Michel, V., 2004. *Multiscale Potential Theory: With Application to the Geoscience*. Basel, Berlin, Boston: Birkhäuser.

Gaina, C., and the CAMP-GM group, 2011. Circum-Arctic mapping project: new magnetic and gravity anomaly maps of the Arctic. In Spencer, A., Embry, A., Gautier, D., Stoupakova, A. & Sørensen, K. (eds.), *Arctic Petroleum Geology*. Memoir of the Geological Society of London No. 35, (in press).

Geese, A., Mandea, M., Lesur, V., and Hayn, M., 2010. Regional modelling of the southern African geomagnetic field using harmonic splines. *Geophysical Journal International*, **181**(3), 1329–1342, doi:10.1111/j.1365-246X.2010.04575.x.

Haines, G. V., 1985. Spherical cap harmonic analysis. *Journal of Geophysical Research*, **90**, 2583–2591.

Hamoudi, M., Thebault, E., Lesur, V., and Mandea, M., 2007. GeoForschungsZentrum Anomaly Magnetic MAp (GAMMA): A candidate model for the world digital magnetic anomaly Map. *Geochemistry, Geophysics, Geosystems*, **8**, Q06023, doi:10-10292007001638.

Hemant, K., Thébault, E., Mandea, M., Ravat, D., and Maus, S., 2007. Magnetic anomaly map of the world: merging satellite, airborne, marine and ground-based magnetic data sets. *Earth and Planetary Science Letters*, **260**, 10–1016200705040.

Holschneider, M., Chambodut, A., and Mandea, M., 2003. From global to regional analysis of the magnetic field on the sphere using wavelet frames. *Physics of the Earth and Planetary Interiors*, **135**(2), 107–124.

Hulot, G., Eymin, C., Langlais, B., Mandea, M., and Olsen, N., 2002. Small-scale structure of the geodynamo inferred from Oersted and MAGSAT satellite data. *Nature*, **416**, 620–623.

Jackson, A., Jonkers, A. R. T., and Walker, M. R., 2000. Four centuries of geomagnetic secular variation from historical records. *Philosophical Transactions of the Royal Society of London. Series A*, **358**, 957–990.

Jakobsson, M., Macnab, R., Mayer, L., Anderson, R., Edwards, M., Hatzky, J., Schenke, H. W., and Johnson, P., 2008. An improved bathymetric portrayal of the arctic ocean: Implications for ocean modeling and geological, geophysical and oceanographic analyses. *Geophysical Research Letters*, **35**, L07602, doi:10.1029/2008GL033520.

Korhonen, J., Fairhead, J., Hamoudi, M., Hemant, K., Lesur, V., Mandea, M., Maus, S., Purucker, M., Ravat, D., Sazonova, T., and Thbault, E., 2007. *Magnetic anomaly map of the world – carte des anomalies magnétiques du monde, scale 1:50, 000, 000*, 1st edn. Paris: Commission for Geological Map of the World, UNESCO.

Korte, M., and Constable, C., 2005. Continuous geomagnetic field models for the past 7 millennia: 2. *Geochemistry, Geophysics, Geosystems*, **6**(1), 1–18.

Korte, M., and Haak, V., 2000. Modelling European repeat station and survey data by SCHA in search of time-varying anomalies. *Physics of the Earth and Planetary Interiors*, **122**, 205–220.

Korte, M., and Holme, R., 2003. Regularization of spherical cap harmonics. *Geophysical Journal International*, **153**(1), 253–262.

Langlais, B., Lesur, V., Purucker, M. E., Connerney, J. E. P., and Mandea, M., 2009. Crustal magnetic fields of terrestrial planets. *Space Science Reviews*, **152**(1–4), 223–249, doi:0.1007/s11214-009-9557-y.

Lesur, V., 2006. Introducing localized constraints in global geomagnetic field modelling. *Earth Planets and Space*, **58**, 477–483.

Lesur, V., Wardinski, I., Rother, M., and Mandea, M., 2008. GRIMM – The GFZ Reference Internal Magnetic Model based on vector satellite and observatory data. *Geophysical Journal International*, **173**, 382–394, doi:10.1111/j.1365-246X.2008.03724.x.

Lesur, V., Wardinski, I., Asari, S., Minchev, B., and Mandea, M., 2010. Modelling the earth's core magnetic field under flow constraints. *Earth Planets and Space*, **62**(6), 503–516.

Lowes, F. J., 1966. Mean-square values on sphere of spherical harmonic vector fields. *Journal of Geophysical Research*, **71**, 2179.
Mandea, M., and Purucker, M., 2005. Observing, modeling, and interpreting magnetic fields of the solid earth. *Surveys in Geophysics*, **26**, 415–459, doi:10.1007/s10712-005-3857-x.
Mandea, M., and Thèbault, E., 2007. *The Changing Faces of the Earth's Magnetic Field: a Glance at the Magnetic Lithospheric Field, from Local and Regional Scales to a Planetary View*. Paris: Commission for the Geological Map of the World, p. 49.
Mandea, M., Holschneider, M., Lesur, V., and Lühr, H., 2010a. The earth's magnetic field at the champ satellite epoch, in system earth via geodetic-geophysical space techniques. In Flechtner, F., Gruber, T., Gntner, A., Mandea, M., Rothacher, M., Schöne, T., and Wickert, J. (eds.), *System Earth via Geodetic-Geophysical Space Techniques*. Berlin, Heidelberg: Springer, doi:10.1007/978-3-642-10228-8.
Mandea, M., Holme, R., Pais, A., Pinheiro, K., Jackson, A., and Verbanac, G., 2010b. Geomagnetic jerks: Rapid core field variations and core dynamics. *Space Science Reviews*, **155**, 147–175, doi:10.1007/s11214-010-9675-6.
Maus, S., Yin, F., Lühr, H., Manoj, C., Rother, M., Rauberg, J., Michaelis, I., Stolle, C., and Müller, R., 2008. Resolution of direction of oceanic magnetic lineations by the sixth-generation lithospheric magnetic field model from CHAMP satellite magnetic measurements. *Geochemistry, Geophysics, Geosystems*, **9**(7), 07021.
Merrill, R. T., McElhinny, M. W., and McFadden, P. L., 1996. *The Magnetic Field of the Earth*. London: Academic Press.
Olsen, N., and Mandea, M., 2008. Rapidly changing flows in the earth's core. *Nature Geosciences*, **1**, 390–394, doi:10.1038/ngeo203.
Olsen, N., Haagmans, R., Sabaka, T. J., Kuvshinov, A., Maus, S., Purucker, M. E., Rother, M., Lesur, V., and Mandea, M., 2006. The Swarm End-to-End mission simulator study: Separation of the various contributions to the Earth's magnetic field using synthetic data. *Earth, Planets and Space*, **58**, 359–370.
Olsen, N., Mandea, M., Sabak, T. J., and Tøffner-Clausen, L., 2009. Chaos-2 – a geomagnetic field model derived from one decade of continuous satellite data. *Geophysical Journal International*, **179**, 1477–1487, doi:10.1111/j.1365-246X.2009.04386.x.
Olsen, N., Mandea, M., Sabaka, T. J., and Toffner-Clausen, L., 2010. The chaos-3 geomagnetic field model and candidates for the 11th generation IGRF. *Earth, Planets and Space*, **62**, 719–729.
Panet, I., Chambodut, A., Diament, M., Holschneider, M., and Jamet, O., 2006. New insights on intraplate volcanism in French polynesia from wavelet analysis of grace, champ and sea-surface data. *Journal of Geophysical Research*, **111**, B09403, doi:10.1029/2005JB004141.
Roberts, P. H., and Scott, S., 1965. On the analysis of secular variation, 1, A hydromagnetic constraint: Theory. *Journal of Geomagnetism and Geoelectricity*, **17**, 137–151.
Robinson, P., Harrison, R. J., McEnroel, S. A., and Hargraves, R. B., 2002. Lamellar magnetism in the haematite–ilmenite series as an explanation for strong remanent magnetization. *Nature*, **418**, 517–520, doi:10.1038/nature00942.
Sabaka, T. J., Olsen, N., and Purucker, M. E., 2004. Extending comprehensive models of the Earth's magnetic field with Ørsted and CHAMP data. *Geophysical Journal International*, **159**, 521–547, doi:10.1111/j.1365-246X.2004.02421.x.
Schott, J. J., and Thébault, E., 2010. Modelling the earth's magnetic field from global to regional scales. In Mandea, M., and Korte, M. (eds.), *Geomagnetic Observations and Models*. Dordrecht: Springer.
Shure, L., Parker, R., and Backus, G., 1982. Harmonic splines for geomagnetic modelling. *Physics of the Earth and Planetary Interiors*, **28**, 215–229, doi:10.1016/0031-9201(82)90003-6.
Spector, A., and Grant, S., 1970. Statistical models for interpreting aeromagnetic data. *Geophysics*, **35**, 10–119011440092.
Thébault, E., Schott, J. J., Mandea, M., and Hoffbeck, J. P., 2004. A new proposal for spherical cap harmonic modelling. *Geophysical Journal International*, **159**, 83–103, doi:10.1111/j.1365-246X.2004.02361.x.
Wardinski, I., and Holme, R., 2006. A time-dependent model of the earth's magnetic field and its secular variation for the period 1980 to 2000. *Journal of Geophysical Research*, **111**, B12101, doi:10–10292006004401.
Wardinski, I., Holme, R., Asari, S., and Mandea, M., 2008. The 2003 geomagnetic jerk and its relation to the core surface flows. *Earth and Planetary Science Letters*, **267**, 468–481, doi:10.1016/j.epsl2007.12.008.
Wicht, J., Harder, H., and Stellmach, S., 2010. Numerical dynamo simulations - from basic concepts to realistic models. In Freeden, W., Nashed, Z., and Sonar, T. (eds.), *Handbook of Geomathematics*. Heidelberg: Springer.

Cross-references

Magnetic Data Enhancements and Depth Estimation
Magnetic Gradiometry
Magnetic Methods, Principles
Magnetic Methods, Satellite
Magnetic Methods, Surface
Magnetic, Global Anomaly Map

MAGNETIC STORMS AND ELECTROMAGNETIC PULSATIONS

Gurbax S. Lakhina[1], Bruce T. Tsurutani[2]
[1]Indian Institute of Geomagnetism, New Panvel (W), Navi Mumbai, India
[2]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

Definition

Electromagnetic pulsations. Geomagnetic field oscillations having periods of 0.2–600 s.
Magnetic storms. Global disturbances in the geomagnetic field lasting from half to several days.
Magnetospheric substorms. A set pattern of disturbances occurring in the magnetosphere, lasting for a period of approximately one to a few hours, in response to an enhanced coupling between the solar wind and the geomagnetic field.

Electromagnetic pulsations

The Earth's magnetic field varies over time scales spanning many orders of magnitudes, from hundreds of thousands of years to a fraction of a second. The variations on the long periods, from hundreds of thousands to tens of years, are believed to be due to the geodynamo processes. The shorter period variations, from about 22 years to a fraction of a second are thought to be caused either directly or indirectly by the solar output. The shortest period oscillations in the geomagnetic field varying from about ~0.2 to 600 s are called electromagnetic pulsations

or simply magnetic pulsations. The frequencies of magnetic pulsations fall in the range of ultra-low-frequency (ULF) electromagnetic waves. The amplitudes of these electromagnetic pulsations decrease as their time periods decrease (or their frequencies increase). At the longest time periods, magnetic pulsations can have amplitudes of several hundreds of nT, whereas at the shortest time periods, their amplitudes are fractions of a nT. Magnetic pulsations have been detected on the ground by magnetometers and in the magnetosphere by in situ satellite magnetic and electric field sensors.

Classification of magnetic pulsations

Magnetic pulsations observed on ground have been classified into two broad classes depending on the waveform and period. This has been done formally by the International Association of Geomagnetism and Aeronomy (IAGA) (Jacobs et al., 1964). Pulsations with quasi-sinusoidal waveform are called pulsation continuous (Pc), and those with an irregular waveform pattern are called pulsation irregular (Pi). Each main class is subdivided into period bands that roughly isolate a particular type of pulsation as shown in Table 1. During disturbed geomagnetic conditions like substorms and magnetic storms, giant magnetic pulsations, called Pc6 or Ps6, are observed with periods ranging from ~600–900 s.

Generation mechanisms of magnetic pulsations

Magnetic pulsations observed on the ground or in space are produced by a variety of plasma processes occurring in the magnetosphere and the solar wind (McPherron, 2005).

The magnetosphere is the region of space where the Earth's magnetic field is confined by the solar wind plasma emanating from the Sun. A schematic of the three-dimensional view of the magnetosphere is shown in Figure 1. It shows various important plasma regions as well as the current systems. Fluctuations in these currents arising from various plasma waves and instabilities excite the magnetic pulsations. Therefore, changes in the plasma and magnetic fields in the solar wind can have dramatic effects on the type of waves seen at a particular location at Earth.

Observations show that continuous pulsations sharing the same frequency range may differ, for example, in their polarization, harmonic structure, or spatial occurrence dependence, and thus can have different generation mechanisms. There are two main mechanisms for the long-period (Pc3–Pc5) magnetic pulsations. The first one is based on field-line resonance theory where a monochromatic surface wave is excited by some plasma instability (e.g., Kelvin–Helmholtz, drift mirror mode, etc.) at the magnetopause. This instability resonantly couples with a shear Alfvén wave associated with local field-line oscillations (Chen and Hasegawa, 1974). The second mechanism is based on cavity modes where global fast modes are excited throughout the entire magnetospheric cavity in response to sudden impulses in the solar wind (Kivelson and Southwood, 1985). The cavity modes couple to field-line resonances that drive currents in the ionosphere producing magnetic pulsations. However, some daytime Pc3–Pc5 pulsations are believed to be driven directly by the solar wind pressure fluctuations. Short-period-pulsations (Pc1 and Pc2) are generated locally in the equatorial region of the magnetosphere by an electromagnetic ion cyclotron instability. The instability is driven by a temperature anisotropy of the energetic protons from the ring-current and plasma-sheet regions. On the other hand, coupling of global cavity modes to field-line resonances is considered as a likely mechanism for the generation of irregular magnetic pulsations (Cheng et al., 1998). Nighttime Pi2 pulsations are probably excited by earthward propagating fast-mode waves launched at substorm onset by large-scale magnetic reconfiguration associated with cross-tail current disruptions.

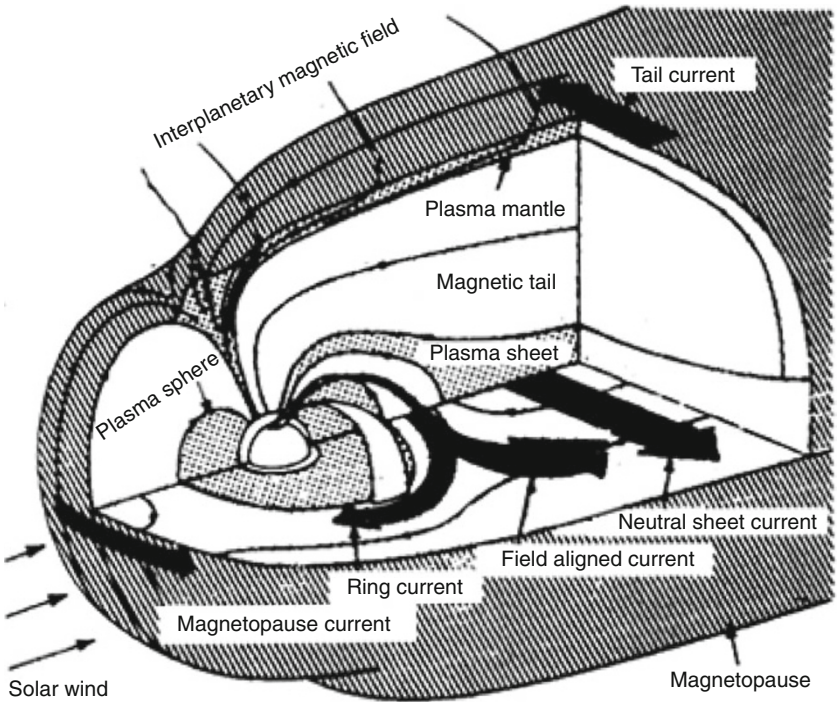
Substorms and magnetic storms

Substorms and magnetic storms are geomagnetic activity phenomena. A substorm (Akasofu, 1964) is composed of a set sequence of evolution of ionospheric auroral forms and magnetospheric disturbances that typically last from ~30 min to 1 h. There are often many substorms in a day. During substorms there is an explosive release of stored magnetotail energy directed toward the midnight sector magnetosphere. This results in the excitation of discrete auroras that become widespread, intense, and much more agitated than normal. The Earth's magnetic field gets disturbed due to intensified field-aligned currents and auroral electrojets. The ultimate source of energy for substorms is the solar wind.

Magnetic storms (Gonzalez et al., 1994) are driven directly by solar drivers like coronal mass ejections (CMEs) and flares, and Corotating Interaction Regions (CIRs) associated with fast streams emanating from coronal holes (Tsurutani et al., 1995, 2006). The cause of magnetic storms is long intervals of southward interplanetary magnetic fields that considerably enhance the efficiency

Magnetic Storms and Electromagnetic Pulsations, Table 1 Classification of magnetic pulsations

	Continuous pulsations					Irregular pulsations	
	Pc1	Pc2	Pc3	Pc4	Pc5	Pi1	Pi2
Period range (s)	0.2–5	5–10	10–45	45–150	150–600	1–40	40–150
Frequency range (mHz)	200–5,000	100–200	22–100	7–22	2–7	25–1,000	7–25



Magnetic Storms and Electromagnetic Pulsations, Figure 1 Schematic three-dimensional view of the earth's magnetosphere formed by the interaction of solar wind from the Sun (from the left) with the geomagnetic field. Small arrows indicate the direction of the magnetic field lines. *Thick arrows* show the direction of electric currents. Various current systems present in the magnetosphere are shown.

of magnetic reconnection process, leading to strong plasma injection from the magnetotail toward the inner magnetosphere. This leads to intense auroras at high-latitude nightside regions and at the same time intensifies the ring current, which causes a diamagnetic decrease in the Earth's magnetic field measured at near-equatorial magnetic stations.

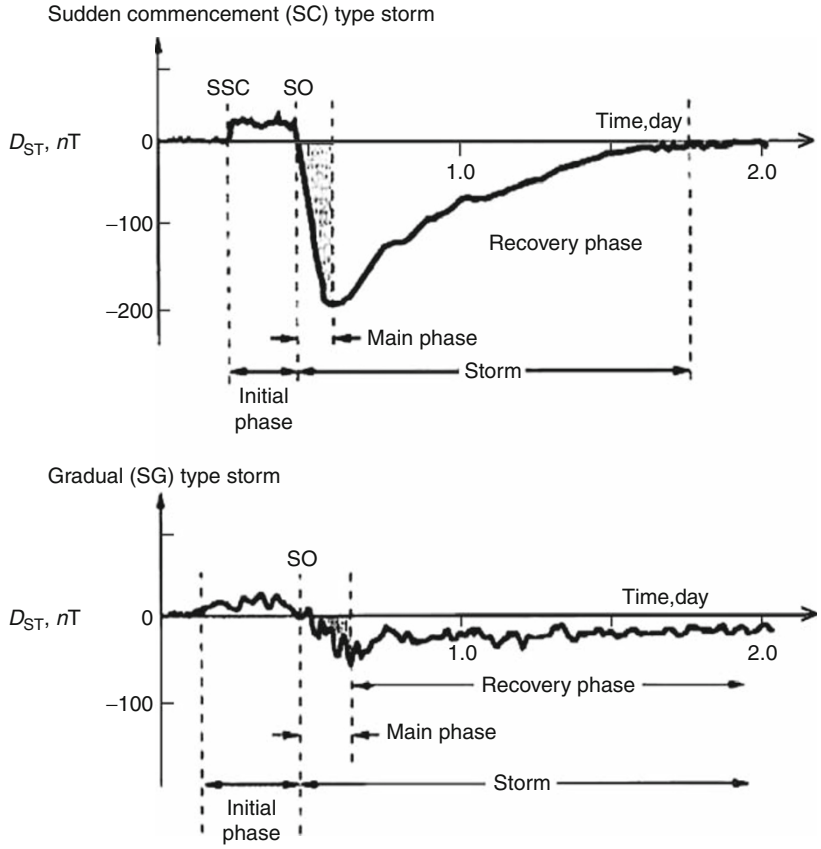
The *sudden commencement* (SC) storms are characterized by a sudden increase in the horizontal magnetic field intensity shortly before the main phase (see top panel of Figure 2). This sudden increase in magnetic field strength is caused by the interplanetary shock compression of the magnetosphere. The period between the sudden commencement and the storm main phase is called the *initial phase*. However, all magnetic storms do not have an initial phase. A geomagnetic storm not accompanied by a SC is called a *gradual geomagnetic storm* (SG) type (bottom panel of Figure 2). The *main phase* where the horizontal components of the low-latitude ground magnetic fields are significantly depressed over a time span of one to a few hours is caused by intensification of the ring-current energetic ions injected into the inner magnetosphere (L ~2–7). This is followed by a *recovery phase*, which may extend for ~10 h or more, associated with the decay of the ring current. The intensity of a geomagnetic storm is expressed in terms of the disturbance storm time index (Dst) or SYM-H (basically the same as Dst except the higher time resolution) index, which is a measure of the intensity of the ring current.

Relationship between magnetic pulsations and substorms and magnetic storms

The magnetospheric magnetic field configuration and plasma populations undergo drastic changes during substorms and magnetic storms. This directly affects the magnetic pulsations that are generated internally, for example, Pc1 and Pc2 by electromagnetic ion cyclotron instability, Pc3 and Pc4 by drift mirror instability, and Pi1 and Pi2 associated with the formation of substorm current wedges. Ps6 magnetic pulsations are also generated during substorms. Injection of energetic oxygen ions during magnetic storms tend to reduce the frequency of the long-period-pulsations, especially Pc5s. Further, the ULF wave power increases substantially during magnetic storms. On the other hand Pc5 magnetic pulsations are considered as a possible factor that can accelerate electrons to relativistic energies during magnetic storm recovery phases.

Magnetic storms and society

Intense (Dst < -100 nT) and super-intense (Dst < -500 nT) geomagnetic storms create hostile space weather conditions that can be hazardous to Earth-orbiting spacecrafts



Magnetic Storms and Electromagnetic Pulsations, Figure 2 Schematic of a magnetic storm sudden commencement (SC) driven by an ICME (*top*) and gradual (SG) type caused by a CIR (*bottom*). All storms may not have initial phases. (Slightly modified from Tsurutani et al., 2006.)

as well as technological systems at ground. Some adverse effects are malfunctioning and even failure of sensitive instruments and satellite command systems due to high fluxes of relativistic "killer" ~MeV electrons, satellite communication failures, satellite data loss and navigational errors, and loss of low earth-orbiting satellites caused by an expanded upper atmosphere. The elevated radiation level is a threat to astronauts and jetliner passengers. Geomagnetically induced currents can cause damage to power-line transformers and corrosion of long pipelines and cables.

Conclusions

Electromagnetic pulsations and magnetic storms form an important component of space weather. The ground-based measurements of magnetic pulsations and magnetic storms offer a unique and simple way of monitoring the conditions in the magnetosphere and solar wind. These measurements can also be utilized for geophysical surveys to probe the subsurface conductivity structure of the Earth.

Acknowledgments

Portions of this research were performed at the Jet Propulsion Laboratory, California Institute of Technology under contract with NASA. GSL thanks Indian National Science Academy, New Delhi, for support under the Senior Scientist Scheme.

Bibliography

Akasofu, S.-I., 1964. The development of the auroral substorm. *Planetary and Space Science*, **12**, 273–282.

Chen, L., and Hasegawa, A., 1974. A theory of long-period magnetic pulsations, 1, steady state excitation of field line resonances. *Journal of Geophysical Research*, **79**, 1024–1032.

Cheng, C. C., Chao, J. K., and Hsu, T. S., 1998. Evidence of the coupling of a fast magnetospheric cavity mode to field line resonances. *Earth Planets and Space*, **50**, 683–697.

Gonzalez, W. D., Joselyn, J. A., Kamide, Y., Kroehl, H. W., Rostoker, G., Tsurutani, B. T., and Vasyliunas, V. M., 1994. What is a Geomagnetic Storm? *Journal of Geophysical Research*, **99**(A4), 5771–5792.

Jacobs, J. A., Kato, Y., Matsushita, S., and Troitskaya, V. A., 1964. Classification of geomagnetic micropulsations. *Journal of Geophysical Research*, **69**, 180–181.

Kivelson, M. G., and Southwood, D. J., 1985. Resonant ULF waves: a new interpretation. *Geophysical Research Letters*, **12**, 49–52.
McPherron, R. L., 2005. Magnetic pulsations: their sources and relation to solar wind and geomagnetic activity. *Surveys in Geophysics*, **26**, 545–592.
Tsurutani, B. T., Gonzalez, W. D., Gonzalez, A. L. C., Tang, F., Arballo, J. K., and Okada, M., 1995. Interplanetary origin of geomagnetic activity in the declining phase of the solar cycle. *Journal of Geophysical Research*, **100**, 21717.
Tsurutani, B. T., Gonzalez, W. D., Gonzalez, A. L. C., Guarnieri, F. L., Gopalswamy, N., Grande, M., Kamide, Y., Kasahara, Y., Lu, G., Mann, I., McPherron, R., Soraas, F., and Vasyliunas, V., 2006. Corotating solar wind streams and recurrent geomagnetic activity: a review. *Journal of Geophysical Research*, **111**, A07S01, doi:10.1029/2005JA011273.

Cross-references

Geoelectromagnetism
Geomagnetic Excursions
Geomagnetic Field, Measurement Techniques
Geomagnetic Field, Polarity Reversals
Magnetic Data Enhancements and Depth Estimation
Magnetic Methods, Airborne
Magnetic Methods, Satellite
Magnetic Methods, Surface
Magnetometers
Magnetovariation Studies

MAGNETIC, GLOBAL ANOMALY MAP

Kumar Hemant Singh
NASA Goddard Space Flight Center, Planetary Geodynamics Laboratory, Greenbelt, MD, USA
Goddard Earth Sciences and Technology Center, University of Maryland Baltimore County, Baltimore, MD, USA

Definition

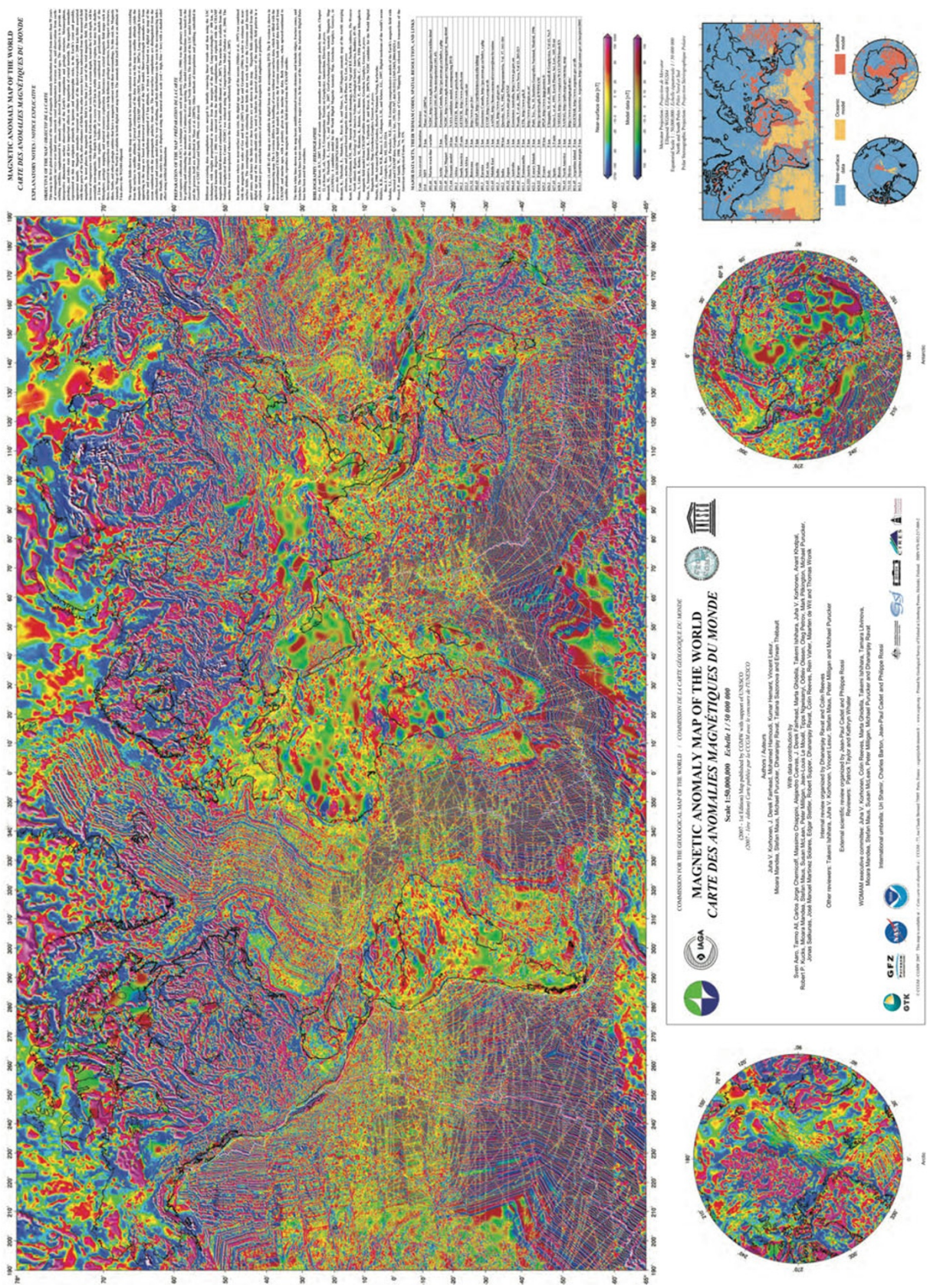
Global Magnetic Anomaly Map is an international effort to integrate near-surface and downward-continued spaceborne maps of the magnetic anomaly field. These global maps are available as grids, and find use in geologic and tectonic interpretation.

Introduction

Magnetic data have been collected from magnetometers towed by aircraft and ships for many decades now. Magnetic anomaly maps derived from regional airborne (see *Magnetic Methods, Airborne*) and marine surveys are beneficial to our understanding of the subsurface structure, chemical composition, and geodynamics of the Earth's crust (see *Earth's Structure, Continental Crust*) and lithosphere (Purucker and Whaler, 2007). Maps derived from marine surveys were instrumental in the discovery of sea-floor spreading and in the development of the theory of plate tectonics. Continental anomaly maps derived from aeromagnetic surveys interpreted with other geophysical information are used to delineate geologic and tectonic provinces and structures like faults, dikes, and lineations (see *Magnetic Anomalies, Interpretation*). These maps are also widely used, for example, in natural resource exploration, in estimating the thermal state of the earth's lithosphere (see *Lithosphere, Continental: Thermal Structure*), and the magnetic characterization (see *Remanent Magnetism*) of volcanic surfaces and deep interiors on both continents and oceans.

Satellite missions (see *Magnetic Methods, Satellite*) have also made significant contributions in enhancing our knowledge of the Earth's magnetic field (cf. Langel and Hinze, 1998). Satellite data are acquired and processed in three-dimensional spherical coordinates; hence, the derived magnetic anomaly maps are on a global scale. The satellite data are measured typically 400–750 km away from the causative sources in the Earth's lithosphere, lack sufficient resolution of geological features smaller than 400 km in wavelengths, but find extensive use in studying the long-wavelength anomalies. In comparison, magnetic anomaly maps prepared from data recorded near the surface (~few kilometers) and therefore near to the sources resolve short-wavelength-anomaly features better (see *Magnetic Anomalies, Interpretation*). There are, however, many geologic problems that could be better solved with the use of the self-consistent global magnetic anomaly maps that take advantage of the strengths of each data set while bypassing its limitation. For example, repeated rifting (see *Continental Rifts*) and movement of tectonic plates (see *Plates and Paleoreconstructions*) over hundreds of millions of years have made once adjacent landmasses drift (see *Continental Drift*) to completely different regions of the world. Reassembly requires consistently prepared magnetic anomaly maps over the entire globe consisting of both short and long wavelengths. The world digital magnetic anomaly map (WDMAM) derived from merging near-surface magnetic data and a global satellite model will aid evaluation of hypotheses about geologic origins of tectonic rifts, fracture zones that run across the continents, and tectonic history of many regions.

The WDMAM is the magnetic field defined on a grid of 3 arc-min angular grid resolution at an altitude of 5 km above the World Geodetic System 1984 (WGS84) reference ellipsoid (Figure 1) (Korhonen et al., 2007). The world magnetic anomaly map is the first international collaborative effort to compile magnetic anomaly information on a global scale. The map is derived from more than 50 years of aeromagnetic, shipborne surveys over land and sea, and it is supplemented by observations from earth-orbiting satellites. The global map shows the variation in strength of the magnetic field (see *Geomagnetic Field, Global Pattern*) after the Earth's dipole field and, external field sources like magnetosphere and ionosphere have been subtracted from the measured field. The remaining variations in the field (few hundreds of nanoteslas (nT)) are due to changes in the magnetic



Magnetic, Global Anomaly Map, Figure 1 The World Digital Magnetic Anomaly Map published by Commission for the Geological Map of the World (CGMW) in July 2007. The map is shown in Mercator Projection and is at an equatorial scale 1:50,000,000. (Courtesy: Korhonen et al., 2007.)

properties of the crustal rocks. In most places, the magnetic anomaly field is less than 1% of the main magnetic field of the Earth.

The magnetic fields derived in this map are internally consistent within the measurement domain, extending from the Earth's surface to satellite altitude (<750 km). The map comprises both the short- and long-wavelength anomalies from different sources. Short-wavelength anomalies are derived from merging continental compilations and marine track line data. Long-wavelength features in this global map are based on a downward-continued grid of the MF5 model, a model derived from CHAMP satellite magnetometer measurements. In oceanic regions where the data gaps exist, the digital age map of the ocean floor (Müller et al., 1997) was combined with the geomagnetic polarity time scale (Gee and Kent, 2007; Kent and Gradstein, 1986) (see *Geomagnetic Field, Polarity Reversals*) to estimate the total field anomaly at 5 km altitude.

The global magnetic anomaly map is available in both digital and print form. Two versions (A and B) of the map are available on DVD. The B version is shown in Figure 1. The map is displayed in natural colors (red = positive values, blue = negative values) and produced from Generic Mapping Tools (GMT) (Wessel and Smith, 1991). The A version differs in its handling of areas with data gaps, which are filled with the downward-continued MF5 model. The map images and the DVD version of the WDMAM 2007 are available at http://ftp.gtk.fi/WDMAM2007/.

Early developments

Magnetic surveys have been carried out on land, sea, and from the air over all continents and oceans. The patchwork of aeromagnetic surveys over the continents and the tracks of oceanographic survey vessels across the oceans, together, make up an enormous but distinct body of data. This incomplete and nonuniform coverage of near-surface data has recently been supplemented by reliable maps from earth-orbiting satellites such as CHAMP (http://www-app2.gfz-potsdam.de/pb1/op/champ/). These satellite-derived models (see *Magnetic Methods, Satellite*) provide uniform coverage over the globe that can help resolve uncertainties in near-surface compilations produced by stitching together disparate sets of data. A new initiative was therefore required to combine these near-surface anomaly maps into a reliable global compilation and make it more accessible for geological mapping, educational purposes, and future research.

The original data is held by a large number of individual organizations around the world and, in some cases, is still commercially used and politically sensitive. In 1977, the International Association of Geomagnetism and Aeronomy (IAGA) (http://www.ngdc.noaa.gov/IAGA/vmod/) realized the importance of magnetic anomaly mapping for geologic and tectonic interpretation (see *Magnetic Anomalies, Interpretation*) and noted the existence of numerous airborne and marine magnetic data. They began a concerted international effort to bring the global magnetic anomaly data set into place alongside other worldwide databases to maximize its use. Through a series of resolutions in 1979, 1995, and 1997, the WDMAM task force was constituted by IAGA at the International Union of Geodesy and Geophysics General (IUGG) Assembly in 2003 to prepare a roadmap to complete the work. The job of the task force included (1) cataloging existing data compilations that cover significant areas of the Earth, (2) negotiating access to such data at a resolution sufficient for a global compilation, (3) producing a preliminary paper map at the IAGA Assembly in Toulouse, France, July 2005, and (4) while addressing the main scientific and practical issues, producing a digital magnetic anomaly map of the world for the 24th IUGG General Assembly, Perugia, Italy, July 2007.

The WDMAM task force urged various organizations and custodians of data to release nonsensitive versions of the classified or confidential aeromagnetic and marine track line digital data to the public domain for the WDMAM project. The organizations holding the data responded to the call and made near-surface data available for the first version of the WDMAM map. In June 2006, a call for submission of WDMAM candidate models was issued. In November 2006, five models were submitted for evaluation. Five teams at the National Geophysical Data Center (NGDC), NASA's Goddard Space Flight Center (GSFC), Geological Survey of Finland (GTK), GeoForschungsZentrum, Potsdam, Germany, and the University of Leeds, UK produced candidate models for the WDMAM project (http://projects.gtk.fi/WDMAM/project/). After careful evaluation, the candidate model developed by NGDC was selected as a base map for representing anomalies over continents while the GAMMA model was selected for oceans. The first version of the official WDMAM was unveiled at the 24th IUGG General Assembly in Perugia, Italy, in July 2007 by the Commission for the Geological Map of the World (CGMW). The United Nations Educational, Scientific and Cultural Organization (UNESCO) kindly supported and provided the funding for the printing.

Aeromagnetic and marine compilations

The first airborne surveys were done over former Soviet Union in 1936 and the first maps of the total intensity were published at 1:2.5 million scale in 1974. These maps were later digitized and made available at 1 and 3 arc-min grid resolution from the NGDC magnetic data archive (http://ols.nndc.noaa.gov/plolstore/plsql/olstore.prodspecific?prodnum=G01284-CDR-A0001). More recently, another magnetic compilation for Russia compiled by the A.R. Karpinsky Russian Geological Research Institute is available as a 5 km resolution of the data grid (VSEGEI, http://www.vsegei.ru/WAY/247038/locale/EN).

The first ever North American continental-scale compilation was derived by merging magnetic anomaly maps

of USA, Canada, and surrounding waters. The map was released in 1987 as a part of the Decade of North American geology. The map had erroneous long-wavelength features, which revealed shortcomings and pitfalls in stitching together residual intensity maps reduced from measured data using different IGRF models (see *Magnetic Data Enhancements and Depth Estimation*). After correcting for these erroneous long wavelengths using Comprehensive Model (CM) (Sabaka et al., 2002), the improved maps for USA and Canada, in addition to the aeromagnetic compilation for Mexico, were combined to produce the revised North American magnetic anomaly map (Bankey et al., 2002). The Arctic and North Atlantic magnetic compilation including nearby landmasses was produced from a patchwork of shipborne and airborne surveys (Verhoef et al., 1996). Because the compilation had incorrect long wavelengths greater than 400 km, it was filtered before releasing the grid to the public domain.

Across the Atlantic, Wonik et al. (2001) prepared the first European compilation by merging magnetic data from North, West, and East Europe. The map retains the long-wavelength features but shows a poor correlation when compared with satellite data. However, there have been many developments in the last decade in the preparation of high-resolution magnetic compilations from marine and airborne surveys. The Geological Survey of Japan compiled the grid for East Asia in 2002. The fourth edition, but a significant improvement over its previous version, the high-resolution magnetic anomaly grid for Australia and its surrounding oceans was released (Minty et al., 2003). The noticeable feature of this map is its control over the long-wavelength anomalies, which was achieved by flying long aeromagnetic traverses around Australia. The maps for China have also been recently made available in digital form by the Chinese National Aerogeophysics Surveys. Another map that was digitized for WDMAM project is a combined map for China, Mongolia, and Russia produced in 1995 by the Geological Survey of Canada. This map also had long wavelengths greater than 400 km corrected. In addition, the Antarctic Digital Magnetic Anomaly Project (ADMAP) provided the continental-scale compilation over Antarctica (Golynsky et al., 2001). The map was derived from merging aeromagnetic and marine data with the satellite magnetic data.

In addition to these data in public domain, there have been efforts by industry-led consortia to produce maps for Africa, South America, and Middle East (http://www.getech.com/data/global_mag_margins.htm). Geological Survey of India also contributed the Indian magnetic compilation. Over oceans, NGDC's archives of track line data sets (http://www.ngdc.noaa.gov/mgg/geodas/geodas.html) were made available. In addition, the Project Magnet ocean traverse data were also available at http://www.ngdc.noaa.gov/geomag/proj_mag.shtml.

Table 1 summarizes the marine and airborne magnetic compilations made available by the WDMAM task force to produce the candidate models. The data density greatly varies between the Northern and the Southern hemispheres. The coverage is especially sparse over oceans in the southern hemisphere where some of the marine track line data were not available for the project. Missing data are noted over parts of the Arabian Peninsula, India, East Asia and many countries within Africa and South America. For processing the compilations, the datasets were grouped in three sets. The NGDC and Project Magnet marine and aeromagnetic track line data comprised the first set. Due to large errors and inconsistencies to be discussed later, the Project Magnet data was not used in the preparation of GAMMA candidate model. The second set included the gridded aeromagnetic and marine magnetic anomaly grids. The downward-continued MF5 model derived from the CHAMP satellite data made the third set. Of the track line profile data, the NGDC marine data and Project Magnet measurements together make 27 million points on 31,000 tracks. With scales ranging from country to continents, a total of 19 gridded data sets were made available.

The metadata information from the compilations, when available, shows the coverage to be in various projections and datums depending on their geographical representation on the globe. Hence, prior to applying any filtering or merging procedures, all grids were preprocessed and transformed from their local reference system to a common representation on the WGS84 global reference ellipsoid, using conversion formulas (Snyder, 1987). The software, Oasis Montaj, Geosoft© was used in the preparation of GAMMA model for upward continuation of the data, while the Leeds model was produced on the ESRI's ARCGIS (www.esri.com).

Satellite compilations

Several global models of the lithospheric magnetic field, based on satellite missions, have been produced in the last 4 decades (see *Magnetic Methods, Satellite*). Although spherical harmonic analyses of data gathered at satellite altitude are typically used to compute the magnetic potential, they follow two different philosophies. The comprehensive approach (CM4), proposed by Sabaka et al. (2004), uses POGO, Magsat, Ørsted, CHAMP, and SAC-C measurements (see *Geomagnetic Field, Measurement Techniques*) to derive a model that accounts for magnetic sources (see *Geomagnetic Field, Theory*) from the Earth's main field to the external field sources in the ionosphere and magnetosphere and their temporal variations, including the crustal magnetic field component up to spherical harmonic degree 60. In contrast to the comprehensive approach, the second philosophy strictly focuses on the lithospheric field representation. This approach requires careful data selection and processing to clean the data from non-lithospheric sources (see *Magnetic Methods, Satellite*). MF series of models are derived from CHAMP scalar and vector magnetometer measurements (http://geomag.org/models/index.html). The MF4 model used 4 years of CHAMP data and is derived up to

Magnetic, Global Anomaly Map, Table 1 The available gridded data and marine track line data sets for WDMAM project. The official WDMAM codes and references are mentioned (Courtsey: Maus et al., 2007b)

Code	Area covered	Resolution	References
701.43	North America	1 km	NAMAG, http://pubs.usgs.gov/sm/mag_map/
302.43	Antarctica	5 km	ADMAP, http://www.geology.ohio-state.edu/geophys/admap/
504.43	Australia	1 km	Geoscience Australia, http://www.ga.gov.au/
601.43	Europe	5 km	BGR, http://www.bgr.bund.de/
121.43	Arctic	5 km	GSC, http://gsc.nrcan.gc.ca/index_e.php
421.43	Middle East	1 km	AAIME, http://home.casema.nl/errenwijlens/itc/aaime/
411.43	East Asia	2 km	CCOP, http://www.ccop.or.th/
442.2	India	50 km	Qureshy (1982)
441.3	India	5 km	GSI, http://www.gsi.gov.in/
201.2	Africa and South America	15 min	GETECH, http://www.getech.com/
625.2	France	15 min	IPGP, http://www.ipgp.jussieu.fr/
627.43	Spain	10 km	Socias et al. (1991)
222.3	South Africa	1.5 min	SADC, http://www.sadc.int/
611.3	Fennoscandia	5 km	GTK, http://www.gtk.fi/
626.2	Italy	5 km	Chiappini et al. (2000)
622.2	Canary Islands	5 km	Socias and Mezcua [1996]
812.3	Argentina margin	5 km	DNA, http://www.dna.gov.ar/
811.45	Argentina inland	5 km	SEGEMAR, http://www.segemar.gov.ar/db/
401.3	Eurasia	2 km	GSC, http://gsc.nrcan.gc.ca/index_e.php
628.3	Russia	5 km	VSEGEI, http://www.vsegei.ru/WAY/247038/locale/EN
101.45	Marine track line	Variable	http://www.ngdc.noaa.gov/mgg/geodas/trackline.html
131.45	Project Magnet	Variable	http://www.ngdc.noaa.gov/geomag/proj_mag.shtml

spherical harmonic degree 90. The map showed inconsistencies in the polar regions due to high data-noise levels. The use of localized basis functions by Lesur and Maus (2006) led to the derivation of MF4× model which showed improvement in lithospheric field at high latitudes. In 2007, the fifth generation of CHAMP models, MF5, was released (Maus et al., 2007a). MF5 extends to spherical harmonic degree 100 (400 km resolution) and showed a significant improvement, especially between 60° and 80° latitude, compared to previous models. After comparing the MF series of satellite compilations, the WDMAM committee, recommended using the MF5 model for the preparation of candidate models. Recently, using only 3 years of CHAMP measurements, the high-resolution MF6 model was derived (Maus et al., 2008). The MF6 model is able to resolve lithospheric magnetic field to degree 120, corresponding to 333 km wavelength. The MF6 model is also the first satellite-data-based global magnetic anomaly map to resolve the direction of oceanic magnetic lineations, which makes a helpful addition to the study of oceanic lithosphere (see *Lithosphere, Oceanic*) from space. The EMAG2 model uses the MF6 model to substitute the long-wavelength components.

Preparation of the global magnetic anomaly map

The NGDC candidate model (Maus et al., 2007b) uses the method of least-squares collocation (LSC) and line-leveling methods to generate a regular grid from irregularly spaced marine and aeromagnetic measurements. Other approaches for computing the candidate model include removal of a regional polynomial, iteratively adjusting a low-degree main field model for different epochs from each of the compilations, upward continuation of the data to 5 km above the WGS84 ellipsoid, and finally knitting, merging, and filtering them together (Hamoudi et al., 2007). Long wavelengths (>400 km or spherical harmonic degrees ≤100) were removed from the individual compilations, and replaced with the MF5 anomaly map of the total field downward-continued to 5 km altitude (Hemant et al., 2007). The following sections describe the preprocessing techniques and methods followed to derive the NGDC candidate model for WDMAM.

The WDMAM requires an accurate estimate of the crustal magnetic field on a 5 km angular resolution grid. Gridded data sets were mostly available in a 5 km grid or even had coarser resolution. In case of track line data, the resolution was variable (Table 1). Therefore, to prevent further loss of information during processing, all grids were decimated to 2.5 km angular resolution using the minimum curvature algorithm of GMT. Most of the compilations, in particular over the Northern Hemisphere, had overlapping regions. The compilations known to have higher accuracy or finer resolution were retained within overlapping parts by rejecting the coarser ones. In the process, the 15 arc-min GETECH data of South America and Africa were discarded whenever a higher resolution grid for the same region was available. Generally, the aeromagnetic surveys over the land are draped, meaning that they follow the surface topography at a constant altitude above terrain. Thus, the derived compilations from these surveys follow topography. Therefore, to accurately process the altitude information with respect to a geoid, a topographic model of the Earth was used.

The NGDC ETOPO-2 (http://www.ngdc.noaa.gov/mgg/fliers/01mgg04.html) surface elevation grid was linearly interpolated to the measurement positions and added to the terrain clearance of all draped surveys, to compute the measurement altitude above the geoid.

The continental-scale compilations (Table 1) are produced after stitching together many regional aeromagnetic anomaly grids (Ravat et al., 2002). These regional surveys are flown at different times and, hence, use different international geomagnetic reference field (IGRF) models (see *Geomagnetic Field, IGRF*) to reduce the measured data. The patchwork grids are therefore prone to offsets and edge effects and mismatch in anomaly shapes and strengths across overlapping regions. They also have the potential to introduce spurious anomalies. These edge effects were corrected by subtracting a linear 2D trend from all smaller sized grids.

It is well known that the genuine crustal anomalies of strengths ~500 nT can easily be contaminated by external field disturbances of amplitudes 100 nT. Thus, it has been a challenge to filter out short-wavelength noise from aeromagnetic and marine track line data. To filter these spurious signals, spatiotemporal variations of the magnetic field of strong but fast external field disturbances and those recorded by instruments on board relatively slow moving ships and aircrafts are compared by computing the along-track derivative of the magnetic field residual. As a selection criterion, if the derivative exceeded 100 nT/km, 20 preceding and following measurement points on a track were removed. This selection procedure removed about 3% of the marine track line profile data and 9% of the Project Magnet data, which possibly were affected by external field disturbances.

Methods

Line-leveling of data

In aeromagnetic surveys, inaccuracy in the measured positions of crossover points of acquisition, line to line differences in flying heights along adjacent tracks, and inadequate compensation for the magnetic field resulting from the aircraft leads to errors covering a region that extends beyond the line spacing in each direction (see *Magnetic Data Enhancements and Depth Estimation*; *Magnetic Methods, Airborne*). These errors are visible as linear anomalies parallel to the flight lines. If uncorrected, the computation of derivatives enhances these noise effects along the lines. This problem is common with acquiring aeromagnetic data. A common processing method to remove these errors is to line-level the data. A set of tie-lines perpendicular to, but spaced greater than that of the main survey lines is normally acquired for line-leveling. The line-leveling procedure computes differences in the field values at the crossover points of the survey and tie-lines, and corrections are applied to minimize these differences in the least-square sense (Reeves, 2005).

The track line data, both marine and aeromagnetic, used for the WDMAM project is also likely to have this kind of problem. These track line data are first corrected for the main field and external fields. Additional navigational errors, instrument biases, and disturbances from the unfiltered external field remain in the data causing uncertainty in the longer wavelength part of the magnetic field. The line-leveling technique addresses these remaining errors in the track line data by minimizing the differences between adjacent or crossing tracks (Maus et al., 2007b).

The technique defines a correction function for each track given by

$$f_j(d) = \sum_{k=0}^{M_j} b_{j,k} \cos\left(\pi k d / X_j\right). \tag{1}$$

Here, X_j is the great circle distance from the first to the last point of the jth track and M_j is the number of parameters required for the estimation of the correction function of this track. The algorithm assumes $M_j = \text{trunc}\ (X_j/400\ \text{km}) + 1$, to correct for offsets with half-wavelength greater than 400 km. The choice of 400 km wavelength is based on the minimum wavelength the MF5 model is able to resolve. To minimize the misfit between all adjacent tracks and the differences of the track line profiles to the gridded data, the parameters $b_{j,k}$ are estimated from inversion. The line-leveling algorithm then minimizes the distance-weighted misfit between all measurement locations. The weight function $W(r_{1,2},z_1,z_2)$ used here is given by

$$W\left(r_{1,2}, z_1, z_2\right) = \max\left(0, 1 - \frac{1}{R_s}\sqrt{\left(r_{1,2}^2 + 2(z_1 - z_2)^2\right)}\right). \tag{2}$$

where R_s is the search radius, $r_{1,2} = \left((x_2 - x_1)^2 + (y_2 - y_1)^2\right)^{1/2}$ is the horizontal separation and z_1 and z_2 are the vertical coordinates of two locations 1 and 2 in Cartesian coordinate system. In Equation 2, the vertical separation (z_1–z_2) is up-weighted by a factor 2 to account for the greater variability of a potential field in the vertical than in the horizontal direction. The weights proportional to the distance from the margins of the grids were chosen. A value of 0.01 on the margin and 1.0 at 200 km inland were used. Although the line-leveling technique was effective in minimizing the misfist between adjacent tracks, it also introduced spurious anomalies with half-wavelengths larger than 200 km. However, these spurious long-wavelength anomalies were later replaced by those from the MF5 model.

The implementation of the technique first required the decimation of 31,000 tracks of NGDC and Project Magnet data into 98,000 segments. These track line data were then line-leveled against each other by estimating 110,000 correction coefficients (b_{jk}) from one grand inversion. This procedure also minimized offsets where tracks overlapped with gridded data. The segments that were affected by strong external field disturbances were removed when

differences between two adjacent segments exceeded 550 nT and misfit against the gridded data exceeded 500 nT. The line-leveling and selection procedures further reduced the rms misfit between two adjacent segments from 380 to 130 nT and against the gridded data from 400 to 90 nT.

Least-squares collocation

LSC is a scheme for interpolation by the weighted average of the measurements similar to the method called Kriging. The weighted averages are obtained in a minimum-variance sense. The technique of LSC has been used to solve a wide variety of problems in physical geodesy (Moritz, 1980). It has been especially useful in certain satellite missions applications, such as in satellite altimetry and satellite-to-satellite tracking. The LSC method is discussed in Langel and Hinze (1998) and has been applied by Maus et al. (2007b) to compute the NGDC global magnetic anomaly. Since locations of the measurement points in both track line and gridded data are not on a regular grid, one requires an interpolation procedure to estimate the field at regular grid nodes.

Following Moritz (1980), the LSC method estimates vector components of magnetic anomaly field values, b, on a predefined grid from a given set of vector magnetic measurements, **c**, at nearby locations. The vector of measurements, **c**, is defined as the sums of the true anomaly vector, **a**, plus an associated noise vector, **e** (i.e., **c = a + e**). To estimate the value of $\underline{b}$, covariance matrices are first computed as follows:

$$V_a = E\left(aa^T\right), V_b = E\left(bb^T\right),$$

$$V_{b,a} = E\left(ba^T\right), V_d = E\left(ee^T\right),$$

where E indicates the expected value. The collocation requires the estimation of anomaly field $\underline{b}$ in the least-square sense from

$$\underline{b} = V_{a,b}^T V_c^{-1} c, \tag{3}$$

where $V_{b,a} = \left(V_{a,b}\right)^T$ and

$$V_c = E\left(cc^T\right) = V_a + V_d. \tag{4}$$

Here, it is assumed that **e** and **a** are uncorrelated, so that $V_{a,e} = V_{e,a} = 0$; and V_d is the error covariance matrix of measurement noise. The efficiency of the LSC method then lies in computation of the covariance matrices V_a, V_b and $V_{a,b}$ from properties of the true signal (**a**) and the estimation of V_d from the properties of the noise.

For a Cartesian approximation of the data volume, Langel and Hinze (1998) suggest approximate correlation functions V_2 and V_3 between two locations at (x_1, y_1, z_1) and (x_2, y_2, z_2) of the form

$$V_2\left(r_{1,2}, z_1, z_2\right) = V_0 \beta \left(r_{1,2}^2 + \left(z_1 + z_2 + \beta\right)^2\right)^{-1/2}, \tag{5}$$

where $\beta = r_c/\sqrt{3}$, and

$$V_3\left(r_{1,2}, z_1, z_2\right) = V_0 \beta^2 \left(z_1 + z_2 + \beta\right)\left(r_{1,2}^2 + \left(z_1 + z_2 + \beta\right)^2\right)^{-3/2}, \tag{6}$$

where $\beta = r_c/0.766$.

Here, the variance V_0 is a linear scale factor for the horizontal axis, while the correlation length r_c is a scale factor on the vertical axis. For magnetic anomaly fields, the use of one of these correlation functions for collocation depends on the variation of the function V_2 and V_3 with distance. Based on correlation functions for NGDC magnetic data, Australia and North America, empirically, it was determined that V_3 gives a better fit than V_2. From V_3, the values of V_0 and r_c are estimated to be 40,000 nT^2 and 15 km respectively.

The value of $\underline{b}$ from Equation 3 can then be estimated by computing the matrix $V_{a,b}$ given by (Maus et al., 2007b)

$$V_{a,b} = V_3\left(r_{a,b}, z_a, z_b\right), \tag{7}$$

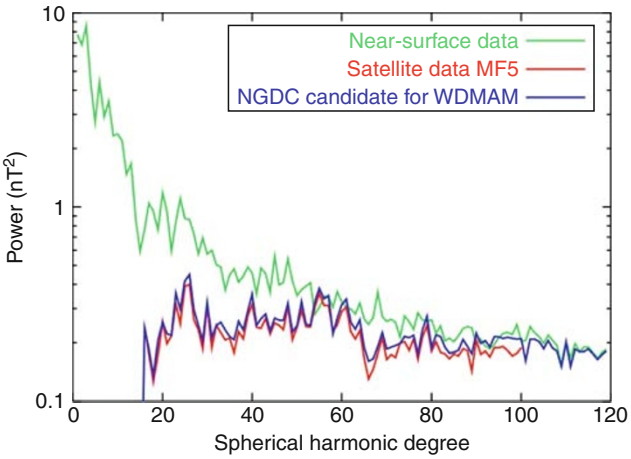
and the matrix V_c from Equation 4,

$$V_c = V_3\left(r_{1,2}, z_1, z_2\right) + E\left(ee^T\right), \tag{8}$$

are evaluated for all pairs of estimation (b) and measurement (a) locations. The error covariance matrix of measurement noise $E(ee^T)$ arise due to differences between adjacent tracks and between a track and a gridded data, which undermine all other sources of errors in the data. After line-leveling, it was assumed the uncorrected difference is 25 nT and the error covariance matrix was set to 625 nT^2 for data pairs from the same set, and 0 nT^2 for data pairs from different sets. The line-leveled data were finally merged with the gridded data using the LSC method. This grid is called the final merged grid.

Merging satellite models

Finally, the NGDC candidate model was derived by combining the final merged grid with MF5 by computing the magnetic potential. As a first step, the magnetic potential of the line-leveled data was computed to spherical harmonic degree 120. The degree correlation (cf. Langel and Hinze, 1998) between this model and MF5 was shown to be about 0.6 at degree 100. This high correlation offered some validation of high-degree coefficients of MF5. The correlation suggests that CHAMP satellite data is able to resolve many small-scale features beyond degree 100, which could be crustal in origin. The power spectra (Figure 2) of line-leveled data show excess power at low-degree. This is possibly caused by unremoved main field in continental-scale grids and by erroneous long-wavelength anomalies introduced in stitching together small-scale grids. At high degrees, the line-leveled data has 10% more power than MF5. Again, considering the different data sources and their uncertainties and various data processing steps and corrections, the disagreement in power is judged small. Following a good spectral agreement between the line-leveled and MF5



Magnetic, Global Anomaly Map, Figure 2 Power spectra of the near-surface (*line-leveled*) data, MF5 satellite model, and the final NGDC candidate model computed from spherical harmonic. (Courtesy: Maus et al., 2007a.)

model at degree 100, a sharp cutoff was used to substitute the long wavelengths of the near-surface data with the MF5 model. The substitution is carried out by subtracting the near-surface model to degree 100 from each grid cell and then adding the prediction from MF5. The final NGDC candidate model for WDMAM is a 3 arc-min angular grid of the total intensity anomaly at 5 km above the WGS84 geoid. The digital grid of the NGDC candidate for WDMAM also referred to as EMAG3 model is available at http://geomag.org/models/wdmam.html and additional plugins can be downloaded from http://www.getech.com/downloads/WDMAM for visualization in NASA World Wind.

Other candidate models for WDMAM

There were four other candidate models for WDMAM. Two of those models, the GAMMA model produced by GeoForschungsZentrum, Potsdam (Hamoudi et al., 2007) and the Leeds model derived by the team at University of Leeds, Leeds (Hemant et al., 2007) are published. These models are described next.

The GAMMA model

Most techniques were designed to improve the compatibility of the overlapping areas of neighboring grids. Hamoudi et al. (2007) applied different procedures to improve this incompatibility. These procedures are useful and easy to implement if only few grids are used, for they require only first order estimation of improved long-wavelength values.

Prior to processing, the gridded and track line data were statistically analyzed to check for their self consistency. For a pure anomaly field, one expects a symmetric distribution with zero mean. A skewed anomaly distribution not centered on zero possibly indicates some more or less severe contamination of other magnetic field contributions. In this respect, Russian, East Asian, and European compilations showed large anomaly intensity means. The standard deviation of anomaly distribution was usually between 100 and 200 nT for continents but could be as large as 930 nT on oceans. Oceanic magnetic anomalies of remament origin have amplitudes larger than on continents but such a wide distribution can only be explained by noise, bad tracks and outliers in the data. Another criterion, based on the arithmetic sum, highlighted some bias in marine, Russian, Australian, and East Asian compilations. These statistics were used to provide the available data and compilation with a quality flag.

The next procedure devoted to the analysis of data compatibility in the overlapping regions was based on visual inspection. By binning the data on a coarse grid, the spurious isolated points or track lines could be easily identified and manually removed. Other track lines, having consistent values but suspicious behavior, like a few Project Magnet data crossing the African continent or possibly biased data offshore Senegal were removed. Some GETECH South American and African compilations showed unusual shapes like over Bangui. Considering that these grids were coarser than any others, it was decided to allow them a low weight in the final process. Similarly, the Indian grid based on repeat stations showed poor resolution, so magnetic map boundaries of Qureshy (1982) on the eastern coast were applied. The European grid showed unexpected geographical shifts by a few kilometers, in particular over Germany, so the grids were used for analysis only when no other data were available in that region. In general, most grids suffered from edge effects, possibly resulting from some kind of filtering at the last stage of their production. Some of these errors were identified but could not be systematically corrected due to lack of complete metadata information.

The third procedure was to carry out an arguably homogeneous correction for the core magnetic field. It has been customary to correct all aeromagnetic surveys with IGRF/DGRF core field models (see *Geomagnetic Field, IGRF*; *Magnetic Anomalies, Interpretation*). Older surveys were corrected for more inaccurate core field models, particularly for those carried out at epochs when rapid core field variations were not yet identified (see *Geomagnetic Field, Secular Variation*). The core field correction was thus often followed by two-dimensional polynomial fitting or a Cartesian Fourier filtering. The DGRF models or polynomial fitting parameters were added to such old surveys whenever they were known, and a more recent core field model correction such as CM4 was (Sabaka et al., 2004). When the core field models were unknown, a better solution was found by trials and errors. A DGRF core field model for different epochs was added to each of these surveys and the CM4 was removed. When it improved the statistics (arithmetic mean, histogram of residuals), this correction was deemed fruitful. This procedure was not applied only for the Russian grid because it was outside of the CM4 time span (1960–2002). This procedure was shown to improve the data statistics by lowering the mean

intensity anomaly toward zero and the continuity in overlapping areas between adjacent compilations and other artifacts caused by the different model biases. For countrywide compilations with a clear knowledge of the acquisition date, in France, Italy, and Spain, this procedure provided a better way to correct for secular variation between different compilations. For some other surveys, however, for which the exact epoch and corrections were poorly known, a regional polynomial adjustment was necessary. In order to avoid Cartesian distortion due to the Earth's curvature a regional second order polynomial from all grids in the WGS84 geographic reference was removed. This polynomial fitting was not applied to the Australian grid, where the long-wavelength control was good, or to compilations over geographic poles (Arctic and Antarctic grids), where the processing procedure was based on using Cartesian reference frame instead. After this correction, the grids were shown to have the expected properties characterizing the anomaly field with the average anomaly intensity and an arithmetic sum zero (Table 2).

The fourth step consisted in upward continuation of the data to the desired altitude of 5 km. Marine track line data sets were also upward continued to mitigate effects of the sharp transition at the ocean-continent boundaries. Grids smaller than 2,000 km width, like Argentina, Austria, Fennoscandia, Italy, France, Mexico, and Spain, were upward continued to 5 km altitude above the geoid without any further processing. However, larger compilations (>2,000 km) were first split into 2,000 × 2,000 km^2 and upward continued individually. This dimension corresponds to the maximum size for which the Earth's curvature can be neglected (Nakagawa et al., 1985). After upward continuation, the edge effects between adjacent grids were not significant.

The preliminary processing discussed earlier reduced the large discontinuities but did not fully remove them. The Grid-knitting tool of GeoSoft was used to smooth the transition between adjacent grids. Some spurious wavelengths were created but were filtered out later. Merging large compilations with small compilations includes a risk, because the adjustment is better constrained by the large compilation even if the small one has apparently a better quality. Before merging the grids, compilations with comparable grid sizes were generated and more weight was given to the grid showing the best statistical properties. Therefore, French, Italian, and Spanish grids were first merged together, followed by building a second grid, the Finland, Fennoscandian, European, and Austrian grids. These two new compilations were merged together and the grids from Russia, Eurasia, Middle East, Antarctica, and North America were successively added to the compilation. The remaining grids that did not overlap with the grids listed above were simply added in the corresponding geographical areas. This generated a 3′ × 3′ global grid using GMT.

Magnetic, Global Anomaly Map, Table 2 Statistics after preprocessing of gridded and track line data (Courtsey: Hamoudi et al., 2007)

Grid name	Number of points in the grid	Minimum anomaly intensity	Maximum anomaly intensity	Standard deviation	Arithmetic mean
Africa and South America	65,030	−1,032	701	66	3.E − 08
Antarctica	969,479	−894	2,113	128	6.E − 05
Arctic	2,225,311	−2,604	5,528	171	−5.E − 05
Argentina inland	571,585	−3,150	1,480	107	−1.E − 07
Argentina margin	44,056	−137	318	44	6.E − 09
Australia	23,636,550	−8,743	18,882	209	8.E − 04
Austria	4,240	−55	137	27	0.E + 00
Canary Islands	328,077	−400	456	87	1.E − 05
India	625,898	−4,001	7,109	102	8.E − 06
Eurasia	1,758,876	−916	1,034	123	4.E − 05
Europe	664,626	−1,712	7,983	179	1.E − 06
Fennoscandia	78,862	−526	1,254	158	6.E − 09
Finland	420,121	−2,110	3,974	240	2.E − 05
France	7,560	−111	251	29	0.E + 00
Italy	36,603	−747	1,402	79	2.E − 09
Mexico	79,699	−562	876	86	−1.E − 07
Middle East	4,915,299	−5,446	2,408	101	5.E − 04
North America	63,447,385	−22,567	26,232	191	3.E − 02
Russia	1,061,053	−1,046	9,545	349	5.E − 05
South Africa	12,324,191	−14,642	11,341	176	−1.E − 04
East Asia	3,554,249	−340	2,196	85	6.E − 05
Spain	81,501	−97	242	20	0.E + 00
Magnet Project	7,814,111	−2,000	2,000	89	0.E + 00
Marine Data	19,455,835	−2,000	2,000	148	1.E − 04

Finally, the spurious intermediate wavelengths introduced by the merging were filtered out by a spherical harmonic transformation. The wavelengths larger than those corresponding to spherical harmonic degree 100 were filtered out of the global grid and were replaced with those from the downward-continued MF5 model. The candidate model along with accompanying documentation is available at http://www.earthref.org.

The Leeds model

To check for inconsistencies in the gridded data, the strength of total field intensities across the two overlaying grids are compared. Within the overlapping regions of Eurasia and Europe, total field intensities disagree in strength by 10–200 nT. At the grid boundaries, these intensities vary by 100–150 nT. These differences indicate possible errors in the long-wavelength content of the two grids. The discrepancies could also be attributable to the relative difference in the anomaly strength and resolution of the original grid. These differences are also observed between the Arctic and the North American compilations. Continental and oceanic compilations also show discontinuities in many regions.

For minimizing differences in the anomalies between various grids, a low-pass Gaussian filter with a cutoff wavelength of 400 km is applied in the 2-D space domain using GMT software (Wessel and Smith, 2004). The cutoff wavelength is set to 400 km because the longer wavelengths are later substituted from the MF5 model, which is robust up to spherical harmonic degree 100 (~400 km). The filtered output is a long-wavelength grid (>400 km) and strongly resembles a satellite-derived magnetic anomaly map. The short-wavelength grid is computed by subtracting the long wavelengths from the original grid and show finer details of the regional geology and structure of the crust (Figure 3).

To analyze the systematic errors in the long wavelengths between adjacent overlapping grids, the difference in the long wavelengths are computed. A visual inspection shows differences in the anomalies exceeding hundreds of nT and often show a lack of continuity across two or more grids. Because, it is difficult to judge the trustworthiness of an overlapping long-wavelength grid, it was best suited to replace the filtered long wavelength with a reliable satellite model.

Techniques that merge two adjacent compilations often employ a weighted average of different-order surfaces such that the two adjoining maps are seamless within the merging area. These processes modify the grids within the overlaying regions and near the edges and alter the strength and pattern of anomaly features. In this process, anomalies are modified, or at times genuine crustal anomaly features are even entirely eliminated. The merging technique applied to merge short-wavelength magnetic anomaly data minimizes abrupt changes along boundaries of overlapping grids. Using the MOSAIC toolbox (http://www.esri.com/), the grid values within the overlapping regions were determined from Hermite Cubic functions (Franke, 1982) and the short-wavelength grids merged.

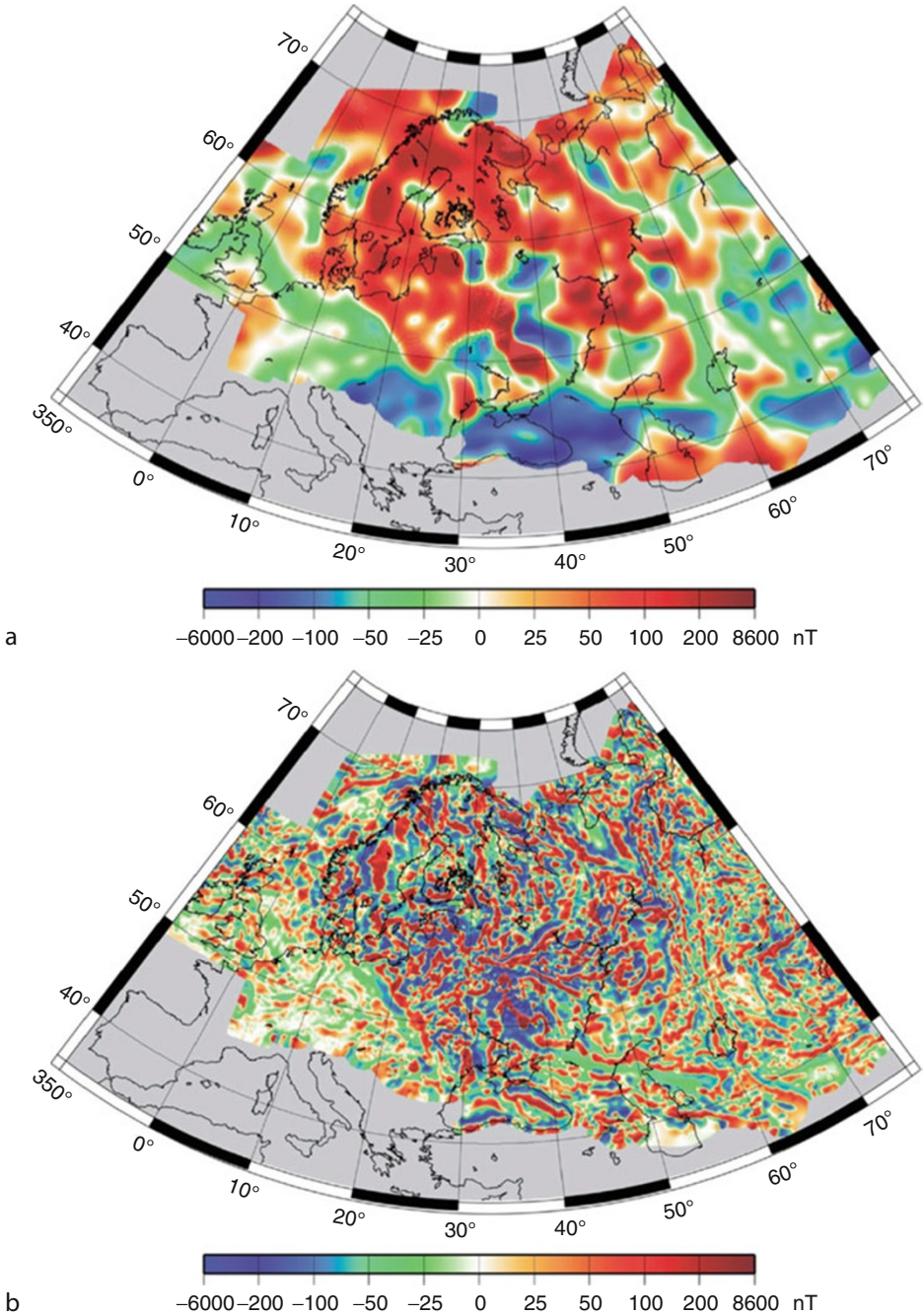
The short-wavelength anomaly differences between grids vary substantially; hence, it is difficult to prescribe the order of precedence for merging. However, the quality of an individual compilation is also reflected in the long-wavelength differences. Therefore, the order of precedence of the grid merging was based on the difference between long wavelengths and the quality of the parent grid. Thus, the short-wavelength grids were merged in the following order: Australia, North America, Austrian, Canary Island, East Asia, Eurasia, Europe, Arctic, Spain, Fennoscandia, France, Russia, Argentina margin and inland, South Africa, Antarctica, Italy, India, Middle East, Africa and South America, NGDC track line, and Project Magnet data. The grids were merged beginning with the highest precedence. The accuracy and reliability of the merged grid is assessed by investigating the effect of merging on the continuity of anomaly patterns across two or more grids and the effects on the anomaly shapes and strengths within and near the grid boundary. The differences for the merged European and Eurasian grid ranged from ± 100 to 200 nT and $< \pm 100$ nT for the remaining grid. A small difference of 8 nT was noted between Australian and East Asian grids indicating consistent anomaly patterns between the pair.

Finally, the total field anomalies of the short-wavelength-merged magnetic anomaly map were added to the downward-continued total field anomaly of the MF5 model to produce the candidate model of the LEEDS-WDMAM. The candidate model is archived at http://earthref.org/cgi-bin/ado.cgi?n=736&dbms=ERDA.

Earth magnetic anomaly grid (EMAG2)

As an improvement to the WDMAM, track line data were added in the southern oceans, where the data gaps were significant. For WDMAM, these, gaps in the ocean were filled with magnetic anomalies computed from ocean age model and geomagnetic polarity time scale. EMAG2 (Maus et al., 2009) was extrapolated into the unsurveyed areas using directional gridding based on the ocean age model of Müller et al. (2008). To further enhance the accuracy of the model over oceans, instead of using precompiled ocean magnetic anomaly grids, original track line data were used where available. Line-leveling algorithm was the primary method to minimize crossover errors. The new Arctic compilation (Circum-Arctic Mapping Project) was used instead of the old Arctic compilation. A new crustal magnetic field model derived from CHAMP data, the MF6 model (Maus et al., 2008), is used for EMAG2 to correct the long-wavelength anomaly field (wavelengths > 330 km) replacing the MF5 model used for WDMAM model.

EMAG2 is 2 arc-min grid defined at a height of 4 km above the geoid. The ocean data includes NGDC's GEODAS marine data archive, Antarctic Digital Magnetic Anomaly Project data (Golynsky et al., 2001), and

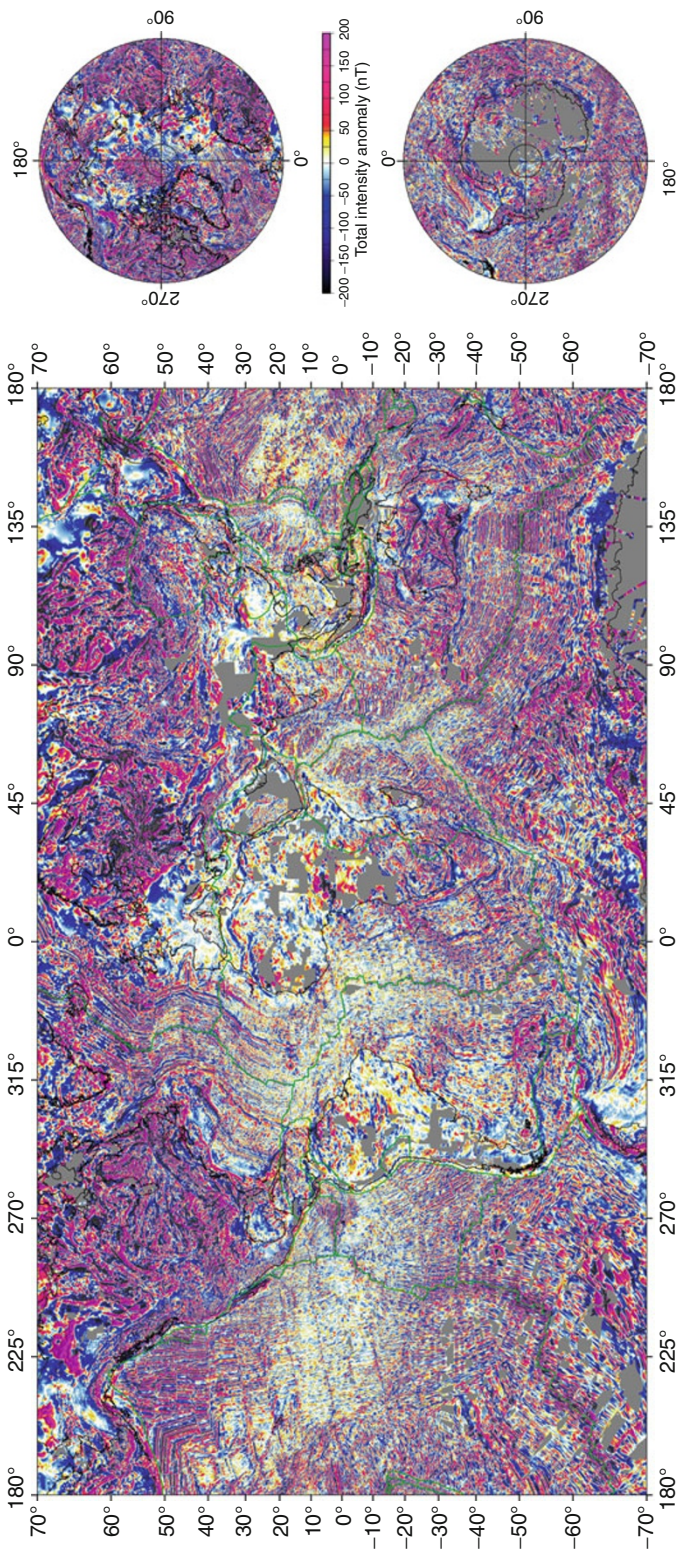


Magnetic, Global Anomaly Map, Figure 3 The effect of Gaussian filtering on the European magnetic compilation is demonstrated. Long-wavelength anomaly features (wavelengths greater than 400 km) in (**a**) resemble a typical satellite-based lithospheric magnetic field model at satellite altitude. The short-wavelength anomalies shown in (**b**) reflect geologic and tectonic features in detail. (After Hemant et al., 2007.)

the Project Magnetic Airborne data of the Naval Research Lab (NRL).

The line-leveling algorithm remains the same as described earlier. However, the search radius was reduced from $R_s = 100–8$ km in order to minimize the crossover effects in processing the track line data. In addition, the number of correction coefficients per track is now $N_i = \text{trunc}\,(X_i/300\ \text{km}) + 1$. These efforts reduced the crossover errors from 92 to 70 nT and reduced the RMS misfit to the merged grid from 121 to 97 nT.

The least-squares collocation method requires estimation of the correlation function. No new correlation



Magnetic, Global Anomaly Map, Figure 4 The Earth Magnetic Anomaly Grid (EMAG2) shown in Mercator projection and polar stereographic projection. (Courtesy: Maus et al., 2009.)

functions were computed for land areas as the gridded compilation remained the same. Notable changes for oceans, however were made. For instance, for Australia, 100 m terrain clearance was used. Due to the inclusion of more accurate grids and original marine and airborne surveys, the assumed variance V_0 was reduced from 40,000 to 33,000 nT and the correlation length r_c from 15 to 14 km.

The methodology of LSC was slightly modified for EMAG2. The correlation analysis now considered the anisotropy of oceanic magnetic anomalies. Track line data, just the way it was processed for NGDC candidate model, were first divided into small windows but along the constant heading direction of the track. Next, for each track, the azimuth for every pair of measurement points was calculated and compared with the direction of the oceanic isochrons at the locations of both points. The pair was removed from further processing if they met one of the following criteria: (1) the azimuth of the isochrons was not well defined, (2) the azimuth differed by more than 5° between the points, or (3) the topographic gradient between the two points exceeded 3%. The NGDC ETOPO-2 bathymetry values were used to compute the topographic gradient. This gradient could detect and remove the magnetic anomalies due to sea-mounts, which otherwise would appear as noise. The anisotropic correlation function was then computed in 10° directional bins; from 0° (parallel) to 90° (perpendicular, i.e., in the spreading direction) to the isochrons. This anisotropic correlation model allowed the LSC method to be used as a directional gridding algorithm.

From the analysis of covariance function in the isochron direction, it was noted that the function value reduced too rapidly. Furthermore, extrapolating the field far from a track line, made the function reduce to zero. This produced erroneous stripes parallel to the track lines. To stabilize the covariance function, the correlation length was doubled when gridding the ocean areas. This made the anisotropy of the oceanic field in EMAG2 to be smoother than it perhaps is.

EMAG2 also suffers inaccuracies from the oceanic crustal age model due to errors in the estimated ocean isochrons. This error could easily have contaminated those estimated oceanic magnetic anomalies of EMAG2, where the uncertainties in isochron age are large. To reduce these uncertainties, empirically, the anisotropy factor was increased four-fold for all oceanic isochrons of age <140 Ma and a linear reduction to a factor 2.25 for ages >150 Ma was used. This procedure helped reduce the errors in EMAG2 within the oceans, in particular over the older crust.

For compiling the final EMAG2 grid (Figure 4), the portion of the original total field anomaly up to degree 120 is replaced by that of the MF6 model. A spectral comparison between the original and modified grids shows a slight discrepancy above degree 120. This is perhaps due to the leakage of long-wavelength power into shorter wavelengths. But the overall comparison generally validates the power content in EMAG2. The EMAG2 grid can be downloaded from http://earthref.org/cgi-bin/er.cgi?s=erda.cgi?n=970 or from http://geomag.org/models/EMAG2.

World digital magnetic anomaly map 2011

The WDMAM is an ongoing project, with plans to produce updates of the map and grid every 4 years. As part of these updates, it is envisaged that more organizations will make additional data available for WDMAM 2011, in particular over regions where large data gaps exist. The call for candidate models for WDMAM 2011 has been announced. WDMAM 2011 will represent the total intensity of the magnetic field at an altitude of 2.5 km above the geoid at a resolution 2.5 km. The teams interested in deriving the candidate models are required to register by April 2010 and submit their models by October 2010. The WDMAM executive committee will choose one candidate as a basis for WDMAM 2011, and will define necessary adjustments required before submitting the manuscript to the CGMW. The final manuscript will consist of a grid file and a map image plus a short explanation. The CGMW will organize a scientific review of the submitted WDMAM 2011 manuscript and will print the map at a 1:50 million scale. Finally, the selected candidate model will be released as the WDMAM 2011 map at the IUGG General Assembly in Melbourne, Australia in July 2011.

Summary

In summary, with methods discussed above, it will be possible to regularly update the WDMAM model without excessive effort to reliably synthesize a map that accurately represents the world's magnetic anomalies. This would also require information on the long-wavelength field in order to continuously improve representation of the world magnetic anomalies. The upcoming Swarm mission of the European Space Agency (http://www.esa.int/esaLP/LPswarm.html) will considerably improve the lithospheric magnetic anomaly maps by providing highly accurate measurements of magnetic field gradients and further improve the long-wavelength anomalies in next generations of WDMAMs and EMAGs. It is our hope that these global magnetic anomaly maps will enjoy widespread usage, and inspire further aeromagnetic and marine surveys, thus helping to produce maps having truly global coverage. Finally, the WDMAM would not have existed without the willingness of institutions to constantly revise and release their national-scale compilation.

Bibliography

Bankey, V. A., Cuevas, D., Daniels, A. A., and Finn, I., 2002. Hermandez and Project Members, Digital Data Grids for the Magnetic Anomaly Map of North America, USGS (2002), Open-File Report 02-414.

Chiappini, M., Meloni, A., Boschi, E., Faggioni, O., Beverini, N., Carmisciano, C., and Marson, I., 2000. On shore-off shore

integrated shaded relief magnetic anomaly map at sea level of Italy and surrounding areas. *Annali di Geofisica*, **43**, 983–989.
Franke, R., 1982. Smooth interpolation of scattered data by local thin plate splines. *Computers and Mathematics with Applications*, **8**(4), Pergamon, pp. 273–281.
Gee, J. S., and Kent, D. V., 2007. Source of oceanic magnetic anomalies and the geomagnetic polarity time scale, Chap. 12. In *Geomagnetism: Treatise on Geophysics*. Kono, M. (ed.), Vol. 5. Amsterdam: Elsevier.
Golynsky, A., et al., 2001. *ADMAP – magnetic anomaly map of the Antarctic, Vol. 10, scale 1:10, 000, 000*. Cambridge: British Antarctic Survey.
Hamoudi, M., Thebault, E., Lesur, V., and Mandea, M., 2007. GeoForschungsZentrum Anomaly Magnetic Map (GAMMA): a candidate model for the World Digital Magnetic Anomaly Map. *Geochemistry, Geophysics, Geosystems*, doi:10.1029/2007GC001638.
Hemant, K., Thebault, E., Mandea, M., Ravat, D., and Maus, S., 2007. Magnetic anomaly map of the world: merging airborne, marine and ground-based magnetic data sets. *Earth and Planetary Science Letters*, **260**, 56–71, doi:10.1016/j.epsl.2007.05.040.
Kent, D. V., and Gradstein, F. M., 1986. Jurassic to recent chronology. In Vogt, P. R., and Tucholke, B. E. (eds.), *The Western North Atlantic Region, Geology of North America Volume M*. Boulder: Geological Society America, pp. 45–50.
Korhonen, J. K., Fairhead, J. D., Hamoudi, M., Hemant, K., Lesur, V., Mandea, M., Maus, S., Purucker, M. E., Ravat, D., Sazonova, T., and Thébault, E., 2007. *Magnetic Anomaly Map of the World – Carte des anomalies magnétiques du monde, Scale: 1:50,000,000l*, 1st edn. Commission for the Geological Map of the World.
Langel, R. A., and Hinze, W. J., 1998. *The Magnetic Field of the Earth's Lithosphere: The Satellite Perspective*. Cambridge: Cambridge University Press.
Lesur, V., and Maus, S., 2006. A global lithospheric magnetic field model with reduced noise level in Polar Regions. *Geophysical Research Letters*, **33**, L13304.
Maus, S., Luehr, H., Martin, R., Hemant, K., Balasis, G., Ritter, P., and Claudia, S., 2007a. Fifth-generation lithospheric magnetic field model from CHAMP satellite measurements. *Geochemistry, Geophysics, Geosystems*, **8**, Q05013, doi:10.1029/2006GC001521.
Maus, S., Sazonova, T., Hemant, K., Fairhead, J. D., and Ravat, D., 2007b. National Geophysical Data Center candidate for the World Digital Magnetic Anomaly Map. *Geochemistry, Geophysics, Geosystems*, doi:10.1029/ 2007GC001643.
Maus, S., Yin, F., Luehr, H., Manoj, C., Rother, M., Rauberg, J., Michaelis, I., Stolle, C., and Muller, R. D., 2008. Resolution of direction of oceanic magnetic lineations by the sixth-generation lithospheric magnetic field model from CHAMP satellite magnetic measurements. *Geochemistry, Geophysics, Geosystems*, **9**, Q07021, doi:10.1029/2008GC001949.
Maus, S., Barckhausen, U., Berkenbosch, H., Bournas, N., Brozena, J., Childers, V., Dostaler, F., Fairhead, J. D., Finn, C., von Frese, R. R. B., Gaina, C., Golynsky, S., Kucks, R., Lühr, H., Milligan, P., Mogren, S., Müller, R. D., Olesen, O., Pilkington, M., Saltus, R., Schreckenberger, B., Thébault, E., and Caratori Tontini, F., 2009. EMAG2: A 2–arc min resolution Earth Magnetic Anomaly Grid compiled from satellite, airborne, and marine magnetic measurements. *Geochemistry, Geophysics, Geosystems*, **10**, Q08005, doi:10.1029/2009GC002471.
Minty, B. R. S., Milligan, P. R., Luyendyk, A. P. J., and Mackey, T., 2003. Merging airborne magnetic surveys into continental-scale compilations. *Geophysics*, **68**, 988–995.
Moritz, H., 1980. *Advanced Physical Geodesy*. Karlsruhe: Herbert Wichmann Verlag.
Müller, R. D., Roest, W. R., Royer, J.-Y., Gahagan, L. M., and Sclater, J. G., 1997. Digital isochrons of the world's ocean floor. *Journal of Geophysical Research*, **102**, 3211–3214.
Müller, R. D., Sdrolias, M., Gaina, C., and Roest, W. R., 2008. Age, spreading rates, and spreading asymmetry of the world's ocean crust. *Geochemistry, Geophysics, Geosystems*, **9**, Q04006, doi:10.1029/2007GC001743.
Nakagawa, I., Yukutake, T., and Fukushima, N., 1985. Extraction of magnetic anomalies of crustal origin from Magsat data over the area of Japanese Islands. *Journal of Geophysical Research*, **90**, 2609–2615.
Purucker, M., and Whaler, W., 2007. In Kono, M. (ed.), *Crustal Magnetism*, Chap. 6, Vol. 5, Elsevier, *Treatise on Geophysics*, pp. 195–237.
Qureshy, M. N., 1982. Geophysical and Landsat lineament mapping – an approach illustrated from west central and south India. *Photogrammetria*, **37**, 161–184.
Ravat, D., Whaler, K. A., Pilkington, M., Sabaka, T., and Purucker, M. E., 2002. Compatibility of the high-altitude aeromagnetic and satellite-altitude magnetic anomalies over Canada. *Geophysics*, **67**, 546–554.
Reeves, C., 2005, Aeromagnetic surveys: Principles, practice and interpretation, published by GeoSoft, p. 155. http://www.geosoft.com/resources/news/2009/090723.asp.
Sabaka, T. J., Olsen N., Langel, R. A., 2002. A comprehensive model of the quiet-time, near Earth magnetic field: phase 3, *Geophysical Journal International*, **151**, 32–68.
Sabaka, T. J., Olsen, N., and Purucker, M. E., 2004. Extending comprehensive models of the Earth's magnetic field with Oersted and CHAMP data. *Geophysical Journal International*, **159**, 521–547, doi:10.1111/j.1365-246X.2004.02421.x.
Snyder, J. P., 1987. Map projections – a working manual, US Geological Survey. Professional Paper, Vol. 1395.
Socias, I., and Mezcua, J., 1996. Levantamiento aeromagneti-co del archipielago canario, Publ. Tec. 35, 28pp, Inst. Geogr. Nacl., Madrid.
Socias, I., Mezcua, J., Lynam, J., and Del Potro, R., 1991. Interpretation of an aeromagnetic survey of the Spanish mainland. *Earth and Planetary Science Letters*, **105**(1–3), 55–64.
Verhoef, J. R., Macnab, R., Roest, W., and Arkani-Hamed, J., 1996. Geological Survey of Canada, Magnetic anomalies of the Arctic and North Atlantic Oceans and Adjacent Land Areas. Open file 3125a.
Wessel, P., and Smith, W. H. F., 1991. Free software helps map and display data, Eos Trans., AGU, **72**, 441.
Wessel, P., and Smith, W. H. F., 2004. The Generic Mapping Tools (GMT), version 4, Technical Reference & Cookbook, SOEST/NOAA.
Wonik, T., Trippler, K., Geipel, H., Greinwald, S., and Pashkevitch, I., 2001. Magnetic anomaly map for Northern, Western and Eastern Europe. *Terra Nova*, **13**, 203–213.

Cross-references

Continental Drift
Continental Rifts
Core Dynamo
Earth's Structure, Continental Crust
Geomagnetic Field, Global Pattern
Geomagnetic Field, IGRF
Geomagnetic Field, Measurement Techniques
Geomagnetic Field, Polarity Reversals
Geomagnetic Field, Secular Variation
Geomagnetic Field, Theory
Lithosphere, Continental: Thermal Structure
Lithosphere, Oceanic
Magnetic Anomalies, Interpretation

Magnetic Data Enhancements and Depth Estimation
Magnetic Domains
Magnetic Methods, Airborne
Magnetic Methods, Satellite
Plates and Paleoreconstructions
Remanent Magnetism

MAGNETOMETERS

Ivan Hrvoic
GEM Systems, Inc, Markham, ON, Canada

Synonyms

Geomagnetic Instruments

Definition

Magnetometers are devices/instruments that measure magnetic field/magnetic flux density in particular Earth's magnetic field either vectorially or scalarly.

Introduction

First measurements of magnetic field intensity have been introduced by Carl Friedrich Gauss (Kaufman et al., 2009) in 1834. He has established a number of magnetic observatories to continuously measure magnetic field of Earth (see *Geomagnetic Field, Measurement Techniques*; *Geomagnetic Field, Secular Variation*). Prior to that, compasses were in use for several centuries mostly for navigation but with no knowledge of the causes of magnetic needle behavior. Since Gauss' times, many methods of magnetic field measurement have been developed but there are two pillars of magnetometry discovered and developed in the twentieth century:

1. Fluxgate magnetometers for measurement of components of the vector of magnetic field (Primdahl, 1979; Korepanov et al., 2007)
2. Scalar, quantum magnetometers with high precision of measurement and a possibility of measurements in motion (Abragam, 1961; Alexandrov and Bonch-Bruevich, 1992; Hrvoic, 2004; Hrvoic, 2008)

Presently, absolute measurements of magnetic field direction by optical methods (DI-Flux or DIM) based on fluxgate technology and the scalar measurements (Proton precession, Overhauser or Potassium) of its magnitude set up the limits of precision of the magnetic field vector measurement.

Magnetometry is a mature science/technology with myriad of applications. The need for better, more precise, smaller magnetometers fuels continuous research in the field.

This review will be treating separately vector and scalar magnetometers describing their principles of operation and main features, and with the view of their application in different fields of geophysical exploration.

Scalar magnetometers with their high precision and absolute accuracy now dominate measurements in mineral and oil exploration, volcanology, ordnance detection, archeology and partially magnetic observatories while vector magnetometer's role is limited to magnetic observatories, partially space measurement (Acuña, 2002) (see *Equatorial Electrojet*; *Magnetic Methods, Satellite*), and in vertical gradiometers for archeological research (see *Archaeomagnetism*). Scalar magnetometers are used to calibrate fluxgates. Some experiments with tensor measurements also include fluxgates and/or SQUIDs (Supercooled Quantum Interference Devices) (Clark, 1993).

All modern magnetometers are computerized instruments with nonvolatile memory for display, storage, and review of data.

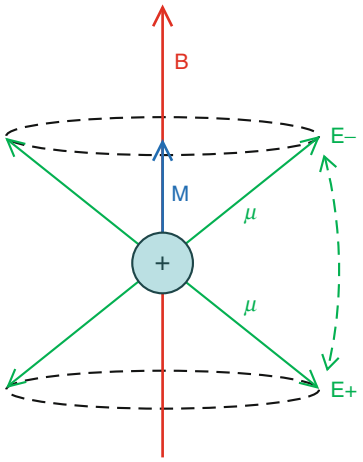
Scalar (quantum) magnetometers

Advent of scalar magnetometers in the second part of the twentieth century made a substantial contribution to measurement of both vector and scalar values of magnetic field. They offer precession frequency as a measure of magnetic field. Scalar magnetometer's sensitivities are determined only by achievable quality of precession signal, value of gyromagnetic constant, width of the spectral line (or time of decay of the precession signal), and signal-to-noise ratio. Scalar magnetometers have high sensitivity and accuracy of readings, virtually no drift with temperature or time. Measurements depend very weakly on sensor orientation or movement allowing for measurement in motion.

Proton, Overhauser, and Alkali metal (optically pumped) magnetometers are now overwhelmingly used in mineral/diamond/oil exploration, weapons detection, volcanology, archeology, magnetic observatories. Most of the current research is done in variations of optical pumping, and the progress in the last 25–30 years is phenomenal. Sensitivities have been improved from some *nT* to fractions of *pT*, perhaps four orders of magnitude, far exceeding requirements for ground (see *Magnetic Methods, Surface*), airborne (see *Magnetic Methods, Airborne*), or marine surveys. Speed of readings increased from perhaps once per second to tens and hundreds of readings per second limited only by increasing noise and/or possibility to usefully store a flood of data. With the development of instrumentation, the use of gradiometers with two or more sensors has increased (see *Magnetic Gradiometry*). It improves the quality of surveys, determination of depth of the anomaly producing body, following the direction and depth of the buried pipelines or electrical power lines etc.

Scalar magnetometer consists of a sensor, separated by a cable from electronics (to avoid its stray magnetic fields). Electronics has an analogue part that generates precession signal from the sensor and digital microprocessor based on part that controls the operation; measures Larmor frequency; converts it into units of magnetic field; and displays, stores, and/or outputs the data. Review of data is often possible (Figure 1).

Magnetometers, Figure 1 Overhauser magnetometer – sensor, cable and electronics console.



Magnetometers, Figure 2 Magnetic moment μ of a proton (or electron) precesses around magnetic induction *B*. Resulting magnetization *M* is collinear with *B* and static. Components orthogonal to *B* are averaged out. Coupling energy $E = -\mu B$.

Global positioning plays a big role in geophysical ground and airborne surveys. In stationary measurements, GPS provides precise timing (1μs accurate pulses referenced to Greenwich standard time).

Background physics

Scalar magnetometers are all based on a spin of subatomic particles – electrons and protons (Abragam, 1961; Schumacher, 1970).

Spinning charged particles make magnetic dipoles. The dipoles are precessing around ambient (applied) magnetic field following Quantum Physics rules (Figure 2).

Precession frequency is proportional to magnetic flux density:

$$\omega_0 = \gamma B$$

ω_0 is an angular precession frequency (Larmor frequency), B magnetic induction (flux density), and γ gyromagnetic constant.

Spinning protons (electrons) orient themselves in magnetic field. Only two angles of orientation are allowed: acute (lower energy level) and obtuse (higher energy level).

Since individual particles precess at random phases, their dipolar magnetic field in the plane perpendicular to the magnetic field is averaged out. A projection of the dipolar magnetic fields in the direction of the applied magnetic field is static. Since orientation at the acute angle to the direction of the field is more populated than the one at the obtuse angle, a minute "polarization" of precessing particles creates a magnetization M_0 collinear with magnetic induction B:

$$M_o = \frac{N\gamma^2 h^2 / 4\pi^2 B}{4kT\mu_o}$$

N is the number of particles, γ gyromagnetic constant, h Planck's constant, T absolute temperature, k, μ_0 constants.

When placed in the flux density B, M_0 will reach its "thermal equilibrium" exponentially with the "longitudinal" time constant T_1.

When deflected by about 90°, it will precess around the field by the Larmor frequency and decay exponentially by the "transversal" time constant T_2. (Time constants T_1 and T_2 depend on the aggregate state of the assembly of spinning particles. In liquids and vapors, they may be several seconds long, while in solids, T_2 time is only in milliseconds and the precession signal disappears very quickly. This is why all scalar magnetometers have liquid or gaseous sensors.) Particles at the lower energy level can accept energy in the form of magnetic field at Larmor frequency and flip to the higher energy level. Leveling of the two energy level populations eliminates M_0. This is saturation of the precession spectral line. Proportionality constant, the gyromagnetic constant (not always a constant), is precisely determined to better than one part per million accuracy only for protons in water (Hrvoic, 1996):

$$\gamma_\rho = 0.2675153362 \mathrm{rad}/nT$$

Magnetization M_0 is too small to produce detectable precession signal when deflected in the plane of precession. All scalar magnetometers therefore need to increase polarization and magnetization M and thus increase the sensitivity of the instrument. Polarization is increased in different ways:

- By temporarily increasing B to few hundreds Gauss (Proton magnetometers)
- Transferring thermal equilibrium polarization of electrons to protons in liquid sensors (Overhauser magnetometers)
- Using light polarization in alkali metals, and ^{4}He magnetometers

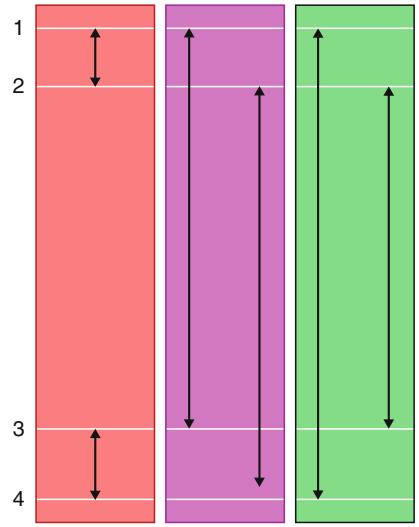
Proton magnetometers

Proton magnetometers are the oldest scalar magnetometers. The first commercial units were produced in early 1960s as portable instruments. In continuation, airborne instruments appeared with optimized speed of readings and sensitivity, large sensors, etc. Later development of Overhauser and optically pumped magnetometers has eliminated Proton magnetometers from airborne surveys. However, they remain very popular in various ground surveys and observatories.

Proton magnetometer's sensor contains liquid rich in protons and polarization/pick-up coil. Polarization is done by increasing flux density B for a time comparable with T_1 of the sensor liquid. Electrical current passing through polarization coil creates strong magnetic field that polarizes protons of the sensor liquid. The coil is usually immersed in liquid to maximize the coupling. Alternatively, omnidirectional toroidal coil is used. Polarization field of few hundred Gauss must be roughly at the right angle to the measured field. Its removal must be fast to leave the newly formed magnetization in the plane of precession. Practical portable Proton magnetometers achieve about $0.1nT$ sensitivity at a few seconds cycle, somewhat less in once per second measurements.

Overhauser magnetometer

Overhauser effect originates from double resonance experiments in metals (Overhauser, 1953). Russian and French scientists have greatly contributed to the development of modern Overhauser magnetometers based on nitroxide free radicals (Abragam, 1961; Pomerantsev, 1968). Practical realization of the magnetometer is as follows: Sensor liquid containing protons is placed in an RF resonator. It has small concentration of a Nitroxide free radical – stable chemical with one unpaired electron. The unpaired electron is dwelling close to the nitrogen nucleus, in its magnetic field of some 24 Gauss. Its electron paramagnetic resonance frequency in the Earth's magnetic field is about 30,000 times higher than the one of protons. Unpaired electrons couple with the protons of the sensor, creating a four energy level system (Figure 3). Coupling can be either scalar or dipole – dipole. If one of them predominates, saturating electron resonance at some 60 MHz, transfers part of electron magnetization to protons. Increased proton polarization (magnetization) needs a 90° deflection to be turned in the plane of precession. The deflection can be either pulsed ($\pi/2$ pulse) or stationary by applying a weak rotating magnetic field of Larmor frequency in the plane of precession. Transferred magnetization is far superior to any achievable by proton magnetometers and the measurement does not need to be interrupted for polarization; RF saturating field can be present continuously. In pulsed mode, repeated $\pi/2$ pulses interrupt measurement by about 25–30 ms. Low power consumption, no warm-up delay, higher sensitivity ($10pT$ for one reading per second), absolute accuracy, reasonable speed of readings (up to five readings per second), and omnidirectional sensors make Overhauser magnetometers very convenient and attractive for magnetic observatories, volcanological exploration, marine surveys, and base stations for airborne surveys. Oersted and CHAMP magnetic satellites use continuous Overhauser magnetometers as reference to vector measurement of magnetic field.



Magnetometers, Figure 3 Four energy levels combined electron–proton system. Transitions 1–4 (weak scalar coupling) and 2–3 (dipole-dipole coupling) involve combined transitions of electrons and protons. 1–3, 2–4, 1–2, and 3–4 are independent transitions of electrons and protons.

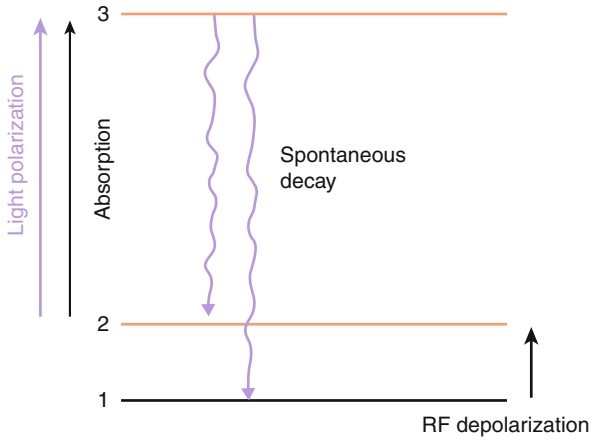
Optically pumped magnetometers

This is the latest to be discovered and the most potent family of magnetometers. It consists of elements of the first column of the table of chemical elements. In gaseous form, alkali metals have an unpaired electron in their valence shell.

Gaseous sensor is subject to a circularly polarized light (D_1 line) (Happer, 1972). Polarization process lifts electrons from a higher energy level to a metastable state from which electrons fall back to both levels.

Eventually, all electrons are in the lower energy level while the upper one is depleted. Absorption of light is then reduced and the sensor is more transparent. Rotating RF magnetic field of the Larmor frequency will depolarize the sensor and increase absorption of the light (Figure 4).

As a result, the polarizing light will be modulated by the Larmor frequency. Its detection produces an electrical signal for the measurement of magnetic field. Helium 4 magnetometers that have two valence electrons need a weak discharge to lift one of them in a metastable state. Laser-pumped Helium 4 magnetometer has sub pT sensitivity. Due to the influence of nucleus (strong scalar coupling with valence electrons), all Alkali metal magnetometers have a number of spectral lines in their EPR spectrum (Breit and Rabi, 1931).



Magnetometers, Figure 4 Optical polarization (lifting an electron from upper energy level to a metastable state) and depolarization (lifting electrons from lower to upper energy level) result in modulation of optical beam by Larmor frequency.

Potassium spectral lines are well spaced and the operation of Potassium magnetometer can be based on a single narrow spectral line. This ensures very high sensitivity and the absolute accuracy similar to the one of Proton/Overhauser magnetometers. Cesium and rubidium lines are very close to each other, and they overlap and make a relatively wide composite single line. Operation of the magnetometer in self-oscillating mode is at the peak of the composite line. Problem is that the position of the peak depends on an angle between sensor axis and magnetic field direction. A heading error due to this effect can be reduced by a "split beam" technique, which symmetrizes the composite line. Reduction of the heading error to some nT or even a fraction of nT is possible. However, the split beam technique reduces the sensitivity by about one order of magnitude. Absolute accuracy of only few nT can be achieved by Cs and Rb magnetometers.

Measurement of frequency

In the magnetic field of Earth (20–65μT), precession frequency of protons is an acoustic frequency in a range of some 850Hz–3kHz, while Alkali metal optically pumped magnetometers produce 70–450kHz, ^{4}He 280kHz to about 2MHz. Considering achieved precisions of measurement – one part per million or a small fraction of that for Proton and Overhauser magnetometers respectively and even few parts per billion for ^{4}He and ^{39}K or ^{41}K – we are resolving Larmor frequency to a millihertz precision. Times of all zero-crossings of precession frequency are taken, and an average period is measured by the least squares fit. Zero-crossing times must be understood as a phase information too (0°, 180°, 360°, etc.), and this allows measurement to the sub Hz precisions. Measurement of phase difference to obtain the frequency results in an unusual noise dependence on the speed of readings. Most of measurements concerned with the noise bandwidth specify noise as per square root of Hertz. Doubling the bandwidth increases noise by square root of two. Not so in precession frequency measurement. When we double bandwidth by doubling the number of readings per second the noise increases by two square roots of two or $2^{3/2}$ instead of $2^{1/2}$. Defining noise as per square root of Hertz is therefore deceiving, although it may be calculated correctly for a particular number of readings per second.

Another peculiarity of precession frequency measurement is a phenomenon of "outliers" or "spikes" that occur when a zero-crossing time is either missing or modified by a spike of phase noise usually showing lower value of magnetic field. Preponderance of outliers increases dramatically as signal/noise ratio of precession signal decreases.

Absolute accuracy and precision

Absolute accuracy of measurement can be determined directly for Proton (Overhauser) magnetometers only. Proton gyromagnetic constant gets periodically refined by national standards institutions of a number of major countries (NIST, USA; NPL, U.K; VNIIM, Russia; NIIM, China).

Absolute accuracy of better than one part per million is possible to achieve. Besides the precise gyromagnetic constant, it requires a number of conditions to be fulfilled: sensitivity must always exceed the absolute accuracy, frequency reference precision must be adequate, phase stability of the signal and time determination of a zero-crossing must be proper, chemical shift of the precession frequency of the sensor liquid must be taken into account (Hrvoic 1996). Since we measure frequency of rotating magnetic moment, rotation of the sensor in the plane of precession introduces error (rotational Doppler). This error, insignificant for optically pumped magnetometers, may be significant for Proton and Overhauser magnetometers (about 23nT offset for one cycle per second rotation of the sensor).

Alkali metal magnetometers have only potassium capable of high absolute accuracy (comparing its readings with Proton/Overhauser magnetometers) (Alexandrov and Bonch-Bruevich, 1992). Cesium and rubidium operate on lumped spectral lines, (and ^{4}He has a very wide spectral line) and their absolute accuracies are in few nT range.

Great majority of nowadays measurements do not require high absolute accuracy but only sensitivity and repeatability or relative accuracy of readings. Exceptions are magnetic observatories, space measurements, and perhaps some standards measurements derived from magnetic measurements (determination of electrical current standard for example).

Sensitivities of scalar magnetometers vary. Details depend heavily on particular design, but an order of magnitude can be assessed. Proton magnetometers are nowadays used for ground exploration or calibration of vector

magnetometers (fluxgates) only. Slow rate of readings can produce about $0.1nT$ sensitivity and absolute accuracy. Up to two readings per second are possible, but at this rate, the noise is in nT range. Overhauser magnetometers achieve $0.01nT$ at one reading per second, maximum rate five per second. High absolute accuracy Overhauser magnetometers are standard at magnetic observatories and very prominent in marine and ground surveys, and base stations for airborne surveys. Composite spectral line Alkali vapor magnetometers (Cesium, Rubidium) have similar sensitivity to Overhauser at one reading per second, but their top speed of readings is over 20/s. Their absolute accuracy is only few nT. Laser-pumped ^{4}He reaches sub pT sensitivities at one reading per second. Potassium has reached $0.05pT$ sensitivity at one reading per second. Practical potassium assemblies for ground and airborne surveys feature sub pT once per second and about $7–10pT$rms noise at ten readings per second. Like cesium and rubidium, maximum speed of readings exceeds 20/s. Increase of noise is rather a practical limit to a number of readings per second. Cesium magnetometers are prominent in airborne surveys while potassium covers top of the field, especially in gradiometry due to it's high absolute accuracy. Potassium is exclusively used in some new high sensitivity magnetic observatories and Earthquake study centers.

Vector magnetometers

Vector magnetometers are generally used for static measurements (observatories) or in vertical suspension (archeology) or rarely in the component airborne surveys where only computed total field may reach $0.1nT$ sensitivity, while components stay useless due to changes in aircraft attitude. In some recent attempt, accelerometers are used to account for attitude changes and apply corrections.

Main characteristics of most of the vector magnetometers are (except for SQUIDs):

1. Good sensitivity of measurement $0.1nT$ common, $0.01nT$ or somewhat lower (Korepanov et al., 2007) the best reported.
2. Very precise determination of the direction of measurement. One arc second change in direction means a fraction of one nT difference in readings, depending on an angle. In some orientations, vector magnetometers can resolve $0.01nT$ field and well below one arc second angle.
3. Relatively substantial temperature dependence ($0.1nT$/°C or more).
4. Relatively substantial time variation – aging ($2nT$ or more per year).

Consequences of the above are difficulties in orienting in the magnetic field and determination of the orthogonality of the three components. Due to their temperature dependence/aging, the vector magnetometers at the magnetic observatories need calibration by Proton/Overhauser magnetometers.

Fluxgate magnetometer

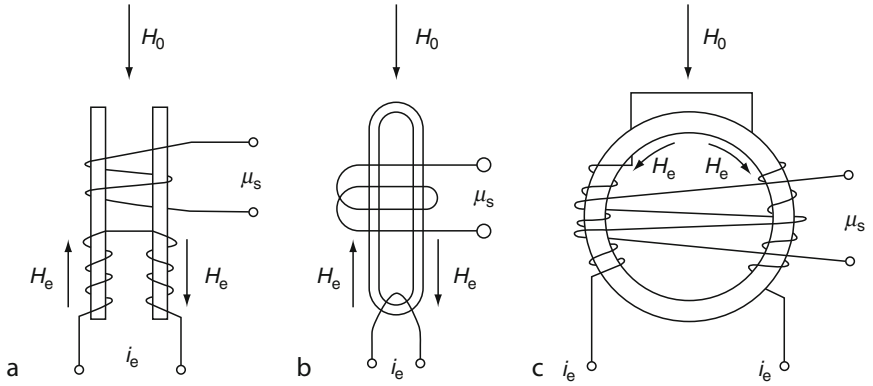
Main vector magnetometer today is the fluxgate. Its operation is based on nonlinearity of magnetic properties of steel (ferromagnetic material) (Primdahl, 1979). When placed in strong magnetic field, steel will lose its high magnetic permeability in saturation, i.e., relative permeability will fall from several hundreds to close to one. A coil is wound around a layered core of steel, and alternating current is driving the core to saturation in two opposite directions alternatively. When not saturated, the core will concentrate the magnetic flux of the applied magnetic field. When saturated, the flux will deconcentrate itself. Pick-up coil wound around the core will detect changes in the flux passing through the core. With no external magnetic field, the saturations on positive and negative sides will be symmetrical i.e. the pick-up voltage will have only odd harmonics. When the external magnetic field is present, the saturations become unsymmetrical and second harmonic appears in the pick-up voltage, its amplitude proportional to the strength of magnetic field. Phase detected second harmonic is then a measure of applied magnetic field in the direction of the core. It is convenient to use two pieces of the core with excitation windings in opposition and the pick-up coil encompassing both cores as this eliminates pick-up voltage due to the original driving current. Instead of double iron core, a ring or "racetrack" shaped cores can be used (Figure 5).

Today's fluxgates can achieve $0.1nT$/°C temperature coefficient with some nT per year aging. Composition of the iron cores plays a big role in the quality of fluxgates. Some recent new developments (Korepanov et al., 2007) claim better noise and aging characteristics.

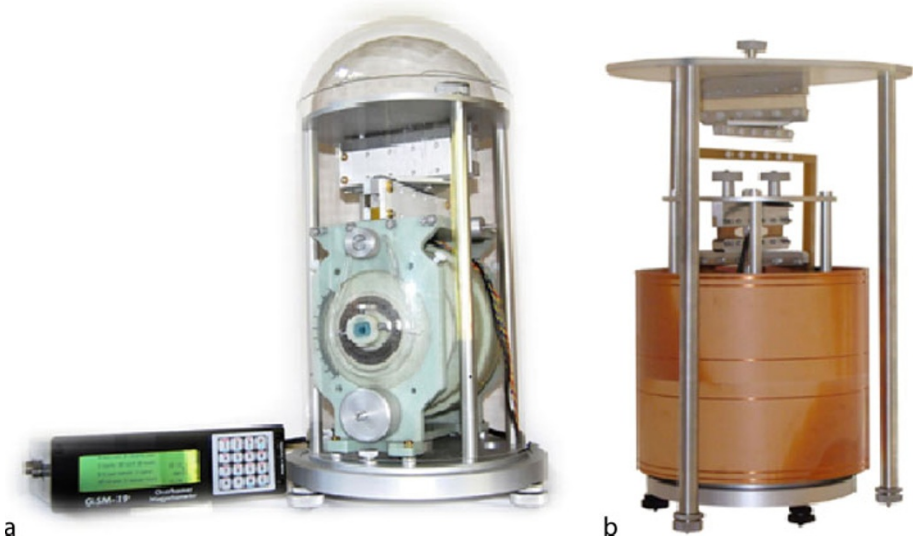
During the Second World War, a total field – oriented single fluxgate was used to detect submarines. ASQ-10 instrument had about $0.1nT$ sensitivity and the total field orientation of the sensor was kept by feedback electronics/mechanics.

Absolute measurement

Declination/Inclination magnetometer (DI-Flux or DIM) is a variation of fluxgate measurement (Jankowski and Sucksdorff, 1996). Fluxgate magnetometer sensor is mounted collinear with the optical axis of (nonmagnetic) theodolyte, in horizontal plane. Theodolyte is rotated, and two zero field angles are marked. The telescope is then reversed and two more zero readings determined. Magnetic north in arbitrary coordinate system is then calculated from angles of the four zero field measurements. The calculated angle is then referred to a known marker to determine true declination angle related to north direction. To determine inclination, the theodolyte is oriented in just determined magnetic declination vertical plane and four zero field positions are determined in that plane. Averaging will give the angle off horizontal, i.e., the inclination. Some attempts have been made (see *Geomagnetic Field, Measurement Techniques*) to automate absolute measurements.



Magnetometers, Figure 5 Excitation current drives the two cores (**a**) in saturation in opposite directions. Pick-up coil detects unsymmetry in saturation by measuring second harmonic. The cores can be connected at the ends to create "race track" (**b**) or ring core (**c**) shapes. Ring core allows for measurement of two orthogonal components, on one core.



Magnetometers, Figure 6 DIdD Overhauser/Potassium suspended system (**a**) and with external vertical coil to measure inclination directly (**b**).

Simplified version of DIM is the Declinometer. It has a suspended magnet on a torsionless fiber, a mirror attached to it at the right angle to its magnetic axis and in front of a telescope. The theodolyte is turned until the telescope becomes at right angle to the mirror. Direction of magnetic meridian is read from the base of the theodolyte. This is referenced to a known reference mark to determine real declination angle.

Weakness of this measurement is an angle mirror-magnetic axis of the magnet that has to be 90°. This is mitigated by using two magnets of different magnetic moments.

Quantum vector magnetometers

SQUID is a vector magnetometer of exceptional fT range sensitivity. It operates at cryogenic temperatures (liquid helium or liquid hydrogen) (Clark, 1993). It is used sporadically in geophysics for tensor calculations or short base gradiometers.

DIdD (Delta inclination delta declination) is a vector magnetometer with two orthogonal bias coils and a scalar magnetometer (Overhauser or Potassium) (Figure 6).

Coils are oriented at right angles to magnetic field direction, one in a horizontal plane, one in vertical magnetic declination plane. A sequence of four biased fields and one unbiased allows for determination of departures of the field from the preset one. Orientation of one bias coil in the vertical direction and the other in the horizontal East-West direction (now possible with the precise GPS direction determination or a North seeking gyroscope) will convert dIdD instrument into an absolute D and I instrument. Potassium magnetometer can provide

superior sensitivity and absolute accuracy. The stability of the instrument as for temperature and aging can be superior to the best fluxgates. This instrument is still in development. It is meant for use in the directional drilling for oil and minerals and possibly in a refined version for magnetic observatories.

Summary

Magnetic measurements are an essential part of geophysical exploration for minerals and oil, archeological, volcanological explorations, research of atmosphere, Sun's influence, crustal studies, earthquake research, etc. Scalar magnetometers are dominant in most fields. They are described in some detail – their physics, principles of operation, details of design, sensitivities and absolute accuracy. Vector magnetometers are essential in magnetic observatories, space measurements, and many scientific investigations. Fluxgate magnetometers and variations of absolute determination of magnetic field direction have been described in some detail. Quantum magnetometers as vector magnetometers are briefly described. Refinement of Potassium magnetometers may improve sensitivity of presently used scalar magnetometers by two orders of magnitude to sub *pT* range.

Bibliography

Abragam, A., 1961. *The Principles of Nuclear Magnetism*. New York: Oxford University Press.

Acuña, M. H., 2002. Space-based magnetometers. *The Review of Scientific Instruments*, **73**, 11.

Alexandrov, E., and Bonch-Bruevich, V., 1992. Optically pumped atomic magnetometers after three decades. *Optical Engineering*, **31**(4), 711.

Breit, G., and Rabi, I. I., 1931. Measurement of nuclear spin. *Physical Review*, **38**, 2082.

Clark, J., 1993. SQUIDs: theory and practice. In Weinstock, H., and Ralston, R. W. (eds.), *The New Superconducting Electronics*. Dordrecht: Kluwer Academic, pp. 123–180.

Happer, W., 1972. Optical pumping. *Review of Modern Physics*, **44**, 169.

Hrvoic, I., 1996. Requirements for obtaining high accuracy with proton magnetometers. In Rasson, J. L. (ed.), *Proceedings of the VIth Workshop on Geomagnetic Observatory Instruments Data Acquisition and Processing 70–72*. Also available online at www.gemsys.ca.

Hrvoic, I., 2004. *Short Review of Scalar Magnetometers* (unpublished) (available online at www.gemsys.ca).

Hrvoic, I., 2008. Development of a new high sensitivity Potassium magnetometer for geophysical mapping. *First Break*, **26**, 81–85. www.gemsys.ca.

Jankowski, J., and Sucksdorff, C., 1996. *Guides for Magnetic Measurements and Observatory Practice* (available online at IAGA iaga_sg@gfz-potsdam.de).

Kaufman, A. A., Kleinberg, R. L., and Hansen, R., 2009. *Principles of Magnetic Methods in Geophysics*. Amsterdam, London: Elsevier Press.

Korepanov, V., Klymovych Ye, Kuznetsov, Pristay, A., Marusenkov, A., and Rasson, J., 2007. NEW INTERMAGNET Fluxgate Magnetometer XII IAGA workshop. *Published by Institute for Geophysics of the Polish Academy of Science*, **v.C-99**(398), 291–298.

Körmendi, A., 2008. Private communication (available online at alpar44@freemail.nu).

Overhauser, W. A., 1953. Polarization of nuclei in metals. *Physical Review*, **92**, 411.

Pomerantsev, N. M., 1968. *Quantum Magnetometry*. Joint Publication Research Service JPRS 46444.

Primdahl, F., 1979. The fluxgate magnetometer. *Journal of Physics E: Scientific Instruments*, **12**, 241–253.

Schumacher, R., 1970. *Introduction to Magnetic Resonance*. New York: Benjamin WA.

Cross-references

Archaeomagnetism
Equatorial Electrojet
Geomagnetic Field, Measurement Techniques
Geomagnetic Field, Secular Variation
Magnetic Gradiometry
Magnetic Methods, Airborne
Magnetic Methods, Satellite
Magnetic Methods, Surface

MAGNETOTELLURIC DATA PROCESSING

Gary Egbert
College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, OR, USA

Definition

Magnetotelluric data processing: statistical estimation and analysis of impedance tensors and other transfer functions from measured electromagnetic time-series data.

Introduction

In the magnetotelluric (MT) method (see *Magnetotelluric Interpretation*), *data processing* generally refers to the initial analysis steps required to use naturally occurring time variations of electromagnetic (EM) fields measured at the surface to map the distribution of electrical conductivity within the Earth (see *Geoelectromagnetism*). The data are acquired as time series, generally including three components of the vector magnetic fields, along with the two components of the horizontal electric field. Earth conductivity variations have a relatively subtle effect on the raw EM data, which most directly reflect the temporal pattern of external source variations. Sophisticated data processing methods have been developed to extract the interpretable Earth conductivity signal from these data. Two distinct steps can be distinguished. In the first, time series processing methods are used to estimate frequency-dependent linear relations or transfer functions (TFs), relating magnetic and electric field components. In the second step, distortion and strike analysis, the full set of TFs is reduced to a simpler set, consistent with simplified one or two-dimensional models of Earth conductivity, typically allowing for shallow near-surface complications. After these data processing steps, TFs are inverted for conductivity, and the resulting images are interpreted in terms of composition, physical state, and Earth structure.

MT transfer functions

The basic assumption underlying the MT method is that the external source fields are spatially uniform, an approximation that can be physically justified provided the spatial scale of external magnetic fields at Earth's surface is large compared to the depth of penetration of the EM fields in the conducting Earth. Except at very long periods ($T > 10^4$ s), for which penetration depths in the Earth can exceed several hundred kilometers, and in the auroral zone where ionospheric current systems vary over short length scales, this assumption holds quite well for natural sources. Local cultural sources of EM noise have much shorter spatial scale, and will not always satisfy the uniform source assumption. These sorts of coherent noise, which violate the most fundamental MT assumption, represent the most serious obstacle to successful MT transfer function (TF) estimation.

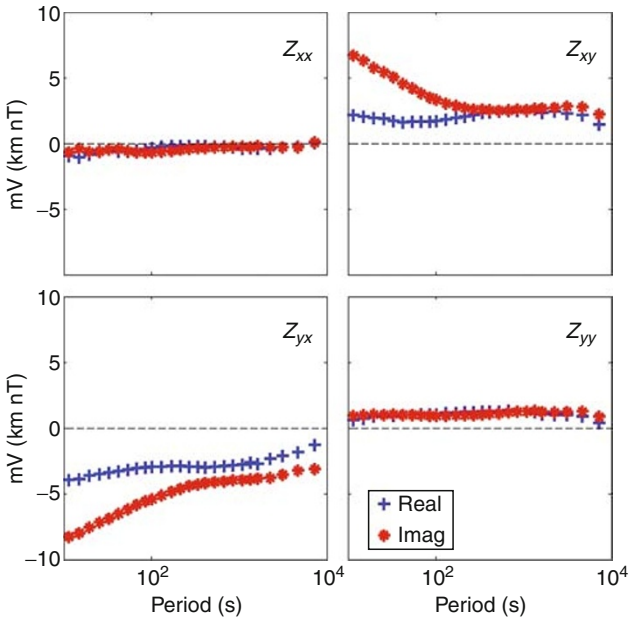
For periodic sources at a fixed frequency, the assumption of spatial uniformity implies that sources can be expressed as linear combinations of two simple modes: unit magnitude uniform sources linearly polarized North–South and East–West. This in turn implies a linear relationship (in the frequency domain) between the horizontal electric and magnetic field vectors measured at a single site.

$$\begin{pmatrix} E_x \\ E_y \end{pmatrix} = \begin{pmatrix} Z_{xx}(\omega) & Z_{xy}(\omega) \\ Z_{yx}(\omega) & Z_{yy}(\omega) \end{pmatrix} \begin{pmatrix} H_x \\ H_y \end{pmatrix} \tag{1}$$

The frequency-dependent 2×2 TF $Z(\omega)$ is referred to as the impedance tensor; see Figure 1 for an example. Equation 1 can be justified by the linearity of Maxwell's equations, and the assumption that all sources can be expressed as linear combinations of two linearly independent polarizations (Egbert and Booker, 1989). Under these circumstances, two independent field components (e.g., H_x and H_y at one location) uniquely determine all other EM field components at any nearby site. The uniform source assumption thus also justifies other sorts of TFs. In the other most important case, the vertical magnetic field component is related to the two horizontal components at the local site

$$H_z = \begin{pmatrix} T_x(\omega) & T_y(\omega) \end{pmatrix} \begin{pmatrix} H_x \\ H_y \end{pmatrix}. \tag{2}$$

Vertical field TFs, which are commonly estimated along with the impedance in most modern MT surveys, are referred to variously as the Tipper, Parkinson vector, or Wiese vector. Inter-station TFs, which relate horizontal magnetic fields at two sites, are also sometimes considered (e.g., Egbert, 2002).



Magnetotelluric Data Processing, Figure 1 Example of an impedance tensor, the 2×2TF relating horizontal electric and magnetic field components. Real and imaginary parts of the complex impedance components are plotted as a function of period. Note that the diagonal components of the impedance are relatively small; for the case of a 2D conductivity distribution, these components would vanish identically, if the impedance were expressed in the proper coordinate system.

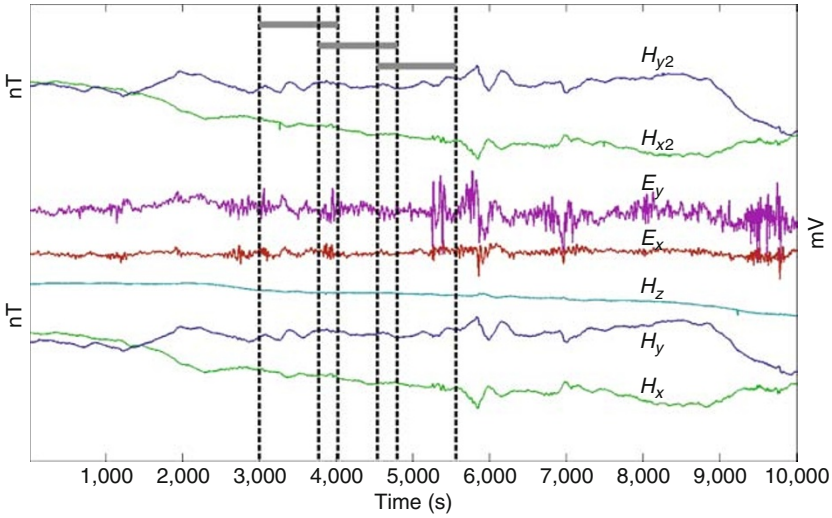
Least squares transfer function estimation

The earliest approaches to MT TF estimation were based on classical time series methods (e.g., Bendat and Piersol, 1971), applying a simple linear least squares (LS) fitting procedure in the frequency domain. To be explicit, consider estimating the MT TF between one component of the electric field (e.g., E_x) and the two horizontal magnetic field components H_x and H_y at a single fixed frequency ω. This TF corresponds to the first *row* of the usual MT impedance tensor. All of the discussion here and in subsequent sections applies equally to the second row of the impedance, and to the vertical field TF. Because all of the TFs are most succinctly described in the frequency domain as in (1–2), the first step in data processing is generally to Fourier transform the data. After possibly de-spiking and/or pre-whitening, time series for each component are divided into M short time windows, tapered, and Fourier transformed (Figure 2). Fourier coefficients for N frequencies in a band centered around ω are used for the TF estimate, for a total of $I = MN$ complex data. Estimates $\hat{\mathbf{Z}}$ of the TF (i.e., impedance elements) are then obtained for frequency ω by least squares (LS) fitting of the linear model

$$\begin{pmatrix} E_1 \\ \vdots \\ E_I \end{pmatrix} = \begin{pmatrix} H_{x1} & H_{y1} \\ \vdots & \vdots \\ H_{xI} & H_{yI} \end{pmatrix} \begin{pmatrix} Z_{xx} \\ Z_{xy} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_I \end{pmatrix}, \tag{3}$$

or in matrix notation

$$\mathbf{E} = \mathbf{HZ} + \varepsilon. \tag{4}$$



Magnetotelluric Data Processing, Figure 2 Example of time series collected with a long-period MT system. From bottom, three vector components of the magnetic field (measured with a fluxgate magnetometer), two colocated orthogonal horizontal electric field components (measured as the potential difference between pairs of electrodes separated by 100 m), and two horizontal components of the magnetic field at a distant remote site. Three overlapping data segments are indicated by the vertical dashed lines. Segments such as these are Fourier transformed to yield a series of complex Fourier coefficients for each data channel, which are then used for the various transfer function estimates discussed in the text. Longer segments are processed to get Fourier coefficients for longer period data, typically after low-pass filtering and decimating the time series.

With standard least squares (LS), this is accomplished by minimizing the sum of the squares of the residuals:

$$\sum_i |E_{xi} - (H_{xi}\hat{Z}_{xx} + H_{yi}\hat{Z}_{xy})|^2 = \sum_i |r_i|^2 \rightarrow \min \quad (5)$$

yielding

$$\widehat{\mathbf{Z}} = (\mathbf{H}^*\mathbf{H})^{-1}(\mathbf{H}^*\mathbf{E}) \quad (6)$$

where the superscript asterisk denotes the conjugate transpose of the complex matrix.

Robust estimation

The simple LS estimator implicitly assumes a Gaussian distribution for the errors ε_i in (3). This assumption often fails for MT data due to the non-stationarity of both signal and noise. Both can be highly variable in time, resulting in a marginal error distribution in the frequency domain that is heavy tailed and contaminated by outliers. As a result, the simple LS estimate all too frequently leads to very poor TF estimates with large error bars (Figure 3a). A number of MT processing methods have been proposed to overcome these difficulties, generally using some sort of automated screening or weighting of the data. Early efforts in this direction used ad hoc schemes, for example weighting data segments based on broad-band coherence between input and output channels. A more rigorously justifiable approach is based on the regression M-estimate (RME; Huber, 1981), a variant on LS that is robust to violations of distributional assumptions and resistant to outliers (Egbert and Booker, 1986; Chave et al., 1987). For the RME, the quadratic loss functional of (5) is replaced by

$$\sum_i \rho(|E_{xi} - (H_{xi}\hat{Z}_{xx} + H_{yi}\hat{Z}_{xy})|/\hat{\sigma}) = \sum_i \rho(|r_i|/\hat{\sigma}) \rightarrow \min \quad (7)$$

where $\hat{\sigma}$ is some estimate of the scale of typical residuals. LS is a special case of (7), with $\rho(r) = r^2$. By choosing the functional form for $\rho(r)$ so that large residuals are penalized less heavily than with the quadratic used for LS, the influence of outliers on the estimate can be substantially reduced. For robust estimation of MT TFs, the so-called *Huber* loss function (Huber, 1981)

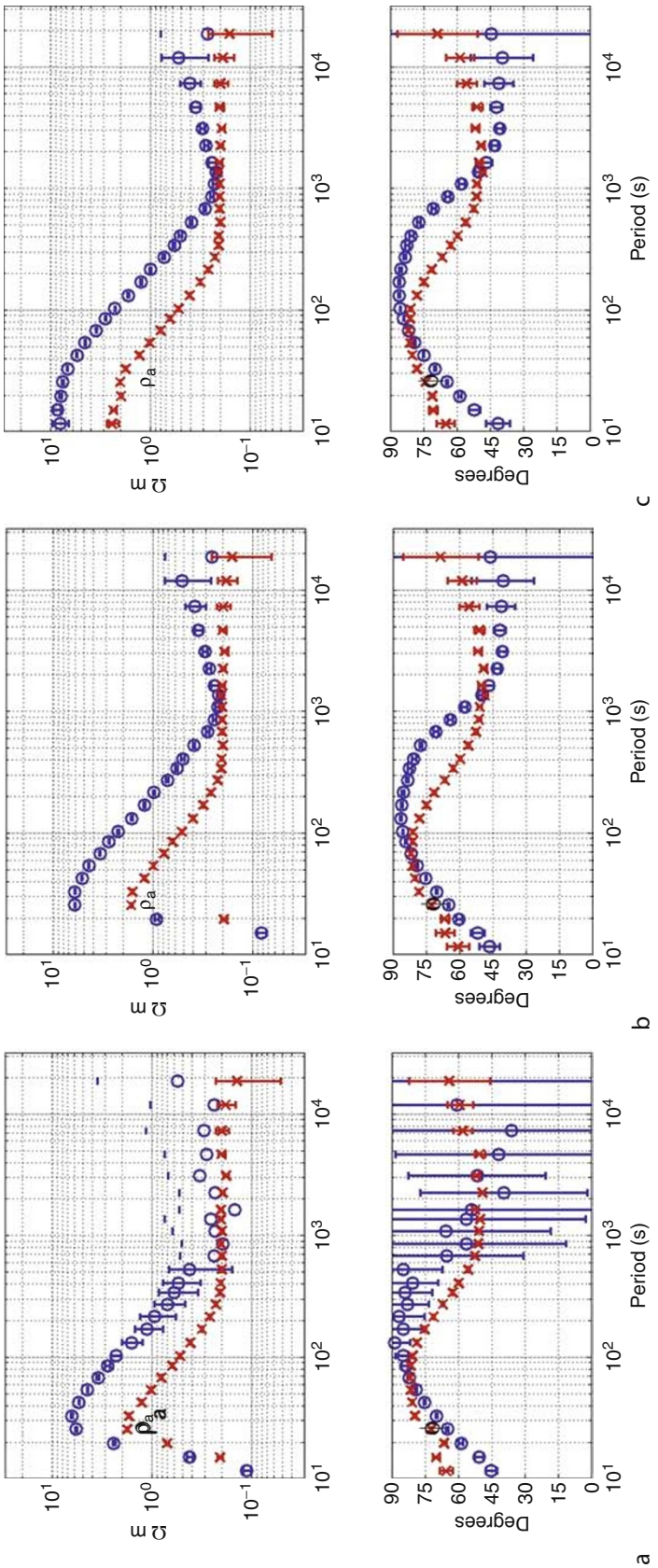
$$\rho(r) = \begin{cases} r^2/2 & |r| < r_0 \\ r_0|r| - r_0^2/2 & |r| >= r_0 \end{cases} \quad (8)$$

is commonly used with $r_0 = 1.5$.

The more general loss function of (7) can be minimized with an iterative scheme, which can be viewed as weighted LS, with weights determined from the data. Define the *influence function* $\psi(r) = \rho'(r)$ (the derivative of the loss function), and set $w(r) = \psi(r)/r$. Then, it is easily shown that the minimizer of (7) satisfies

$$\widehat{\mathbf{Z}} = (\mathbf{H}^*\mathbf{W}\mathbf{H})^{-1}(\mathbf{H}^*\mathbf{W}\mathbf{E}), \quad (9)$$

where $\mathbf{W} = \mathbf{diag}[w_1 \cdots w_I] = \mathbf{diag}[w(|r_1|) \cdots w(|r_I|)]$ is a diagonal matrix of weights. The RME thus corresponds approximately to the weighted LS problem



Magnetotelluric Data Processing, Figure 3 Comparison of TFs estimated from the same time series with three procedures described in the text (**a**) standard unweighted LS; (**b**) the regression M-estimate applied to data from a single site; and (**c**) the robust remote reference estimate. Here, apparent resistivities $\rho_a = (\omega\mu_0)^{-1}|Z|^2$ and phases $\phi = \tan^{-1}(Z)$ are plotted, computed from the dominant off-diagonal impedances Z_{xy} (*circles*) and Z_{yx} (*crosses*). The LS estimates are too noisy to be useful over much of the frequency range, while both of the robust estimates vary smoothly with frequency and have small error bars. However, the single site apparent resistivities of (**a**) and (**b**) are severely biased to low values for periods shorter than about 20 s. This bias is eliminated by the remote reference processing.

$\sum_i w(|r_i|)|r_i|^2 \rightarrow \min$. The weights depend on the normalized residuals $r_i = (E_{xi} - [Z_{xx}H_{xi} - Z_{xy}H_{yi}])/\hat{\sigma}$, and hence on the TF estimate, so an iterative procedure is required. Given an estimate of the TF, and of the error scale $\hat{\sigma}$, normalized residuals can be used to calculate weights, and the weighted LS problem can be solved for a new TF estimate. This procedure can be started from a standard LS estimate of the TF (and some robust estimate of error scale, for example, based on the median absolute deviation of residuals) and then repeated until convergence.

For convex loss functions (e.g., the Huber function of (8)), convergence of this procedure to the unique minimizer of (7) is guaranteed (Huber, 1981). For the Huber loss function in (8), the weights are

$$w(r) = \begin{cases} 1 & |r| > r_0 \\ r/r_0 & |r| \geq r_0 \end{cases}, \tag{10}$$

i.e., data corresponding to large normalized residuals get smaller weights. To allow for a sharp cutoff, with data exceeding a hard threshold discarded, the loss function $\rho(r)$ must be non-convex, and convergence of the iterative minimization algorithm cannot be guaranteed. The standard practice is to iterate with a convex loss function such as (8) to convergence, followed by a final step or two with a hard cutoff to completely discard extreme outliers. This completely automated procedure frequently results in significant improvements in TF estimates, as indicated by improvements in smoothness and physical realizability of apparent resistivity and phase curves (Figure 3b), and reproducibility of results.

More ad hoc schemes are also frequently used for down-weighting noisy or inconsistent data. These can be viewed as special cases of the weighted LS estimate of (9) but with weights now determined by some criteria other than residual magnitude. For example, in coherence weighting, broad-band coherence of input and output channels is used to down-weight time segments with low signal-to-noise ratios. Such methods can of course be combined with the RME.

Leverage

The RME can still be excessively influenced by a small number of data sections with large signal (or more typically noise) amplitude. In the terminology of linear statistical models, these large amplitude data are *leverage points*. For the LS estimate, the predicted data are

$$\hat{\mathbf{E}} = \left[\mathbf{H}(\mathbf{H}^*\mathbf{H})^{-1}\mathbf{H}^*\right]\mathbf{E} \tag{11}$$

with the matrix expression in brackets (which maps observed data to predicted) referred to as the "hat matrix." It is readily shown that the sum of the diagonal elements of the hat matrix satisfy $\sum_{i=1}^{I} h_{ii} = 2$, and that the individual diagonal elements h_{ii} can be interpreted as the fraction of total magnetic field signal power in the ith Fourier coefficient. Large values of h_{ii} indicate data points with inordinate influence on the TF estimates. In extreme cases, leverage points are used heavily to predict themselves, making the iterative RME ineffective. To deal with leverage points in routine MT processing, it is thus useful to include an additional weighting term (a function of the magnitude of h_{ii}) to reduce the influence of any observations of unusually large amplitude (e.g., Chave and Thomson, 1989).

Remote reference

The linear statistical model (4) is strictly appropriate to the case where noise is restricted to the output, or predicted electric field channels. Violation of this assumption results in the downward bias of estimated impedance amplitudes. These biases are proportional to the ratio of noise power to signal power, and can be quite severe in the so-called MT "dead bands" centered at frequencies 1–5 kHz and 0.1–1 Hz (examples of such biases are seen in Figure 3a, b, at periods shorter than 20 s). To avoid these bias errors, horizontal magnetic fields recorded simultaneously at a remote reference site are correlated with the EM fields at the local site (Gamble et al., 1979). Letting R_{xi} and R_{yi} be the Fourier coefficients for the two remote site components for the ith data segment, and $\mathbf{R}$ the corresponding $I \times 2$ matrix, then the remote reference analog of the LS estimate is

$$\hat{\mathbf{Z}} = [\mathbf{R}^*\mathbf{H}]^{-1}[\mathbf{R}^*\mathbf{E}]. \tag{12}$$

It is can be shown that the remote reference estimate is unbiased if noise at the remote site is not coherent with noise in magnetic and electric fields at the local site. Thus, the remote reference estimate can eliminate or significantly reduce the effects of coherent cultural EM noise, provided the reference site is sufficiently distant. Very large scale sources of cultural noise, in particular DC electric railways, still present a significant challenge to collecting high-quality MT data in many parts of the world. These sources can be coherent over large enough scales to contaminate both local and remote signals, but still have significant power at wavelengths which are short enough to violate the MT uniform source assumption.

To generalize the RME to remote reference, one can iterate the weighted analog of (12), i.e., $\hat{\mathbf{Z}} = [\mathbf{R}^*\mathbf{W}\mathbf{H}]^{-1}[\mathbf{R}^*\mathbf{W}\mathbf{E}]$, with the weights on the diagonal of $\mathbf{W}$ determined from the residual magnitudes, exactly as for the single site robust estimator. It is also useful to add some additional weighting to allow for outliers at the remote site, and as with the single site estimates, for leverage. When arrays of simultaneously operating EM instruments are available, more complex procedures, which use data from multiple sites to define the reference fields are possible (Egbert, 2002).

Error estimates

With a statistical approach to TF estimation, one also obtains error variances that define the precision of the TF estimates. These error bars are required to assess the

adequacy of fit of models derived by inversion, and thus play a critical role in the interpretation process. The covariance of the linear LS TF estimate of (6) is readily derived from standard theory, based on linear propagation of errors:

$$\begin{aligned}\mathbf{Cov}\left[\hat{\mathbf{Z}}\right] &= \hat{\sigma}^2[\mathbf{H}^*\mathbf{H}]^{-1} \\ \hat{\sigma}^2 &= (I-2)^{-1}\sum\nolimits_i |r_i|^2.\end{aligned} \tag{13}$$

where r_i are again the residuals. An analogous expression for the error covariance for the remote reference estimate is

$$\begin{aligned}\mathbf{Cov}\left[\hat{\mathbf{Z}}\right] &= \hat{\sigma}^2[\mathbf{R}^*\mathbf{H}]^{-1}[\mathbf{R}^*\mathbf{R}][\mathbf{H}^*\mathbf{R}]^{-1} \\ \hat{\sigma}^2 &= (I-2)^{-1}\sum\nolimits_i |r_i|^2.\end{aligned} \tag{14}$$

For both (13) and (14), the diagonal components of the 2×2 covariance matrices are the estimation error variances for the two impedance elements.

For the RME, variances of the estimates are complicated by nonlinearity, but asymptotic expressions (valid in the limit of large sample sizes) can be obtained from standard theory (Huber, 1981):

$$\mathbf{Cov}[\hat{\mathbf{Z}}] = \frac{(I-2)^{-1}\sum_i w_i|r_i|^2}{\left[I^{-1}\sum_i \psi'(|r_i|/\hat{\sigma})\right]^2}[\mathbf{H}^*\mathbf{H}]^{-1} \tag{15}$$

where r_i, w_i, $i = 1, \ldots, I$ and $\hat{\sigma}$ are the residuals, and weights and error scale from the final iteration, and the prime denotes the derivative of the influence function. An analogous asymptotic covariance for the robust remote reference estimates can be given with $[\mathbf{H}^*\mathbf{H}]^{-1}$ in (15) replaced by $[\mathbf{R}^*\mathbf{H}]^{-1}[\mathbf{H}^*\mathbf{H}][\mathbf{H}^*\mathbf{R}]^{-1}$.

As an alternative to asymptotic error estimates such as (15), the nonparametric jackknife method (Efron, 1982) has also been frequently used for computing TF error estimates (Thomson and Chave, 1991). The jackknife approach can also be applied to compute error bars for complicated nonlinear functions of TFs, which arise, for example, in some forms of distortion analysis. The jackknife approach significantly increases computational effort required for TF estimation, unless some approximations are used (Eisel and Egbert, 2001).

Strike and distortion analysis

As of this writing, MT data are most often acquired in profiles across a geologically appropriate strike, and the data are interpreted with two-dimensional (2D) Earth conductivity models that vary only with depth, and across strike, a reasonable approximation for many geologic structures. If the 2D assumption holds exactly, and the coordinate axes (used to give the components of the horizontal electric and magnetic fields) are chosen so that one of x or y aligns with the geoelectric strike (i.e., the direction along which there is no variation in conductivity), the diagonal components of the impedance tensor vanish.

$$\mathbf{Z}(\omega) = \begin{bmatrix} 0 & Z_{xy}(\omega) \\ Z_{yx}(\omega) & 0 \end{bmatrix}. \tag{16}$$

This special form only holds when the impedance is expressed in the proper coordinate system. Thus, once impedances are estimated, a key step in 2D MT interpretation is to determine an appropriate geoelectric strike where (16) holds approximately. The two off-diagonal components in this coordinate system (the transverse electric or TE mode, in which electric currents flow along the geologic strike, and transverse magnetic or TM mode, in which electric currents flow across the geologic strike; *Magnetotelluric Interpretation*), are then fit using 2D modeling and inversion programs.

Even if the 2D approximation is reasonably justified for large-scale deep structure, more complex (3D) small-scale near-surface complications in conductivity will occur. These near-surface features result in distortion of the impedance tensor, so that even when the correct regional strike is used to define the coordinate system, the impedance tensor does not have the simple form of (16). Most modern approaches to 2D MT interpretation allow for these distortion effects in strike determination, and subsequent inversion. Early approaches to strike determination, such as choosing the strike angle to minimize the magnitude of the diagonal components, without accounting for distortion effects are seldom used anymore.

In the limit where the penetration depth of the EM fields greatly exceeds the depth to the bottom of the near-surface distorting structures, the electric fields (only) are distorted in a static (frequency-independent) manner. Allowing for rotation of the measurement coordinate system by an angle θ, the distorted impedance tensor for a 2D Earth can be decomposed as

$$\begin{aligned}\mathbf{Z}(\omega)z &= \mathbf{U}\mathbf{D}\tilde{\mathbf{Z}}(\omega)\mathbf{U}' \\ &= \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}\begin{bmatrix} D_{xx} & D_{xy} \\ D_{yx} & D_{yy} \end{bmatrix} \\ &\times \begin{bmatrix} 0 & \tilde{Z}_{xy}(\omega) \\ \tilde{Z}_{yx}(\omega) & 0 \end{bmatrix}\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix},\end{aligned} \tag{17}$$

where $\mathbf{D}$ is the distortion matrix, which is real, and $\tilde{\mathbf{Z}}(\omega)$ is the undistorted 2D impedance, which carries all of the frequency dependence. It is readily verified that not all of the parameters in (17) are uniquely identifiable – e.g., the distortion parameter D_{xx} cannot be determined independent of the absolute level of the impedance $\tilde{Z}_{xy}$; only the product $D_{xx}Z_{xy}$ is uniquely determined. However, (17) provides a suitable model to determine an optimal geoelectric strike (defined by the rotation angle θ), and this has ultimately become the most common use for strike and distortion analysis in most 2D MT interpretations. Note that in principal the same $\mathbf{D}$ and θ should apply to all frequencies at one site, while the strike angle should be the same over all sites.

Nonlinear least-squares estimation schemes can be used to fit the parameters in (17) (e.g., Smith, 1995; McNeice and Jones, 2001). These are often applied site by site, for various subsets of frequencies, but impedances from multiple sites can also be fit simultaneously. Finding a consistent strike provides justification, and an optimal coordinate system, for 2D interpretation. Sometimes, the analysis is also used to eliminate sites, or frequency ranges, which are inconsistent with the generally preferred strike. Once the strike is determined and the impedance tensors are rotated, the off-diagonal components can be used as input for 2D inversion. Because the overall level of the impedance curves are still potentially effected by near-surface distortion, but the phases are not, it is common to fit phases more tightly, and relax fit to amplitudes (i.e., apparent resistivities). It is also possible to directly allow for distortion by introducing additional parameters at the inversion stage.

Summary

Data processing represents the first step in MT data interpretation, resulting in a reduction of the raw time-series data (approximately 10^6 observations) to frequency domain transfer functions such as impedances (a few hundred numbers). These transfer functions contain information about the response of the Earth to simple plane-wave sources, and provide input data for subsequent inversion and interpretation steps. Robust remote reference processing substantially improves reliability of transfer functions estimates. Except in areas of strong cultural noise (which unfortunately are increasingly common), modern processing produces smooth unbiased response curves. Distortion and strike analysis represents a further step in MT processing, essential for 2D interpretation of MT profile data. Such analysis is used to test for consistency with the assumed 2D geoelectric geometry, to estimate the strike, and to reduce the full impedance to the TE and TM mode responses appropriate to 2D interpretation.

Bibliography

Bendat, J. S., and Piersol, A. G., 1971. *Random data: Analysis and Measurement Procedures*. New York: Wiley.

Chave, A. D., and Thomson, D. J., 1989. Some comments on magnetotelluric response function estimation. *Journal of Geophysical Research*, **94**, B10, doi:10.1029/JB094iB10p14215.

Chave, A. D., Thomson, D. J., and Ander, M., 1987. On the robust estimation of power spectra, coherence and transfer functions. *Journal of Geophysical Research*, **92**, 633–648.

Efron, B., 1982. *The Jackknife, the Bootstrap and Other Resampling Methods*. Philadelphia: Society for Industrial and Applied Mathematics.

Egbert, G. D., 2002. Processing and interpretation of electromagnetic induction array data: a review. *Surveys in Geophysics*, **23**, 207–249.

Egbert, G. D., and Booker, J. R., 1986. Robust estimation of geomagnetic transfer functions. *Geophysical Journal of the Royal Astronomical Society*, **87**, 173–194.

Egbert, G. D., and Booker, J. R., 1989. Multivariate analysis of geomagnetic array data I: the response space. *Journal of Geophysical Research*, **94**, 14227–14248.

Eisel, M., and Egbert, G. D., 2001. On the stability of magnetotelluric transfer function estimates and the reliability of their variances. *Geophysical Journal International*, **144**, 65–82.

Gamble, T., Goubau, W., and Clarke, J., 1979. Magnetotellurics with a remote reference. *Geophysics*, **44**, 53–68.

Huber, P., 1981. *Robust Statistics*. New York: Wiley.

McNeice, G. W., and Jones, A. G., 2001. Multisite, multi-frequency tensor decomposition of magnetotelluric data. *Geophysics*, **66**, 158–173.

Smith, J. T., 1995. Understanding telluric distortion matrices. *Geophysical Journal International*, **122**, 219–226.

Thomson, D., and Chave, A., 1991. Jackknifed error estimates for spectra, coherence, and transfer functions. In Haykin, S. (ed.), *Advances in Spectrum Analysis and Array Processing*. Englewood Cliffs: Prentice Hall.

Cross-references

Electrical Properties of Rocks
Electrical Resistivity Surveys and Data Interpretation
Geoelectromagnetism
Instrumentation, EM
Magnetotelluric Interpretation

MAGNETOTELLURIC INTERPRETATION

John F. Hermance
Department of Geological Sciences, Brown University, Providence, RI, USA

What is magnetotellurics?

The magnetotelluric (MT) method is a geophysical technique in which the time-varying electric and magnetic fields from distant natural or artificial sources are measured at the Earth's surface, and analyzed to determine the electrical properties of the subsurface.

Workers have known from the 1800s that naturally occurring transient voltages of significant magnitude observed on grounded telegraph wires were associated with rapid fluctuations of the geomagnetic field following *solar flares* and *auroral displays*. These natural voltage fluctuations – so-called telluric fields – have since been explained as the product of eddy currents induced in the finite conducting earth by transient magnetic field disturbances in the ionosphere and magnetosphere, even the global-wide propagation of radiated signals from lightning strokes in the distant atmosphere (see *Magnetic Storms and Electromagnetic Pulsations*).

Early work by Kato and Kikuchi, and Rikitake in Japan, and Tikhonov in the then, Soviet Union, laid the foundation for using these natural electric and magnetic field variations to study the electrical structure of the solid earth (see the collection of classic reprints by Vozoff, 1985; see also *Geoelectromagnetism*). However, it was the seminal paper by Louis Cagniard (1953) that first described the comprehensive procedure for using the ratio of the

horizontal electric – that is, the *telluric* – field to the horizontal *magnetic* field, for which he coined the name the *magnetotelluric method*.

Fundamental concepts: surface impedance and apparent resistivity

The magnetotelluric method, as generally employed by geophysicists, invokes several fundamental concepts from the theory of the propagation of electromagnetic waves. The first concept is the physical attribute known as the *electromagnetic wave impedance* of the medium (see *Magnetotelluric Data Processing*), provided by computing the ratio of the electric field intensity to the orthogonal magnetic field intensity. For example, Figure 1 shows an electromagnetic plane wave vertically incident on the earth's surface; the electric field is polarized in the x-direction, and, to be consistent with the Poynting vector $\mathbf{S} = \mathbf{E} \times \mathbf{H}$, the magnetic field is polarized in the y-direction. Alternatively, for the case where the electric field might be polarized in the y-direction, the magnetic field will be polarized in the $-x$-direction. The ideal natural field for magnetotellurics is one which is randomly polarized in *all* horizontal directions, so that usually both sets of electric and magnetic data, (E_x, H_y) and (E_y, H_x), will be employed, leading to two estimates of the *surface impedance*:

$$Z_{xy}^{(\mathrm{surf})} = \frac{E_x}{H_y} \text{ and } Z_{yx}^{(\mathrm{surf})} = -\frac{E_y}{H_x} \tag{1}$$

The impedance terms in Equation 1 can be looked upon as the coefficients of linear coupling relations between the electric and respective magnetic field components having the forms:

$$E_x = Z_{xy}^{(\mathrm{surf})} H_y \text{ and } E_y = -Z_{yx}^{(\mathrm{surf})} H_x \tag{2}$$

The subscripts on the impedance coefficients denote the ordering of the respective electric and corresponding orthogonal magnetic field components.

Under appropriate conditions, the *resistivity* of the uniform half-space in Figure 1 may be determined from the measured surface impedance, since the *characteristic impedance* that a wave encounters traveling through a homogeneous medium is given by $Z_{\mathrm{char}} = E_x/H_y = -E_y/H_x = \sqrt{i\omega\mu\rho_{\mathrm{Tot}}}$ where $i = \sqrt{-1}$, ω is the radian frequency ($\omega = 2\pi f$, and $f = 1/T$ is the frequency in Hz, where T is the period in seconds), μ is the free-space magnetic permeability ($\mu = 4\pi \times 10^{-7}$ henries/m), and ρ_{Tot} is the total resistivity that accounts for both ohmic and dielectric effects in the medium.

The latter expressions for the ratios of the electric field to its orthogonal magnetic field counterpart, can be rearranged to estimate the resistivity of the medium according to $\rho_{\mathrm{obs}} = (1/\omega\mu)\left|E_x/H_y\right|^2 = (1/\omega\mu)\left|E_y/H_x\right|^2$.

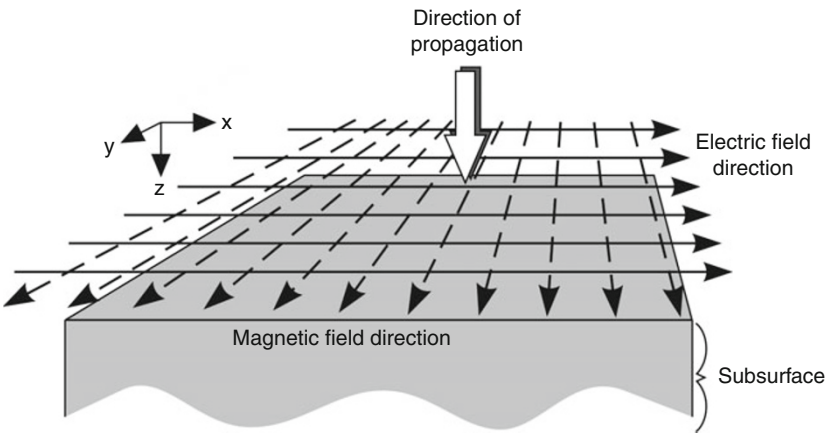
In general, however, the earth is *not* uniform, and as proposed by Cagniard (1953) in his classic paper, it is best to refer to the observed resistivity as an *apparent resistivity*, which might be determined using either combination of the respective field components, (E_x, H_y) or (E_y, H_x). The corresponding apparent resistivities are commonly referred to as *Cagniard pairs*, and are given by

$$\rho_{xy}^{(\mathrm{app})} = \frac{1}{\omega\mu}\left|\frac{E_x}{H_y}\right|^2 \text{ and } \rho_{yx}^{(\mathrm{app})} = \frac{1}{\omega\mu}\left|\frac{E_y}{H_x}\right|^2 \tag{3}$$

For a uniform earth as well as for a plane-layered earth, as discussed below, one should expect that $\rho_{xy}^{(\mathrm{app})} = \rho_{yx}^{(\mathrm{app})}$. Exceptions to the latter denote the presence of lateral heterogeneities, also discussed below.

Principles of depth sounding

The *skin depth* or *depth of penetration* of an electromagnetic wave impinging on a uniform conductor (the earth) is given by $\delta \cong 500\sqrt{\rho \cdot T}$ meters, where ρ is the electrical resistivity of the medium in Ohm.m, and T is the period (inverse of frequency) of the signal in seconds. The consequence of the *skin-depth effect* is that low frequency (long period) waves penetrate to greater depth in a given



Magnetotelluric Interpretation, Figure 1 Schematic of a uniform plane wave vertically incident on the earth's surface.

conductor than high frequency (short period) waves (see *Geoelectromagnetism*).

The intrinsic resistivity of typical earth materials varies over many orders of magnitude, depending on mineralogy, water content, and temperature (see *Electrical Properties of Rocks*). Figure 2 summarizes the range of characteristic depths of penetration over the typical range of earth resistivities and frequencies employed in magnetotellurics. As an aside, note in the upper right corner of the figure, the roll-over of the linear isolines for high frequencies and high resistivities due to dielectric effects coming into play.

The *apparent resistivity* as a function of increasing period can be determined using spectral analysis of the measured electric and magnetic time series (see Vozoff, 1985; see *Magnetotelluric Data Processing*). Then, from measurements of the apparent resistivity as a function of period, one can use the variable depth of penetration to electrically "probe" the earth to greater depth by progressively "tuning" the observed signals to progressively longer period.

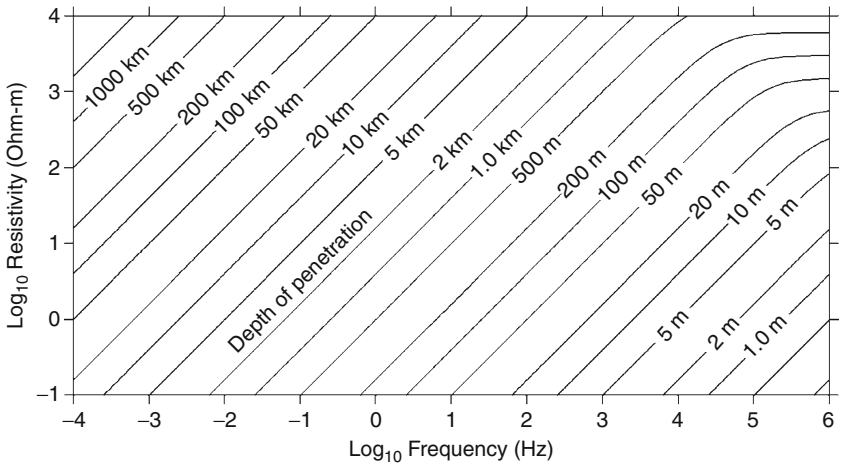
For conventional magnetotelluric applications, at the highest frequencies (≅1 MHz) and the lowest resistivities (close to that of seawater, 1/3 Ω-m, or lower), the signal (often from commercial or government radio transmitters, or, in the audio band, from distant lightning strikes) would penetrate to a few meters or few tens of meters depth. This range is useful for shallow geotechnical and groundwater investigations, and perhaps shallow geothermal work.

At the other end of the spectrum, at frequencies of a Hertz or lower ($T \geq 1$ s), workers usually rely on natural sources in the magnetosphere and ionosphere (such as micropulsations, magnetic substorms, and diurnal variations), (see *Magnetic Storms and Electromagnetic Pulsations*) to probe to depths of tens to hundreds of kilometers.

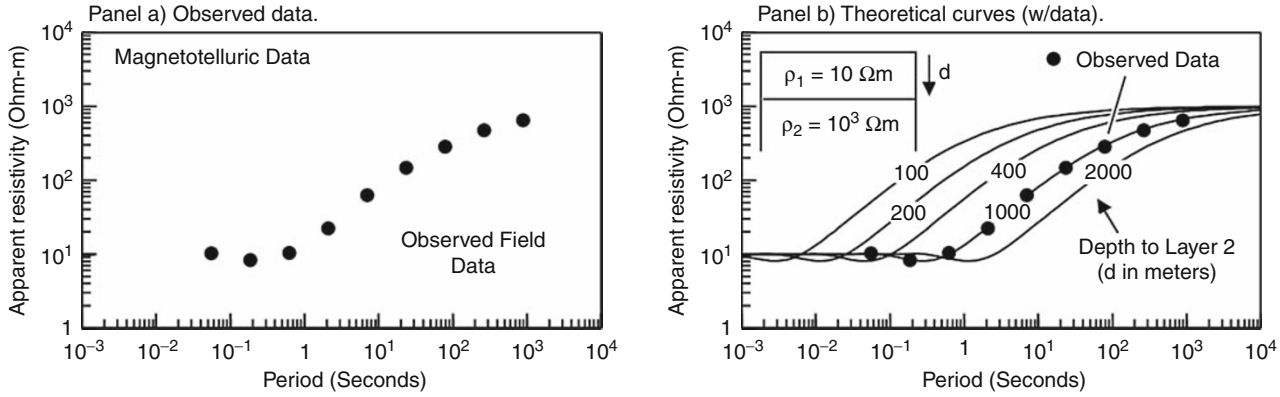
An example of the apparent resistivity that might be measured on the surface of a broad sedimentary basin as a function of period, as envisioned by Louis Cagniard in his 1953 paper, is shown in Figure 3. The exploration geophysicist would have in mind that the relatively conductive sediments are underlain at some unknown depth by more resistive crystalline or metamorphosed basement (see *Electrical Properties of Rocks*). At an early stage, in the exploration of such a basin, the interpreter would like to determine three types of information from the observed MT data: *first*, the physical properties of the sedimentary basin (ρ_1: the resistivity of layer 1); *second*, the depth d to the base of the sediments; *third*, the physical properties of the underlying basement (ρ_2: the resistivity of layer 2). From the various models seen in Figure 3b, a theoretical model with a surface resistivity of 10 Ω-m, underlain by a more resistive 1,000 Ω-m layer at a depth of 1,000 m, "fits" the observed data quite well. The mathematical framework through which one determines a model that best fits the set of observed data is known as *inverse theory* (see *Magnetotelluric Data Processing*; *Inverse Theory, Global Optimization*; *Inverse Theory, Linear*; *Inverse Theory, Monte Carlo Method*; *Inverse Theory, Singular Value Decomposition*).

Effects of lateral heterogeneities

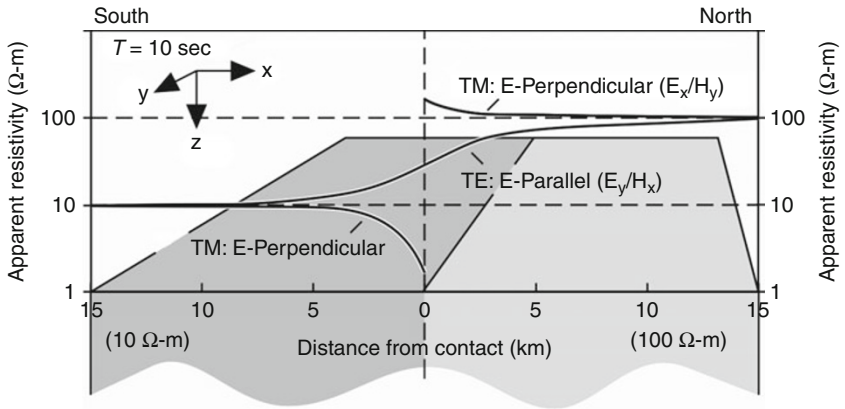
Early workers discovered the effects of lateral heterogeneities on MT data when they found that the Cagniard pairs were unequal: $\rho_{xy}^{(\mathrm{app})} \neq \rho_{yx}^{(\mathrm{app})}$ in Equation 3. That is to say, the observed apparent resistivity often depended on the direction in which the respective E and H fields were measured, as well as on the location of the measurements relative to local geologic features. Consider, for example, the case of the simple 2D vertical contact shown in Figure 4. The observed apparent resistivities can be resolved into two principal modes: the incident electric source field may be polarized *perpendicular* to the strike of the contact, or it may be *parallel* to the strike of the contact in the y-direction.



Magnetotelluric Interpretation, Figure 2 Depth of penetration as a function of frequency and resistivity of the host medium for geophysical applications.



Magnetotelluric Interpretation, Figure 3 (Panel a) An example of field data from magnetotelluric measurements. (Panel b) Field data overlaid on a set of theoretical curves representing various combinations of model parameters, in this case the resistivity of each of two layers and the depth of layer 2.



Magnetotelluric Interpretation, Figure 4 Apparent resistivities along profiles transecting a vertical contact for two polarizations – TM and TE – of the source field relative to strike.

Computing the corresponding apparent resistivities from the electromagnetic impedances for both cases, we obtain for the *TM* (Transverse Magnetic), or E-perpendicular mode:

$$Z_{xy}^{(\mathrm{surf})} = \frac{E_x}{H_y}, \text{ leading to } \rho_{xy}^{\mathrm{app}} = \frac{1}{\omega\mu}\left|\frac{E_x}{H_y}\right|^2 \tag{4}$$

and for the *TE* (Transverse Electric), or *E*-parallel mode

$$Z_{yx}^{(\mathrm{surf})} = -\frac{E_y}{H_x}, \text{ leading to } \rho_{yx}^{\mathrm{app}} = \frac{1}{\omega\mu}\left|\frac{E_y}{H_x}\right|^2 \tag{5}$$

Impedance matrix. Of course in actual field surveys one usually does not know ahead of time the trend or principal directions of local subsurface geologic features, so the electrode lines and magnetic sensors are usually deployed in standard geographic or geomagnetic coordinates. This leads to a "*mixing*" or "*cross-coupling*" of the field components. With reference to the linear coupling forms in Equation 2, it is convenient to express these relations in the compact matrix form

$$\begin{bmatrix} E_x^{(\mathrm{p})} \\ E_y^{(\mathrm{p})} \end{bmatrix} = \begin{bmatrix} 0 & Z_{xy}^{(\mathrm{p})} \\ Z_{yx}^{(\mathrm{p})} & 0 \end{bmatrix} \begin{bmatrix} H_x^{(\mathrm{p})} \\ H_y^{(\mathrm{p})} \end{bmatrix} \tag{6}$$

where the superscript "p" denotes measurements in the principal coordinates of the feature (such as in the *x*- and the *y*-directions in Figure 4). Rotating relation (6) through an arbitrary angle θ clockwise from the principal coordinates in Figure 4 leads to a relation between the measured electric and magnetic field components of the form

$$\begin{bmatrix} E_x \\ E_y \end{bmatrix} = \begin{bmatrix} Z_{xx} & Z_{xy} \\ Z_{yx} & Z_{yy} \end{bmatrix} \begin{bmatrix} H_x \\ H_y \end{bmatrix} \tag{7}$$

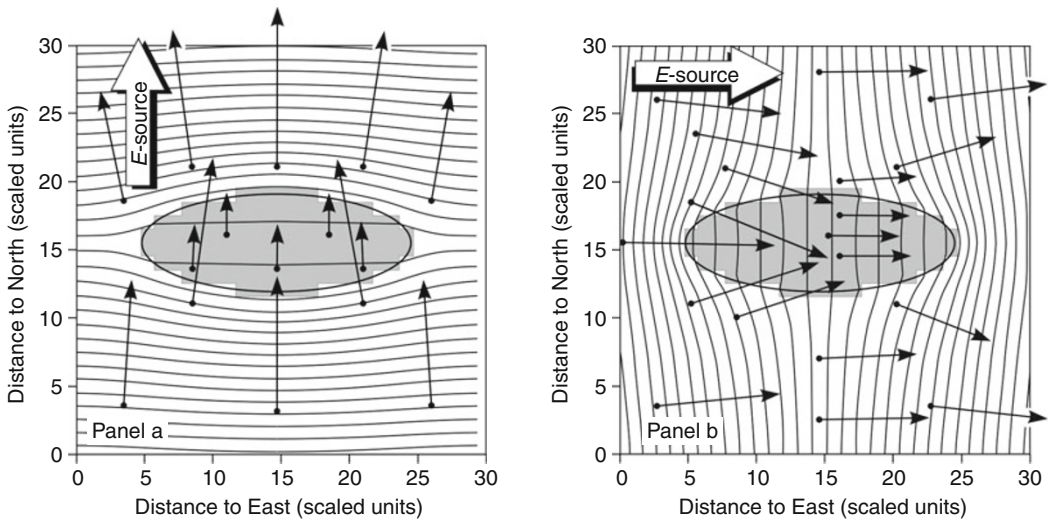
where the *impedance matrix* is often referred to as a *tensor impedance*.

Three-dimensional (3D) heterogeneities. Recognizing that the earth is, in general, 3D, electromagnetic workers have developed various approximate methods to dealing with it since the 1930s (see Keller and Frischknecht, 1966). Moreover, the theory of electromagnetic induction in 3D laterally heterogeneous material has evolved significantly over the decades since 1980, largely through implementing physically representative numerical techniques (see the work cited for Stodt, Wannamaker, Mackie and Madden among others, in Simpson and Bahr, 2005; see also Siripunvaraporn et al., 2005). Several of the principal effects of three dimensionality are illustrated in the example in Figure 5 which is a plan view map of telluric (i.e., electric) field distortions in the vicinity of a small idealized, elliptically shaped conductivity anomaly in the surface layer of an otherwise uniform or plane-layered earth. The embedded "elliptical" feature in the figure is a modest four times more conductive than the host material (although it is not uncommon for lateral contrasts to be an order of magnitude or more). The important point is that the electric field (and the associated magnetic field) can change dramatically in magnitude and direction over very small distances.

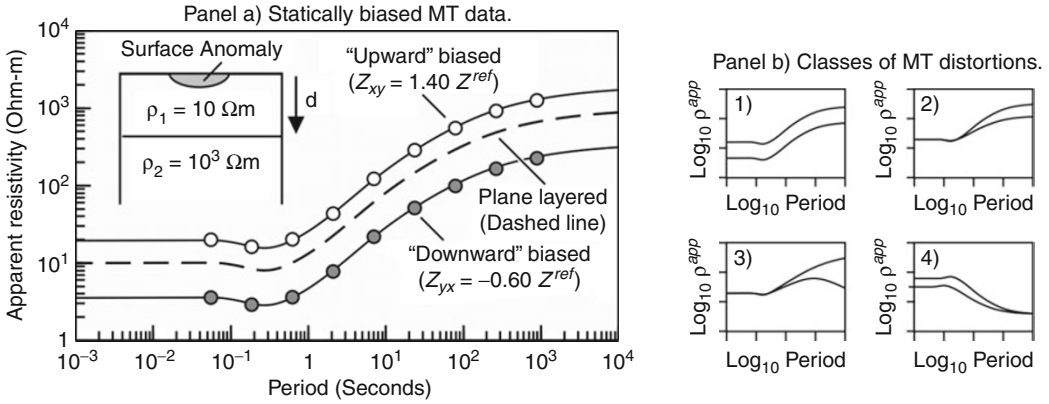
Effects of bias of MT data from lateral heterogeneities. Workers realized early in the development of magnetotellurics that linear coupling relations like Equation 7 are most often the rule in field data (see the collection of reprints assembled by Vozoff, 1985). The first step, therefore, in interpreting MT data often involves decomposing multiple independent sets of observations of both of the electric field components $\left(E_x^{(\mathrm{obs})}, E_y^{(\mathrm{obs})}\right)$ and the associated magnetic field components $\left(H_x^{(\mathrm{obs})}, H_y^{(\mathrm{obs})}\right)$ to extricate estimates of the principal values of apparent resistivity, ρ_{xy}^{app} and ρ_{yx}^{app}, in the principal directions. It is common practice to plot up both resistivity estimates as a function of period, as was done for the plane-layered case in Figure 3. However, the effects of even a small heterogeneity on the MT response of an otherwise plane-layered earth is illustrated in Figure 6, using as the background the plane-layered example in Figure 3 (shown as the dashed line in Figure 6).

The dashed curve in Figure 6a represents the original plane-layered response of Figure 3, along with the hypothetical effects that might be encountered in MT field data acquired at a site within, or in the vicinity of, a small, local embedded surficial anomaly. For one of the principal polarizations, the affected apparent resistivity relative to the original plane-layered curve is "*upward-biased*" by almost a factor of 2, and for the orthogonal polarization the apparent resistivity is "*downward-biased*" to 36% of its undistorted value. (There is no reason, of course, that for some cases *both* $\rho_{xy}^{\mathrm{app}}(T)$ and $\rho_{yx}^{\mathrm{app}}(T)$ might be biased to higher values, or both biased to lower values.)

Such bias effects are endemic to magnetotelluric surveys, and are particularly challenging to those investigations that are intended to be "*deep-sounding*" of the crust and upper mantle (and the mantle transition zones), since it is so easy for an undetected static offset to bias the long period asymptotic values to larger, deeper, and more resistive – or smaller, shallower, and more conductive – values. If, along with the "static" offsets described above, there are 3D induction effects as well, the field data become increasingly more complicated to untangle. Figure 6b shows several selected classes of MT responses often observed in field data due to quasi-static and inductive coupling effects for various combinations of shallow and deep (or distant) lateral heterogeneities. While one must



Magnetotelluric Interpretation, Figure 5 Plan view map of telluric (or electric) field distortions in the vicinity of an idealized embedded conductivity anomaly. (Panel a) Voltage contours and selected electric field directions for a north-directed electric source field. (Panel b) The same parameters as panel a, but for an east-directed source field.



Magnetotelluric Interpretation, Figure 6 Examples of static bias. (Panel a) Example to demonstrate the mechanism of "quasi-static shift" (referred to the dashed plane-layered case). (Panel b) Selected classes of "distortions" of MT data due to various combinations of shallow and deep lateral heterogeneities, with static and inductive coupling effects.

be cautious in drawing too much from a single set of sounding curves, these panels might illustrate the following generic phenomena. Panel b1 illustrates the type of static shift from a shallow surface feature discussed above. Panel b2, on the other hand, suggests a fairly uniform layer at shallow depth, but as the period and depth of penetration increases, the principal resistivities begin to separate indicating a deeper-seated inhomogeneity. One needs to be careful here, as the inhomogeneity may lay off to the side, and be a lateral effect, rather than a simple depth effect. Panel b3 suggests a similar, but more complicated phenomenon in that the two resistivity modes do not simply separate at longer period, but their respective shapes are frequency dependent, indicating that there are significant induction effects as well as static shifts coming into play. Panel b4 is an idealized version of situations commonly encountered in real data that might be theoretically explained by a true anisotropic resistivity layer at the surface, with a more conductive layer at depth. Often, however, for real world data, the merging of the principal resistivity curves with increasing period is simply a precursor to their crossing and then shifting apart again at still longer periods. In all of these cases, drawing 1D inferences from data that are clearly affected by 2D and 3D structure is questionable. For other examples, see Vozoff (1985) and Simpson and Bahr (2005).

Current strategies for MT interpretation

Much of the progress on MT interpretation over the last 3 decades is due to a progressive improvement in data acquisition, robust processing of multi-site data, and careful 2D and 3D numerical modeling. Simultaneously, recording multi-station MT arrays have become the standard practice for industry, academia, and government surveys. Recognizing the importance of high-density coverage using simultaneous observations of MT fields at multiple sites, several national programs are developing multi-station MT arrays. Typical instrumentation for such facilities is described by Smirnov et al. (2008), with some of the preliminary results of the Earthscope MT array in the United States reported by Patro and Egbert (2008) (See also *Instrumentation, EM*).

There is an increasing emphasis, when interpreting MT data, on either explicitly using 3D models, or at least on using 3D models as guides when using the more efficient, less computationally intensive, 2D inverse techniques. The earth, however, *is* three-dimensional, and most workers agree that any attempt to simplify its inherent complexity with 2D or even 1D plane-layered models, needs to be assessed at some level with effective and representative 3D models. Such models have progressively evolved from the initial contributions reported in the reprint collection of Vozoff (1985). A number of notable reports on 3D numerical models – modeling, inversion, and applications – are contained in the compendium of Oristaglio and Spies (1995), with additional developments by Mackie et al. (1993), Siripunvaraporn et al. (2005), and Tong, et al. (2009).

In spite of the promise of full 3D MT interpretations, most recognize that the inversion of data from a large regional array of broadband stations in terms of generalized 3D models tends to be computationally intensive (Newman and Alumbaugh, 2000; Zhdanov and Tolstaya, 2004). Thus, most surveys interpret array data in terms of 2D cross-sections. In many cases, the response of a 3D feature is locally quite 2D in nature. In other cases, it may be possible that the type of localized static shift discussed in an earlier section can be compensated for, and the observed coupling matrix can be decomposed into regionalized principal elements (Groom and Bailey, 1989). Berdichevsky et al. (1998) summarize several of the major concepts that many interpreters employ when analyzing MT data from 2D (and 3D) environments. First, while the *TM* (*E*-perpendicular) mode is often quite sensitive to local near-surface features, the *TE* (*E*-parallel) mode is often more indicative of deeper, more regional features. Second, while the *TM* mode is more susceptible to the effects of static shift, it is not uncommon that the

TE mode may be relatively undistorted. Both modes, however, are essential for a comprehensive interpretation of a study area, but must be used prudently. Ledo et al. (2002) discuss the limitations as well as the computational advantages – when appropriate – of applying 2D model inversions to actual 3D field situations.

For many investigations, the concomitant 2D and 3D effects might be considered unwanted "distortions" of deep-sounding data, and a bane to those studies intended to better understand the deep crust, or to delineate the depth to the lithosphere–asthenosphere boundary, the mantle transition zones, and so forth, on regional or subglobal scales. On the other hand, there is indeed a great deal of information in MT data that could be essential for characterizing the full 3D character of the earth, and the method, when properly executed, is proving to be a powerful means for delineating geologic features having application to a variety of fundamental geophysical and geological problems. However, it should be clear from the example in Figure 5 that in order to identify, characterize, and account for the effects of localized surficial features, one needs to deploy a sufficient density of sites to spatially characterize the surface field patterns. The situation is quite analogous to the "aliasing" problem that governs the strategy for properly sampling a time series; only in MT we are not dealing with simply a 1D sequence, but rather with 2D areal coverage. Berdichevsky (1999) offers a thoughtful critique of several of the underpinning paradigms and procedures of modern magnetotellurics.

Case studies

A number of early applications of the MT method are described in the texts of Patra and Mallick (1980), and Vozoff (1985). More recent case studies are described in the text by Simpson and Bahr (2005). The most exciting of these involve multidisciplinary geophysical field efforts. Particularly, noteworthy is the series of integrated MT, seismic refraction, and gravity investigations initiated by the US Geological Survey in the Pacific Cascades volcanic belt in the Northwest US (see the review by Stanley et al., 1990). Early phases of this work led to the first major international offshore/onshore MT and geomagnetic deep-sounding (GDS) project: the EMSLAB project that extended from the offshore Juan de Fuca Rise to the east side of the Cascades volcanic system (*c.f.*, Wannamaker et al., 1989). Work in this area continues. For example, in a reanalysis of data from the EMSLAB Project, Vanyan et al. (2002) articulate the experience of many that when interpreting a profile of widely spaced, multi-site MT data in such complicated terrains, one needs to be selective in which elements of the MT impedance matrix are most consistent when developing a 2D inversion. They found that under the complex geologic conditions of the region, the strict inversion of both the TE and TM modes simultaneously at each and every site yielded an intricate, but what is likely a geophysically implausible, pattern of conductive and resistive features resulting in a large misfit of the model response to the data. The analysis improved significantly, they felt, when they were more selective in which of the principal impedance components were employed at a site. They also found that using lateral variations of the transient magnetic field, in conjunction with the MT data, significantly complemented their analysis. This point of view probably captures the spirit of most current MT investigations, and certainly underscores the importance of multiple station arrays of closely spaced sites.

In other regions, Sarma et al. (2004) report on a comprehensive MT study of an active seismic zone in western India. Solon et al. (2005) describe a large array MT study of the deep crust beneath the Himalayas in central Tibet. Munoz et al. (2008) investigate mid-crustal conductivity anomalies using 3D models of SW Europe. Wannamaker et al. (2008) describe the application of MT to understanding lithospheric deformation and magmatic processes along the Great Basin-Colorado Plateau transition zone. Continuing the work in the US Pacific Northwest, Patro and Egbert (2008) report on their analysis of data from deployments in 2006 and 2007 of the still-evolving US Earthscope MT Transportable Array. Data were sequentially acquired in the period range from 10 to 10,000 s at 110 sites, covering the region at a nominal station spacing of 75 km. Han et al. (2009), using a 3D model to interpret MT data on the Korean peninsula, emphasize the importance of accounting for the effects of the nearby ocean and coastline. Eaton et al. (2009), in pulling together a variety of geophysical evidence, present a recent update on MT investigations of the lithosphere and asthenosphere beneath cratons, particularly for North America. Similar applications of MT to understanding significant solid earth problems are ongoing in China, Japan, Europe, Great Britain, South America, Antarctica, Africa, New Zealand, and Australia. Clearly, magnetotellurics is becoming an essential tool for a better understanding of earth systems on a variety of scales.

Summary

The magnetotelluric (MT) method is based on the physical theory of electromagnetic wave propagation in electrically heterogeneous materials. Using the "skin-depth" principle that, in a given conductor, long period signals penetrate to greater depth than short period signals, the method typically employs the rich spectrum of natural electromagnetic energy from the ionosphere, magnetosphere, and atmosphere (distant lightning storms) to "probe" the electrical structure of the earth from a few hundred meters depth to tens even hundreds of kilometers. For near-surface applications, active transmitters are used throughout the audio band (100 Hz to 20 kHz) to above a megahertz. The last decades have seen significant developments in the type of low-noise, large dynamic range instrumentation needed for such studies, along with robust real-time data processing, 3D numerical modeling, and geophysical inverse theory. The MT method involves recording and

analyzing the spectral character of the complex amplitudes and phases of the electric and magnetic field components over a wide range of periods from an array of closely spaced sites on the earth's surface. From the spatial and spectral character of such data, one can, in a sense, deconvolve the signatures of lateral and vertical variations in the bulk properties of the subsurface to develop insight on fundamental processes in the earth.

Bibliography

Adam, A., (Editor-in-Chief), 1976. *Geoelectric and Geothermal Studies (East Central Europe, Soviet Asia)*, KAPG Geophysical Monograph. Budapest: Akademiai Kiado.

Berdichevsky, M. N., 1999. Marginal notes on magnetotellurics. *Surveys in Geophysics*, **20**, 341–375.

Berdichevsky, M. N., Dmitriev, V. I., and Pozdnjakova, E. E., 1998. On two-dimensional interpretation of magnetotelluric soundings. *Geophysical Journal International*, **133**(3), 585–606.

Cagniard, L., 1953. Basic theory of the magnetotelluric method of geophysical prospecting. *Geophysics*, **18**, 605–635.

Eaton, D. W., Darbyshire, F., Evans, R. L., Gruetter, H., Jones, A. G., and Yuan, X., 2009. The elusive lithosphere-asthenosphere boundary (LAB) beneath cratons. *Lithos*, **109** (1–2), 1–22.

Groom, R. W., and Bailey, R. C., 1989. Decomposition of magnetotelluric impedance tensors in the presence of local 3-dimensional galvanic distortion. *Journal of Geophysical Research – Solid Earth and Planets*, **94**(B2), 1913–1925.

Han, Nuree, Nam, M. J., Kim, H. J., Lee, T. J., Song, Y., and Suh, J. H., 2009. Three-dimensional inversion of magnetotelluric data including sea effects obtained in Pohang, Korea. *Journal of Applied Geophysics*, **68**(4), 533–545.

Kaufmann, A. A., and Keller, G. V., 1981. *The Magnetotelluric Sounding Method*. New York: Elsevier.

Keller, G. V., and Frischknecht, F. C., 1966. *Electrical Methods of Geophysical Prospecting*. New York: Pergamon.

Ledo, J., Queralt, P., Marti, A., and Jones, A. G., 2002. Two-dimensional interpretation of three-dimensional magnetotelluric data: an example of limitations and resolution. *Geophysical Journal International*, **150**, 127–139.

Mackie, R. L., Madden, T. R., and Wannamaker, P. E., 1993. 3-Dimensional magnetotelluric modeling using finite difference equations and comparisons to integral-equations solutions. *Geophysics*, **58**(2), 215–226.

Munoz, G., Mateus, A., Pous, J., Heise, W., Santos, F. M., and Almeida, E., 2008. Unraveling middle-crust conductive layers in Paleozoic Orogens through 3D modeling of magnetotelluric data: the Ossa-Morena Zone case study (SW Iberian Variscides), *Journal of Geophysical Research – Solid Earth and Planets*, **113**(B6), Art. No. B06106.

Newman, G. A., and Alumbaugh, D. L., 2000. Three-dimensional magnetotelluric inversion using non-linear conjugate gradients. *Geophysical Journal International*, **140**(2), 410–424.

Oristaglio, M., and Spies, B. (eds.), 1995. Three-dimensional electromagnetics. In *Proceedings of an International Symposium in Honor of Jerry Hohmann*, Schlumberger-Doll Research, Cambridge, MA.

Patra, H. P., and Mallick, K., 1980. *Geosounding Principles, 2; Time-Varying Geoelectric Soundings*. New York: Elsevier.

Patro, P. K., and Egbert, G. D., 2008. Regional conductivity structure of Cascadia: preliminary results from 3D inversion of U.S. Array transportable array magnetotelluric data. *Geophysical Research Letters*, L20311, doi: 10.1029/2008GL035326.

Sarma, S. V. S., Prasanta, B., Patro, K., Harinarayana, T., Veeraswamy, K., Sastry, R. S., and Sarma, M. V. C., 2004. A magnetotelluric (MT) study across the Koyna seismic zone, western India: evidence for block structure. *Physics of the Earth and Planetary Interiors*, **142**(1–2), 23–36.

Simpson, F., and Bahr, K., 2005. *Practical Magnetotellurics*. Cambridge: Cambridge University Press.

Siripunvaraporn, W., Egbert, G., Lenbury, Y., and Uyeshima, M., 2005. Three-dimensional magnetotelluric inversion: data space method. *Physics of the Earth and Planetary Interiors*, **150**, 3–14.

Smirnov, M., Korja, T., Dynesius, L., Pedersen, L. B., and Laukkanen, E., 2008. Broadband magnetotelluric instruments for near-surface and lithospheric studies of electrical conductivity: a fennoscandian pool of magnetotelluric instruments. *Geophysica*, **44**, 31–44.

Solon, K. D., Jones, A. G., Nelson, K. D., Unsworth, M. J., Kidd, W. F., Wei, W., Tan, H., Jin, S., Deng, M., Booker, J. R., Li, S., and Bedrosian, P., 2005. Structure of the crust in the vicinity of the Banggong-Nujiang suture in central Tibet from INDEPTH magnetotelluric data. *Journal of Geophysical Research – Solid Earth and Planets*, **110**(B10), 1–20.

Stanley, W. D., Mooney, W. D., and Fuis, G. S., 1990. Deep crustal structure of the Cascade Range and Surrounding Regions from seismic refraction and magnetotelluric data. *Journal of Geophysical Research – Solid Earth and Planets*, **95**(B12), 19419–19438.

Tong, X. Z., Liu, J. X., Xie, W., Xu, L. H., Guo, R. W., and Cheng, Y. T., 2009. Three-dimensional forward modeling for magnetotelluric sounding by finite element method. *Journal of Central South University of Technology*, **16**(1), 136–142.

Vanyan, L. L., Berdichevsky, M. N., Pushkarev, P. Y., and Romanyuk, T. V., 2002. A geoelectric model of the Cascadia subduction zone. *Izvestiya Physics of the Solid Earth*, **38**(10), 816–845.

Vozoff, K. (ed.), 1985. *Magnetotelluric Methods, Geophysics Reprints Series No. 5*. Tulsa: Society of Exploration Geophysicists.

Wait, J. R., 1982. *Geo-Electromagnetism*. New York: Academic.

Wannamaker, P. E., Booker, J. R., Filloux, J. H., Jones, A. G., Jiracek, G. R., Chave, A. D., Tarits, P., Waff, H. S., Egbert, G. D., Young, C. T., Stodt, J. A., Martinez, M., Law, L. K., Yukutake, T., Segawa, J. S., White, A., and Green, A. W., 1989. Magnetotelluric observations across the Juan de Fuca Subduction System in the EMSLAB Project. *Journal of Geophysical Research – Solid Earth and Planets*, **94**(B10), 14111–14125.

Wannamaker, P. E., Hasterok, D. P., Johnston, J. M., Stodt, J. A., Hall, D. B., Sodergren, T. L., Pellerin, L., Maris, V., Doerner, W. M., Groenewold, K. A., and Unsworth, M. J., 2008. Lithospheric dismemberment and magmatic processes of the Great Basin-Colorado Plateau transition, Utah, implied from magnetotellurics, *Geochemistry Geophysics Geosystems*, **9**, Art. No. Q05019.

Zhdanov, M., and Tolstaya, E., 2004. Minimum support nonlinear parametrization in the solution of a 3D magnetotelluric inverse problem. *Inverse Problems*, **20**(3), 937–952.

Cross-references

Electrical Properties of Rocks
Electrical Resistivity Surveys and Data Interpretation
Geoelectromagnetism
Instrumentation, Electrical Resistivity
Instrumentation, EM
Inverse Theory, Global Optimization
Inverse Theory, Linear
Inverse Theory, Monte Carlo Method
Inverse Theory, Singular Value Decomposition
Magnetic Storms and Electromagnetic Pulsations
Magnetotelluric Data Processing

MAGNETOVARIATION STUDIES

Nandini Nagarajan
National Geophysical Research Institute, Hyderabad, India

Definition and introduction

Magnetovariation studies identify the induced component of magnetic variations, from a spatial array of magnetic variation measurements and thereby infer the distribution of subsurface electrical conductivity in the Earth, through a linear transfer function between inducing and induced variations. These induction studies, which investigate the secondary magnetic fields induced in the solid Earth by external current systems, are undertaken to provide constraints on the electrical conductivity of the upper 1,000 km of the finitely conducting earth.

The techniques involved in magnetic variation studies are as follows:

Measurement of magnetic variations over an array of stations

Variations of the earth's magnetic field, caused by processes in the ionosphere and beyond, viz., external variations, and the secondary fields induced by them in the Earth are recorded at observatories or at magnetometer stations explicitly set up for the purpose (see *Geomagnetic Field, Measurement Techniques*). The sampling interval and lengths of data recorded are determined by the depth of investigation desired, limitations of instrumentation, size of the array, etc. Electric currents flow through the Earth's magnetosphere and ionosphere causing a wide spectrum of external magnetic variations ($\sim$ 1 s –1 day), (see *Magnetic Methods, Surface*). These magnetic variations induce secondary electric currents in the conducting layers of the Earth below the source current. The presence of such conducting layers is identified from spatial gradients in magnetic variations measured at that point or nearby stations (Egbert, 2002). The normal field in spectral content is defined as the sum of contributions from the external source field and that part of the internal field that is due to the regional (1D, radial variation) electrical conductivity structure (Parkinson, 1987).

Spectral analysis of magnetic variations to determine sufficiency of signal content based on assumptions regarding uniform inducing fields

Spectral content of magnetic variation is estimated, for desired lengths of time windows, after which, spectra are smoothed, stacked over several time windows, and significant peaks in frequency bands are estimated. To compute these spectra from magnetic variations at a single site, certain assumptions are made about both the external sources and the underlying conductivity distribution (Egbert, 2002). The external variations are generated over large-scale lengths relative to the area of the study and the incident magnetic fields are assumed to be a plane wave. For very large arrays, some modifications to the analysis are required since this assumption may not be valid. Since a large volume of data is collected, where the properties of external variations may be nonstationary, a careful selection of data intervals of suitable characteristics, conforming to the assumptions above, is made. Several techniques to eliminate effects of nonuniform, time-varying characteristics of inducing fields have also been evolved (Egbert, 2002). Further, this spectral analysis provides insights into the character of external variations and their measurement over large spatial extent on Earth. Spectral estimates are also examined to eliminate sources of nonuniformity and man-made noise.

The geometry, intensity, and period of induced anomalous variations depend both on the source and on the distribution of electrically conducting materials within the Earth. Inducing external magnetic variations cover a wide spectrum from minutes to days, thus creating induced responses over a wide spectral range. The longer period responses derive from deeper penetration of induced currents within the crust/mantle. The spectral content of magnetic variations is therefore a useful natural tool to probe the earth's interior.

Computation of transfer functions

For magnetovariation studies, lateral gradients of the magnetic fields are used to obtain estimates of electrical conductivity contrasts in the subsurface. The induced, anomalous field is generated by telluric currents flowing in non-one-dimensional conductivity structures (2D/3D) and, thus, entirely of internal origin. This anomalous field is not uniform and affects both the horizontal and the vertical components – evident as spatial gradients, and may be subjected to noise from man-made sources. Transfer functions, usually estimated by least squares, indicate the relation between inducing and induced variations at a single site, multiple sites, and for signals not recorded simultaneously (hypothetical event analysis). Improved methods of determining transfer functions using robust statistical approaches that incorporate the actual properties of the data, rather than ideal distributions, have refined estimation of transfer functions, improving their reliability. Simultaneous observations also provide information on the spectral space of the data, in each array study, and indicate whether new or modified interpretation parameters are to be evolved for a particular study. Fitting these with a formal inversion procedure, along with more traditional single site parameters such as MT impedances and vertical field TFs, can enhance resolution of some features (Simpson and Bahr, 2005).

Modeling to obtain induced magnetic variations from the conductivity distribution: Interpret in terms of geology and structure

Magnetovariation studies primarily delineate contrasts in electrical conductivity in the subsurface. The depth and

conductivity are poorly resolved. Therefore this method is widely used as preliminary investigation of large- and small-scale anomalies, which may later be intensively investigated using more precise electromagnetic methods (e.g., Magnetotellurics). The conductivity anomalies investigated using magnetovariation often extend over thousands of kilometers across continents, correlate tectonic signatures to deeper crust/mantle properties, and link events of the vast geologic timescale. Many examples of continental-scale structures, as well as sharply defined shallow crustal anomalies that are directly related to tectonic events (faulting, seismicity) are discussed in good reviews (see *Earth's Structure, Continenal Crust*; *Electrical Resistivity Surveys and Data Interpretation*; *Magnetotelluric Data Processing*; *Magnetotelluric Interpretation*; *Electrical Properties of Rocks*) (Gough, 1989). The principles of electromagnetic induction have also been applied to satellite data of denser spatial coverage. Here, the forward response of conductivity distribution is directly modeled and fitted to potential data. In this manner, induced contributions to crustal magnetization of continents and ocean crust have been estimated (Hemant and Maus, 2005).

Interpretation

Magnetometer array studies provide geometry of induced currents, in the Earth's subsurface. The interpretation is not confined to any predetermined shape or single conductive feature, rather delineates broadly any lateral inhomogeneities in the upper 1,000 km of the Earth's crust/mantle. The conductivities are poorly constrained in this method and need to be verified using other methods (e.g., MT).

The significance of these electrical conductivity contrasts are threefold:

1. Indicate change in composition within depth of penetration, more graphite/conducting minerals
2. Changes in temperature, at appropriate depth
3. Indicates accumulation of fluids, interstitial or within the metamorphic mineral structures (Jones, 1992).

The above conditions and geometry of the conductivity anomaly are evaluated or eliminated, for the region under study. The significance of any or either of these conditions can impact the tectonics and evolutionary history of the region. The complementary nature of conductivity anomalies and other geophysical studies were established in the course of several Geotransects over ~1,000 km continental-scale studies. The role of fluids in the crust has been a major contribution of these studies (see *Earth's Structure, Continental Crust*; *Heat Flow, Continental*; *Seismic Velocity-Temperature Relationships*; *Isostasy, Thermal*; *Electrical Properties of Rocks*; *Differential Rotation of the Earth's Inner Core*; *Lithosphere, Continental: Thermal Structure*).

Summary

At the Earth's surface, magnetometers measure the composite of external (from the external source currents) and internal (from the induced currents) field components. The observed magnetic field can be considered as the sum of a normal plus an anomalous induced field. Induced currents in the finitely conducting earth manifest as gradients in magnetic variations over spatial scales of ~10–1,000 km.

The interpretation of electrical conductivity anomalies in the crust, provide additional insights into the nature of the Earth's crust. In the upper crust (<15 km), most conductivity anomalies were correlated with deeper extensions of tectonic elements (faults, sutures) and compositional changes. Simultaneously, the phenomenon of seismic reflectors within the lower crust were found and combined with this, the significant role of fluids in the lower crust/upper mantle came to be recognized. A number of Geotransects across continents were carried out in the 1990s and the complexities of evolutionary history on the lower crust were revealed. The relationship between electrical conductivity, high heat flow, seismic parameters, and tectonic history were formulated (Jones, 1992). Induction studies and later magnetotelluric soundings have contributed to an understanding of processes in the Earth's crust and upper mantle. This was also possible because of the refinements in instrumentation, data processing, and computational approaches in magnetovariation studies.

Bibliography

Artemieva, I. M., and Mooney, W. D., 2001. Thermal structure and evolution of Precambrian lithosphere: A global study. *Journal of Geophysical Research*, **106**, 16387–16414.

Egbert, G. D., 2002. Processing and interpretation of electromagnetic induction array data. *Surveys in Geophysics*, **23**, 207–249.

Gough, D. I., 1989. Magnetometer array studies, Earth structure, and tectonic processes. *Reviews of Geophysics*, **27**(1), 141–157, doi:10.1029/RG027i001p00141.

Hemant, K., and Maus, S., 2005. Geological modeling of the new CHAMP magnetic anomaly maps using a geographical information system technique. *Journal of Geophysical Research*, **110** (B12103), 23, doi:10.1029/2005JB003837.

Jones, A., 1992. Electrical conductivity of the lower continental crust. In Fountain, D. M., Arculus, R., and Kay, R. W. (eds.), *Continental Lower Crust*. Amsterdam: Elsevier, pp. 81–143.

Mooney, W. D., and Meissner, R., 1992. Multi-genetic origin of crustal reflectivity: continental lower crust and Moho. In Fountain, D. M., Arculus, R., and Kay, R. W. (eds.), *Continental Lower Crust*. Amsterdam: Elsevier, pp. 45–79.

Parkinson, W. D., 1987. The global conductivity distribution. *Surveys in Geophysics*, **9**, 235–243.

Schmucker, U., 1970. Anomalies of geomagnetic variations in the Southwestern United States. *Bulletin of the Scripps Institution of Oceanography*. San Diego: University of California, Vol. 13, 165 pp.

Simpson, F., and Bahr, K., 2005. *Practical Magnetotellurics*. Cambridge: Cambridge University Press, p. 270.

Cross-references

Deep Seismic Reflection and Refraction Profiling
Earth's Structure, Continental Crust
Electrical Properties of Rocks
Electrical Resistivity Surveys and Data Interpretation

Geoelectromagnetism
Geomagnetic Field, Measurement Techniques
Heat Flow, Continental
Lithosphere, Continental
Lithosphere, Continental: Thermal Structure
Magnetic Storms and Electromagnetic Pulsations
Magnetotelluric Data Processing
Magnetotelluric Interpretation
Magnetovariation Studies
Seismic Velocity-Temperature Relationships

MANTLE CONVECTION

David Bercovici
Department of Geology & Geophysics, Yale University, New Haven, CT, USA

Synonyms

Mantle circulation; Mantle dynamics

Definition

Mantle convection. Thermal convection in the terrestrial planetary mantles, the rocky layer between crust and core, in which hot material rises, cold material sinks, and the induced flow governs plate tectonic and volcanic activity, as well as chemical segregation and cooling of the entire planet.

Mantle convection

Introduction and history

All planetary bodies retain some heat from their early formation but are inexorably cooling to space. Planetary surfaces are therefore cold relative to their hotter interiors, and thus undergo thermal convection wherein cold material is dense and sinks while hot material is light and rises (liquid water near freezing being one of the rare exceptions to this process). Planetary atmospheres, oceans, rocky mantles, and metallic liquid cores convect and are subject to unique patterns of circulation in each domain. Silicate mantles however tend to be the most massive and sluggish part of terrestrial planets and therefore govern how planetary interiors evolve and cool to space (See Figure 1).

The theory of mantle convection was originally developed to understand the thermal history of the Earth and to provide a driving mechanism for Alfred Wegener's theory of Continental Drift in the 1930s (see Bercovici, 2007; Schubert et al., 2001). Interest in mantle convection waned for decades as Wegener's theory was criticized and apparently discredited. However, the accumulation of sea-floor sounding data during War World II and refinement of paleomagnetic techniques paved the way for the discovery of sea-floor spreading (Hess, 1962; Vine and Matthews, 1963) and the birth of the grand unifying theory of plate tectonics in the 1960s; this consequently revived interest in mantle convection as the driving mechanism for plate motions (Runcorn, 1962a, b) as well as non-plate-tectonic volcanism such as the possible Hawaiian plume (Morgan, 1971). The success of mantle convection theory in explaining plate velocities, sea-floor subsidence, volcanism, gravity anomalies, etc., led to its further application to other terrestrial planets such as Venus and Mars, which also sustained unique forms of mantle convection, evident from volcanic activity.

Basics of thermal or free convection

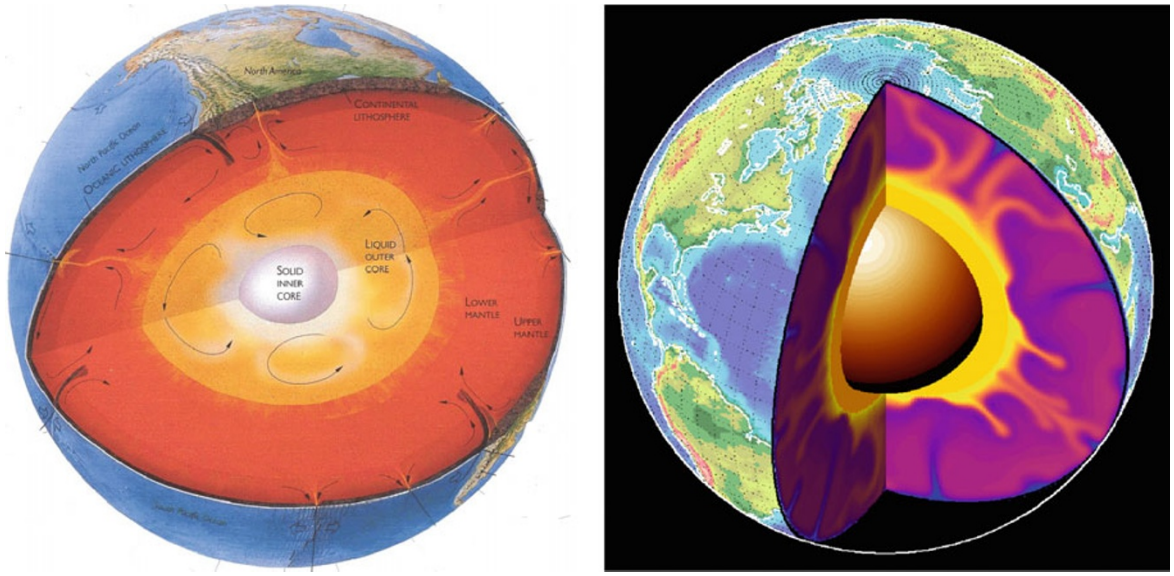
Rayleigh–Bénard convection

The simplest form of thermal convection is referred to as Bénard convection named after the French experimentalist Henri Bénard who in 1900 performed the first systematic experiments on convection in thin layers of oil (spermaceti) and recognized both the onset of convection from a static conductive state and the regular patterns formed in a convecting layer (Bénard, 1900, 1901). Fifteen years later, the British theoretical physicist and mathematician Lord Rayleigh (William John Strutt), attempted to explain Bénard's results for the onset of convection (Strutt, 1916) – the delay in communication between them being caused by World War I. However, the mismatch between theory and experiment was profound, and not resolved until the late 1950s (Pearson, 1958) when it was inferred that Bénard's experiments were strongly influenced by surface tension or Marangoni effects not included in Rayleigh's theory (although Bénard himself was aware of these effects). Because Rayleigh's work provided the framework for nearly all thermal convection theory to follow, the simple Bénard convective system is also referred to as Rayleigh–Bénard convection.

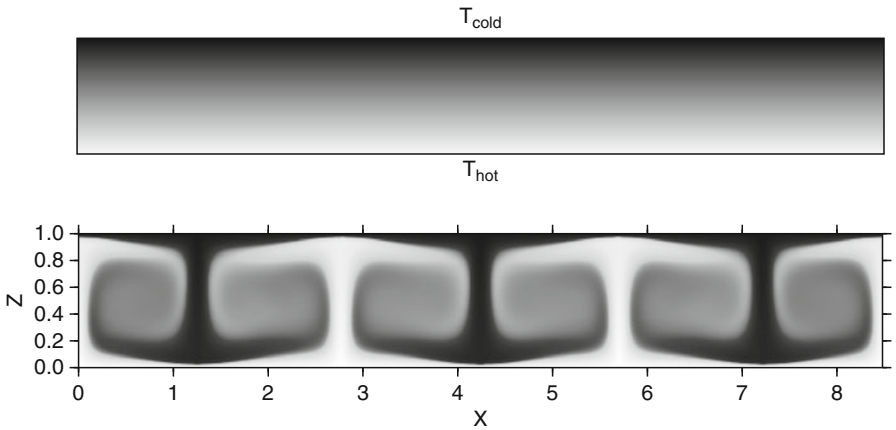
Although Bénard's experiments were in a metal cavity, Rayleigh–Bénard convection actually refers to Rayleigh's idealized model of a thin fluid layer infinite in all horizontal directions such that the only intrinsic length scale in the system is the layer thickness. The Rayleigh–Bénard system is heated uniformly on the bottom by a heat reservoir held at a fixed temperature (i.e., the bottom boundary is everywhere isothermal) and the top is likewise held at a fixed colder temperature by another reservoir (see Figure 2). If the layer were not fluid, heat would flow from the hot boundary to the cold one by thermal conduction. But since the fluid near the hotter base is (typically) less dense than the fluid near the colder surface, the layer is gravitationally unstable, that is, less-dense material underlies more-dense material. To release gravitationally potential energy and go to a minimum energy state, the layer is induced to turn over.

Convective onset and the Rayleigh number

While the fluid in a Rayleigh–Bénard layer might be *gravitationally* unstable, it is not necessarily *convectively* unstable. Convective overturn of the layer is forced by heating but resisted or damped in two unique ways. Clearly, the thermal buoyancy (proportional to density contrast times gravity) of a hot fluid parcel rising from the bottom surface through colder surroundings acts to



Mantle Convection, Figure 1 Graphic renditions of cut aways of Earth's structure showing crust, mantle, and core (*left*), and of the convecting mantle (*right*). The relevant dimensions are that the Earth's average radius is 6,371 km; the depth of the base of the oceanic crust is about 7 km and continental crust about 35 km; the base of the lithosphere varies from 0 km at mid-ocean ridges to about 100 km near subduction zones; the base of the upper mantle is at 410 km depth; the transition zone sits between 410 and 660 km depths; the depth of the base of the mantle (the core–mantle boundary) is 2,890 km; and the inner core–outer core boundary is at a depth of 5,150 km. (*Left frame* adapted from Lamb and Sington (1998). *Right frame*, provenance unknown).



Mantle Convection, Figure 2 Rayleigh–Bénard convection: initially conducting layer (*top*), and numerical simulation of convection (*bottom*).

drive convective overturn. However, viscous drag acts to slow down this parcel, and thermal conduction, or diffusion, acts to erase its hot anomaly (i.e., it loses heat to its colder surroundings). Thus while the fluid layer might be gravitationally unstable, hot parcels rising might move too slowly against viscous drag before being erased by thermal diffusion. Similar arguments can be made for cold material sinking from the top surface through warmer surroundings. The competition between forcing by thermal buoyancy, and damping by viscosity and thermal diffusion, is characterized in dimensionless ratio called the Rayleigh number

$$Ra = \frac{\rho g \alpha \Delta T d^3}{\mu \kappa} \tag{1}$$

where ρ is fluid density, g is gravity, α is thermal expansivity (units of K^{-1}), ΔT is the difference in temperature

between the bottom and top surfaces, d is the layer thickness, μ is fluid viscosity (units of Pa s), and κ is fluid thermal diffusivity (units of m^2/s).

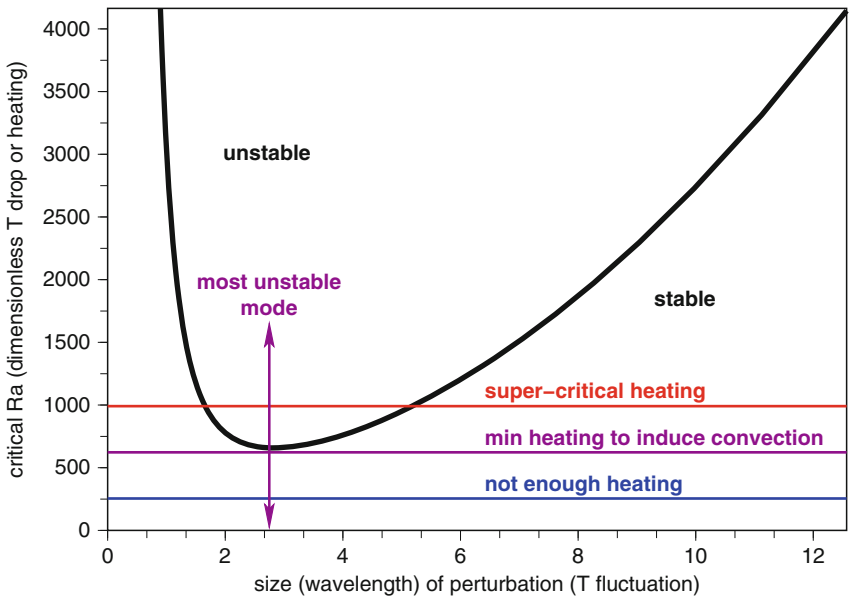
Even though $\Delta T > 0$ (i.e., heating is from below and causes gravitational instability), Ra still must exceed a certain value, called the critical Rayleigh number Ra_c, for convection to occur. For $Ra < Ra_c$ the layer is stable and transports heat by conduction; for $Ra > Ra_c$ the layer will be convectively unstable and transport heat more rapidly via convection. See Figure 3.

Although Ra_c varies depending on the mechanical nature of the horizontal boundaries (whether rigid or a free surface), it is typically of order 1,000. This value is easily understood by considering the fate of a hot (or cold) parcel of size a and temperature anomaly ΔT. Dimensional analysis readily shows that the typical ascent rate of the parcel is $\rho g \alpha \Delta T a^2/\mu$ (with units of m/s). However, the rate that heat diffuses out of the parcel is κ/a (smaller parcels lose heat faster). The critical state occurs when these two rates are equal; that is, if the buoyant ascent rate just exceeds the diffusion rate, the parcel should rise without being erased, but if the ascent rate is less than the diffusion rate it will not rise very far before being lost. Therefore, the critical state occurs if $\rho g \alpha \Delta T a^3/(\mu\kappa) \approx 1$. Scaling purely by orders of magnitude, a small parcel of fluid can be assumed to be of an order ten times smaller than the entire layer; thus assuming $a \approx d/10$ leads to a critical condition for the onset of convection of $\rho g \alpha \Delta T d^3/(\mu\kappa) \approx 1{,}000$.

For the Earth's mantle, the typical average properties from which the Rayleigh number is constructed are $\rho \approx 4{,}000$ kg/m^3, $g = 10$ m/s^2, $\alpha = 3 \times 10^{-5}$ K^{-1}, $\Delta T \approx 3{,}000$ K, $d = 2{,}900$ km, $\mu = 10^{22}$ Pa s (dominated by the lower mantle), and $\kappa = 10^{-6}$ m^2/s (see Schubert et al., 2001). Taken together, these lead to a Rayleigh number of approximately 10^7, which is well beyond super-critical; although the mantle viscosity is extremely high, the mantle is also very hot and very large, and hence convecting vigorously.

Thermal boundary layers and the Nusselt number

For a Rayleigh number Ra above the critical value Ra_c, convective circulation will mix the fluid layer, and the mixing and homogenization of the fluid will become more effective the more vigorous the convection, that is, as Ra is further increased. With very large Ra and vigorous mixing, most of the layer is largely uniform and isothermal. (In fact, if the layer is deep enough such that the pressures are comparable to fluid incompressibility, the fluid layer is not isothermal but adiabatic, wherein even without any heating or cooling, the temperature would drop with fluid expansion on ascent and increase with fluid compression on descent.) Most of the fluid in the Rayleigh–Bénard system is at the mean temperature between the two boundary temperatures. However, the temperature still must drop from the well-mixed warm interior to the cold temperature at the top, and to the hotter temperature at the

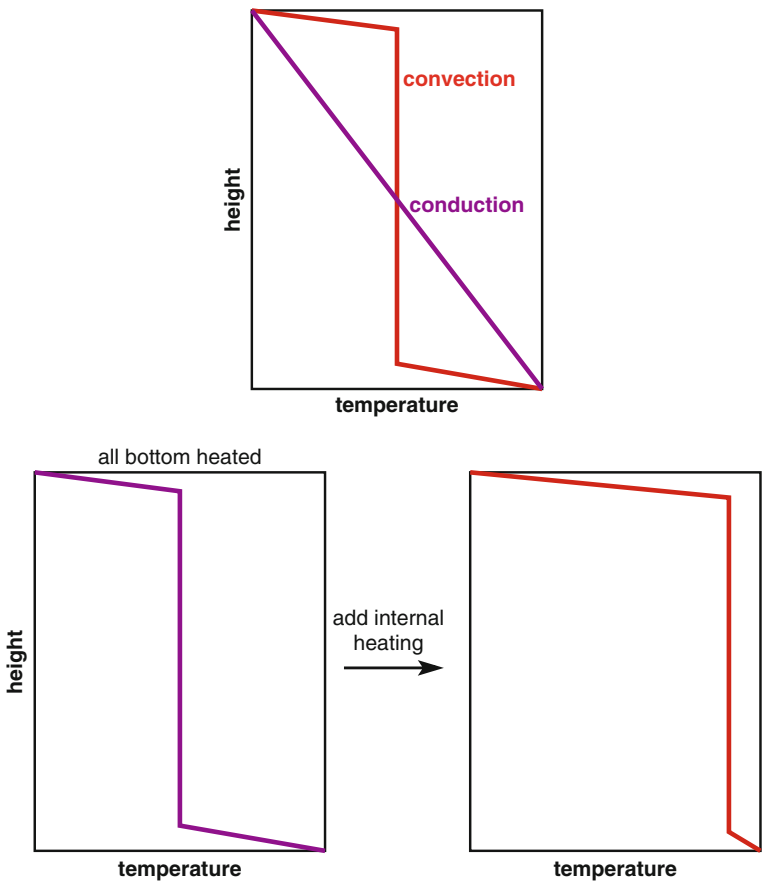


Mantle Convection, Figure 3 The critical Rayleigh number Ra for the onset of convection is a function of wavelength or size of the thermal perturbation to a static conductive state. For a layer with isothermal and free-slip (shear-stress free) top and bottom boundaries, the relationship is $Ra_c = (k^2 + \pi^2)^3/k^2$ where $k = 2\pi/\lambda$ and λ is wavelength (Chandrasekhar, 1961). Values of Ra above the Ra_c curve are associated with the conductive layer being convectively unstable (perturbations grow), while below the curve the layer is stable (perturbations decay). The minimum in the Ra_c curve occurs at the wavelength of the first perturbation to go unstable as heating and Ra is increased, often called the most unstable mode.

bottom. The narrow regions accommodating these jumps in temperature are called thermal boundary layers (Figure 4).

Thermal boundary layers are of great importance in thermal (and mantle) convection for two reasons. First, most of the buoyancy of the system is bound up in thermal boundary layers since these are where most of the cold material at the top and hot material at the bottom reside and from where hot and cold thermals or currents emanate. Moreover, with regard to convection in the mantle itself, the top cold thermal boundary layer is typically associated with the Earth's lithosphere – the 100 km or so thick layer of cold and stiffer mantle rock that is nominally cut up into tectonic plates.

Second, since fluid in these boundary layers is near a horizontal boundary, most of the fluid motion is horizontal and thus heat is only transported vertically across these thin layers by conduction; but since the layers are very thin, such conductive transport is rapid. Indeed, the entire cycle of heat transport occurs by heat conducted in rapidly through the bottom boundary layer, after which hot fluid in this layer will, in various spots, become unstable and rise to form a convective upwelling that carries heat out of the boundary layer rapidly across most of the fluid layer and deposits it at the upper boundary, where the heat is then transported out of the system by conduction across the top boundary layer. The eventual heat flow (power output per unit area) out of the well-mixed layer is essentially $k\Delta T/\delta$ where k is thermal conductivity (units of W K^{-1} m^{-1}), $\Delta T/2$ is the temperature drop from the well-mixed interior to the surface and we define $\delta/2$ is the thickness of the boundary layer. By comparison, the thermal conduction across a static non-convecting layer is $k\Delta T/d$. The ratio of heat flow in the convectively well-mixed layer to the purely conductive layer is thus d/δ, which is called the Nusselt number Nu (named after the German engineer, Wilhelm Nusselt 1882–1957). The relation between Nu and convective vigor parameterized by Ra is important for understanding how convection transports heat and cools off bodies including planets.



Mantle Convection, Figure 4 Sketch of temperature profiles, showing how convective mixing homogenizes the conductive mean temperature into a nearly isothermal state (if the fluid is incompressible) with thermal boundary layers connecting it to the cold surface and hot base (*top frame*). With no internal heating the interior mean temperature is the average of the top and bottom temperatures; the effect of adding internal heating (*bottom frames*) is to increase the interior mean temperature and thus change the relative size and temperature drop across the top and bottom thermal boundary layers.

Convective heat transport is often written as $Nu(k\Delta T/d)$ and in considering this relation Howard (1966) argued that vigorous convective heat transport across the bulk of the well-mixed fluid layer is so fast that it is not the rate limiting factor in releasing heat (only conduction across the thermal boundary layer is), and thus heat flow should be independent of fluid depth d; this implies that since $Ra \sim d^3$, $Nu \sim Ra^{1/3}$, which yields a convective heat flow $Nu(k\Delta T/d)$ that is independent of d. In general, since the fluid is conductive for $Ra \leq Ra_c$, one often writes that $Nu = (Ra/Ra_c)^{1/3}$ (although $Nu = 1$ for $Ra < Ra_c$), which is a reasonably accurate relationship borne out by simple experiments and computer modeling. This relationship also implies that the ratio of thermal boundary width to fluid layer depth is $\delta/d \sim Ra^{-1/3}$, which shows that the boundary layers become increasingly thin as convective mixing of the layer becomes more vigorous.

The relation of $\delta \sim Ra^{-1/3}$ applies to the horizontally averaged boundary layer thickness. However, boundary layers change with time or distance from their first formation, for example, where an upwelling impinges on the top boundary. As the fluid in the boundary layer moves from an upwelling to a downwelling it cools and the boundary layer thickens as more material cools next to the cold surface. The thickening depends on the thermal diffusivity κ (with units of m^2/s) and the residence time or age t near the cold boundary (i.e., time since leaving the upwelling). Simple dimensional considerations show that the boundary layer thickness goes as $\sqrt{\kappa t}$; this corresponds to the well-known $\sqrt{\text{age}}$ law for subsidence of ocean sea floor with age since formation at mid-ocean ridges, implying that sea floor gets deeper because of the cooling and thickening lithosphere.

Patterns of convection, structure of upwellings and downwellings: plumes and slabs

When convection occurs, upwellings and downwellings will be separated horizontally by some optimal distance. If they are too close to each other, they can induce too much viscous drag on each other and/or lose heat rapidly to each other; if they are too far apart, they must roll too much mass between them. The separation distance is also determined by heat transport in the thermal boundary layer between upwellings and downwelling. When hot upwelling fluid reaches the surface, it spreads laterally into the thermal boundary layer. As it travels horizontally, it cools to the surface and eventually gets cold and sinks into the downwelling; thus the upwelling-downwelling separation is also determined by the distance it takes for fluid to cool off and become heavy enough to sink.

The upwelling-downwelling separation distance or convection cell size is predicted by convective stability theory to be approximately equal to the layer depth d (a bit larger at the onset of convection but identically d as Ra becomes very large); that is, the cell that is either least stable and thus most likely to convect – or equivalently the cell that optimizes gravitational potential energy (and thus heat) release – is usually the cell that is as wide as it is deep.

Viewing a convecting layer from above, the upwelling and downwellings may be separated in various patterns, such as 2-D rolls (sheets of upwelling rolling over into sheets of downwelling, and each cell counter-rotating with its neighboring cell). In the Rayleigh–Bénard layer, which is infinite horizontally, no one location of the layer should be different than any other one, and so ideally the pattern should be a regular repeating tile; as there are only so many polygons that can form a repeating tile, the patterns usually involve convection cells in the shapes of rolls (already mentioned), squares, hexagons, and triangles (Figure 5). Of course nonideality and irregularities can occur due to tiny imperfections, for example, in the boundaries, leading to irregular patterns.

In many instances, in the 3-D pattern of convection, especially in fluids where hot fluid is less viscous than cold fluid (as is true in many materials, including the mantle), the upwelling is in the form of a cylindrical plume at the center of a canopy of sheet-like downwellings, much like a fountain.

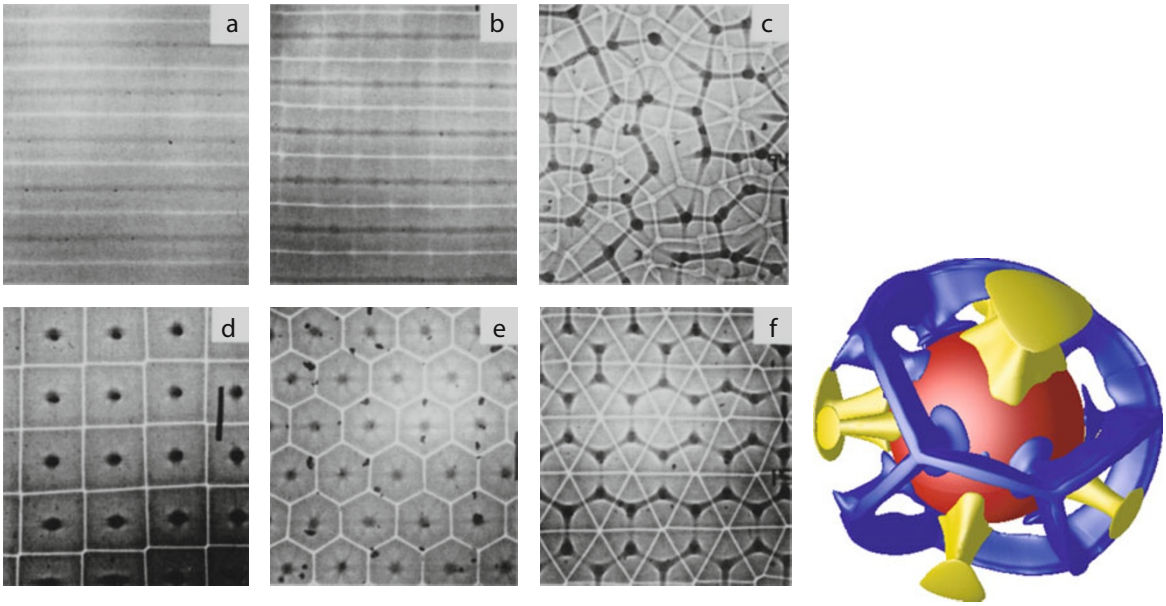
Plumes and slabs in the mantle: simple view

The common occurrence of sheet-like downwellings and columnar upwellings in simple convection is crudely applicable to mantle convection. Subducting slabs are where tectonic plates sink into the mantle; these are analogous to the cold, sheet-like downwellings seen in 3-D convection, although much more complicated by rheology as discussed below. Deep hot narrow upwelling plumes are inferred to explain anomalous intraplate volcanism as in Hawaii, as well as the fixity of these volcanic hotspots relative to each other (which suggests deep anchoring in the mantle). These mantle plumes are ostensibly analogous to the pipe-like upwelling in simple 3-D convection, but again more complicated by unique mantle properties. (See *Mantle Plumes* entry by Farnetani and Hofmann.)

While mid-ocean ridges or spreading centers transport material vertically to surface, they are very narrow features and for the most part involve shallow upwelling (inferred from their weak gravity anomalies that suggest shallow isostatic compensation, that is, they are floating on a shallow buoyant root). Ridges are likely best explained as being pulled and rifted apart passively from a distant force (ostensibly slabs) rather than involving a deep convective upwelling that pries them open.

Energy sources for mantle convection and Earth's thermo-chemical history

Heatflow coming out of the Earth is measured by heatflow gauges (measuring conductivity of rocks first and then thermal gradients in boreholes) both in continents and oceans (see Turcotte and Schubert, 1982). The total heat flowing out from beneath the Earth's surface is approximately 46 TW (46 trillion Watts) (Jaupart et al., 2007), which is in fact not a large number given the surface area of the Earth, and is actually tens of thousands



Mantle Convection, Figure 5 Patterns of convection from laboratory experiments in rectangular tank by White (1988) (*left*), and numerical simulations in a spherical shell by Zhong et al. (2000) (*right*).

of times smaller than the heat absorbed from the Sun. Nevertheless, it represents the source of power driving dynamics inside our planet, including mantle convection (and hence tectonic, volcanological, and other geological activity) as well as core cooling and flow.

The source of the Earth's internal heat is a combination of primordial heat, that is, left over from accretion (gravitational potential energy from formation and collisions), and heat generated by unstable radioactive isotopes, especially the isotopes of uranium (^{238}U), thorium (^{232}Th), and potassium (^{40}K), although the ^{40}K half-life is much shorter than the others and so generated a large heat pulse primarily in the early Earth. Because continental crust is originally formed by partial melting and chemical segregation of early mantle material (indeed the chemical separation allows another energy source in the release of gravitational potential energy, but other than early core formation this is a relatively minor contribution), these radioactive elements tend to be concentrated in crust (i.e., melt more readily dissolves these elements than does solid rock, so they partition toward the melt). Thus the crust itself produces a significant fraction of the net heat output through the surface; removing the crustal component leaves approximately 31 TW emanating from the mantle and core (Jaupart et al., 2007; Schubert et al., 2001).

The relative contributions of primordial cooling and radiogenic heating to the mantle (and core) heat output remains an active area of debate even today and leads to various quandaries. The most direct estimate of the concentration of radiogenic sources (U, Th, K) is by looking at the concentration in chondritic meteorites, which come from the solar systems' main asteroid belt, and because they have been largely unaltered for 4.5 Gyr (i.e., unmelted), they are thought to be the same as the original building blocks of the terrestrial planets. The chondritic concentrations of U, Th, and K would allow for radiogenic heating to contribute 50% or less of the total heat output (Korenaga, 2008); this is often called the Urey ratio, that is, the radiogenic heat output to the total output. With radiogenic heating this small, the only way the mantle could be as hot as it is today – while also transporting heat as it does presently – is if it were very hot and nearly molten in the geologically recent past (a few hundred million years); this is geologically untenable since petrological and geochemical analysis demonstrate the presence of solid rock and even liquid water in the early Earth (see Halliday, 2001), which therefore ceased to be molten not long after its formation 4.5 billion years ago. This paradox has led some researchers to assume that the radiogenic sources are super-chondritic, that is, to allow heat to be produced continuously throughout the past rather than by rapid cooling from a molten state. Alternatively, researchers have sought ways to keep chondritic concentrations of U, Th, and K by arguing that heat transport in the past was different than it is today; for example, higher temperatures in the past might have allowed for more melting and thus more buoyant crust and/or stronger dehydrated lithosphere that kept the top part of the convecting mantle sluggish or immobile, hence bottling up primordial heat for much later release (Korenaga, 2008).

Finally, the release of both radiogenic and primordial heat, again termed collectively "internal heating" (i.e., heat coming from the bulk of the fluid) means that the idealized Rayleigh–Bénard model of convection, where

heating is only along the base, is inaccurate. The effect of internal heating in addition to "basal heating" is relatively straightforward to understand. While in the Rayleigh–Bénard model the temperature of the well-mixed interior is at the average of the two boundary temperatures, the addition of internal heating acts to heat up the well-mixed interior to a higher mean temperature. This puts the interior temperature closer to the hot bottom temperature, but further from the cold top temperature; the effect is to create a larger temperature drop across the top thermal boundary layer than across the bottom one; in essence, the top boundary layer must conduct out heat injected through the bottom plus heat generated or released from the interior (Figure 4). These very different boundary layers tend to cause more negatively buoyant and stronger downwelling currents and smaller and weaker upwelling currents; this effect seems to be borne out in the Earth by the presence of many large cold slabs driving plate tectonic motion with large thermal anomalies (of order 500 K) relative to fewer upwelling plumes with weaker (200 K) thermal anomalies (although this is still somewhat a matter of debate).

Effects of mantle properties

Mantle rheology

The entire mantle of the Earth is potentially a convecting and overturning fluid. However, the mantle is almost entirely solid (with some small portions of melting near the surface delineated by volcanism, and possibly much smaller areas of melting at depth) and thus flows extremely slowly. (Indeed, the term fluid does not suggest a liquid state but refers to how a medium deforms, as opposed to elastic or brittle deformation; these do not necessarily correlate with states of matter, that is, gas, liquid, solid, each of which can display, e.g., either fluid flow or elastic behavior depending on the nature of the deformation; e.g., atmospheric sound waves are elastic behavior in gas.)

The solid-state viscous flow of the mantle is a complex process and there are various mechanisms that permit such "irrecoverable" deformation (one of the definitions of viscous flow). A survey of mantle deformation mechanisms or "rheology" is beyond the scope of this entry (see Karato, 2008; Ranalli, 1995). However, in brief, the two primary deformation mechanisms are called diffusion and dislocation creep. In any solid-state creep mechanism, mobility depends on the statistical-mechanical probability of a molecule in a crystal lattice leaving the potential well of its lattice site; the potential well itself is defined by electrostatic or chemical bonds inhibiting escape, and Pauli-exclusion pressure preventing molecules squeezing too closely to each other. Thus mobility depends on the Boltzmann distribution measuring the probability of having sufficient energy to overcome the lattice potential well barrier, which is often called the activation energy (or allowing instead for pressure variations the activation enthalpy). This probability depends on the (Arrhenius) factor $e^{-E_a/RT}$ where E_a is the activation energy (J/mol), R is the gas constant (J/K/mol) and T is temperature; RT represents the thermal excitation energy of the molecule in the well. As T goes to infinity, the probability of escaping the well goes to 1, while as T goes to 0, the probability of escape goes to 0.

Stress imposed on the medium effectively changes the shape of the potential well, such that compressive stress steepens the walls of the well (squeezing molecules closer makes coulomb attraction in the chemical bonds stronger) while tension lowers the walls (separating molecules weakens the bonds); thus the probability of escape is preferred in the direction of tension and away from compression, thereby allowing the medium to stretch in the tensile direction by solid-state diffusion of molecules.

Simple diffusive creep works much in this way whereby differential stress (i.e., nonuniform stress) causes slow diffusion of molecules to allow the entire substance to deform accordingly. However, such deformation occurs at the grain or mineral level inside a rock, with either diffusion through the grains or along the grain boundaries; thus the response depends significantly on grain size.

Dislocation creep is more complicated. A dislocation can be in the form of a truncated row of molecules in a crystal lattice; shortening of the crystal under compression perpendicular to the row can be accomplished by simply removing that row. Thus differential compressive stress would act to force molecules to diffuse out of that dislocated row into other parts of the lattice. However, stress not only governs the preferential diffusion of molecules (as in diffusion creep) but also the geometry of the dislocations (their spacing and directions), hence the multiple actions of stress are compounded into a nonlinear response.

The viscosities for diffusion and dislocation creep mechanisms can be written as

$$\mu = \begin{cases} Ba^m e^{\frac{E_a}{RT}} & \text{for diffusion creep} \\ A\sigma^{1-n} e^{\frac{E'_a}{RT}} & \text{for dislocation creep} \end{cases} \quad (2)$$

where A and B are proportionality constants, a is grain size, σ is stress (in fact since stress is a tensor σ is the scalar second invariant of the stress tensor), and m and n are exponents, typically both equal to 3 (with variations depending on exact deformation mechanisms; see Karato, 2008). It should be emphasized that diffusion and dislocation creep occur independently of each other depending on stress and grain size: for high stress and large grains dislocation creep dominates; for low stress and small grains diffusion creep dominates.

Dislocation creep allows for moderate softening as stress increases; diffusion creep potentially allows for significant softening if stress can reduce grain size, although mechanisms to allow this are still controversial (see section below on generating plates), and significant hardening via grain growth by standard coarsening of the material (i.e., what happens to all grained materials under the action of grain-surface energy reduction).

The strongest rheological effect is clearly that of temperature; the temperature dependence of viscosity allows for many orders of magnitude variations in viscosity. For example, while this rheological effect allows subducting slabs to keep their strength and integrity to great depths as they sink, it would make hot upwelling mantle plumes more fluid and, if they have a conduit structure, the plume flow would be relatively rapid, of order 100 cm/year or more. However, this effect is most profound in the cold top thermal boundary layer or lithosphere. If viscosity is strongly temperature dependent, as it is in Earth, the lithosphere can become so stiff that it becomes immobile; in this case, convection in the mantle would proceed beneath the lithosphere, which in turn would act like a rigid lid to the mantle (Solomatov, 1995). If mantle rheology obeyed only diffusion or dislocation creep laws, then the lithosphere should be locked and immobile, and there should be no plate tectonics. While this scenario might be relevant for Venus and Mars (and Moon and Mercury) which have no plate tectonics, obviously it is missing a vital ingredient to allow plate tectonics on Earth. This paradox underlies the fundamental question and mystery about why Earth has plate tectonics at all and how it is generated on our planet but not others in our solar system (Bercovici, 2003).

Compressibility, melting, and solid phase changes

Pressures deep inside the Earth's mantle are so large they are sizable fractions of rock incompressibility or bulk modulus (e.g., mantle pressures reach 140 GPa, or 1.4 million atmospheres, while bulk modulus – which has the same units as pressure – are typically a few to several 100 GPa). As downwelling mantle material travels from near the surface to the base of the mantle its density and temperature increase due to compression, called "adiabatic compression and heating" (and likewise upwelling material undergoes "adiabatic decompression and cooling"), although these increases are not large (of the order of several degree Celsius). The compression and decompression of circulating material establishes a weak adiabatic temperature and density increase with depth, which has a slight stabilizing effect on convection; however, because the mantle is so viscous, the thermal anomalies needed to get it to move – and in particular the temperature variations across the thermal boundary layers – are so large (of order several 100–1,000°C) that the adiabatic variations are small in comparison.

Where compressibility and pressure play a dual important role is in phase changes. First, as hot upwelling mantle material approaches the surface, it actually travels along a gradually cooling adiabatic temperature profile. The upwelling does eventually melt when it gets near the surface but not because it gets hotter. Melting occurs because the melting temperature T_m drops with decreasing pressure faster than the upwelling adiabat (in essence, decreasing confining pressure makes it easier for molecules to mobilize into a melt); thus at a certain (usually shallow depth of a few tens of kilometers to 100 km) the upwelling mantle crosses the melting temperature from solid to liquid phase and undergoes melting; however, the mantle is not a single pure substance so in fact only partially melts. Such "pressure-release" melting is a shallow process but is vital for chemical segregation of the mantle and development of oceanic and continental crust. In particular, melting is sequential in that the most easily melted material (usually more silica-rich material with lower melting temperature) melts first, freezes last, and is typically chemically less dense, and thus comes to the surface as lighter crust eventually gathers, after more weathering and reactions into continental crust. (Continental crustal rocks like sandstone and granite have typical densities of 2,300 and 2,700 kg/m^3, respectively.) The more refractory (harder to melt, silica poor, and heavier) material melts last, freezes first, and either stays in the mantle or lithosphere or sits in the heavy basaltic oceanic crust. (Oceanic crustal rocks like basalt have densities of 3,000 kg/m^3, while mantle peridotites at near-surface pressure have densities of around 3,400 kg/m^3.)

Extreme pressures with depth can also overcome a mineral's elastic resistance to compression and cause solid–solid phase changes where the minerals change their crystallographic structure to a more compact and incompressible state (but of course their chemistry remains the same). Such mineralogical phase changes have been observed in laboratory experiments in olivine, which is the major component mineral of the upper mantle (at about 60% by weight, the remainder being mostly pyroxene at shallow depths, and garnet at slightly greater depth); moreover, the pressures at which they are predicted to occur have been verified seismologically, wherein the seismic wave speeds and density undergo a jump at the predicted pressures. The first major phase change to occur with depth is from olivine to the same material with a wadsleyite structure, at 410 km depth. Wadsleyite changes slightly to a similar ringwoodite structure at 510 km depth. The largest phase change occurs from ringwoodite to perovskite/magnesiowüstite at 660 km depth.

The 410 and 660 km phase changes are the two most remarkable and global phase changes in the mantle, and the region between them is called the transition zone, since it is where most of the mineralogical transitions occur, over a relatively narrow region. The mantle above the transition zone is typically identified as the Upper Mantle, although in some papers and books Upper Mantle includes the transition zone. Below the transition zone is the Lower Mantle and that is universally agreed upon in the literature.

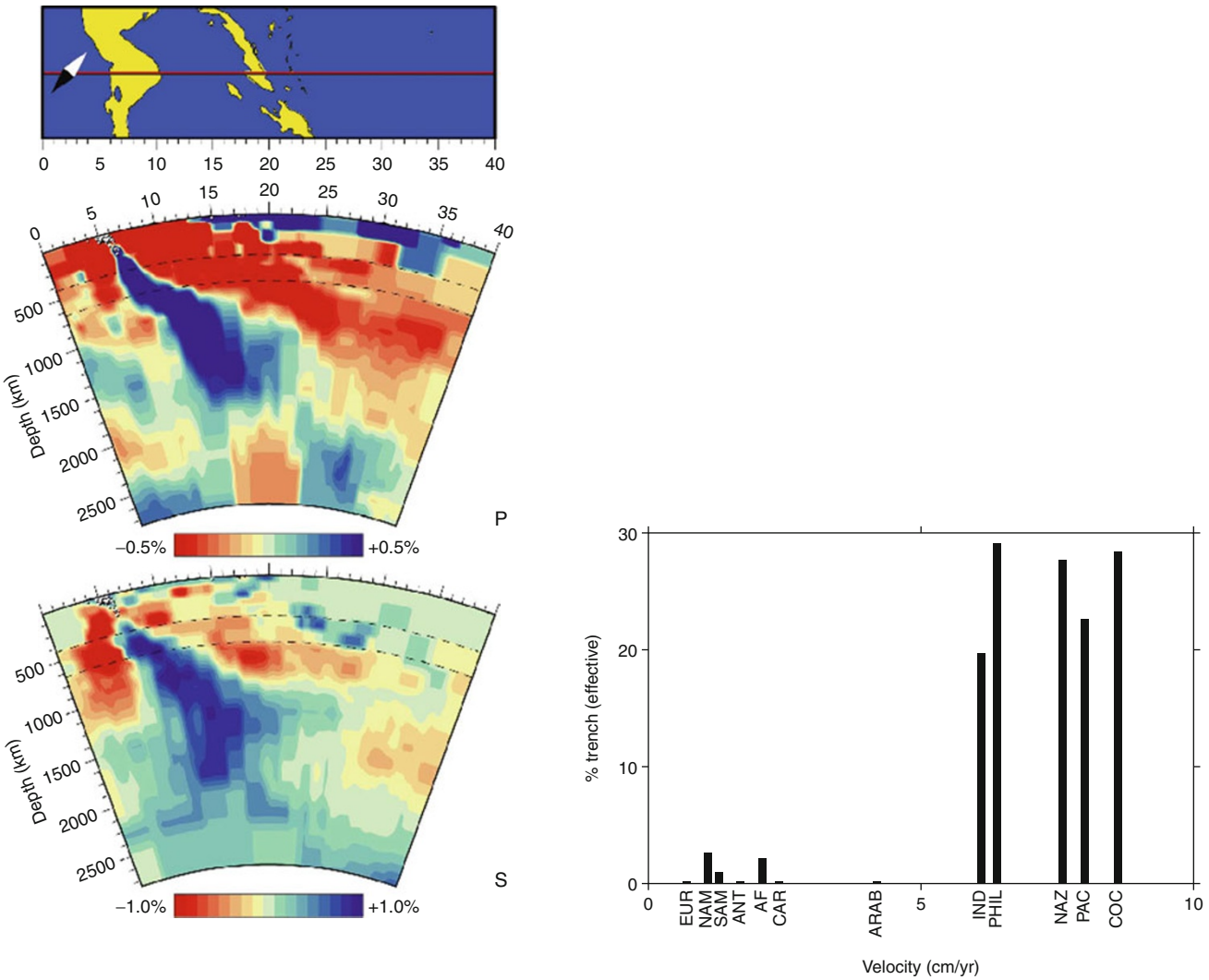
The transition zone has anomalous properties due to mixing and transitions in mineral organization; for example, it is thought to be able to absorb an order of magnitude more water (per kg) than the mantle both above and below it (although this issue is still somewhat controversial).

Other phase changes are thought to occur with depth, although these are less well resolved, and in some instances do not appear to be global. Recently, a new

phase change has been inferred in the lowest part of the mantle (the bottom few hundred kilometers), called the perovskite-post-perovskite (or just the post-perovskite) transition (Murakami et al., 2004). This transition is still an active area of exploration.

The effect of phase transitions on mantle convection has been an active area of research since the 1960s. The discovery of the major phase change at 660 km depth coincided with the fact that deep earthquakes along subducting slabs (the Wadati–Benioff Zone) go no deeper than 700 km. This seemed to imply that subducting slabs, the mantle's equivalent of cold convective downwellings (see below), did not extend into the lower mantle. Recent seismological studies using tomographic techniques to resolve ostensibly "cold" and "hot" areas of the mantle (really seismically fast and slow regions), implied that many slabs might stall or pool temporarily at the 660 km boundary but many penetrate into the lower mantle (Figure 6).

Whether the density jump due to a phase change could impede vertical flow has been a key question in studying the interaction of phase changes and convection. In particular, the 660 km phase change was inferred mineralogically to be "endothermic" whereby the entropy of the deeper heavier phase (below 660 km) increases, or more simply latent heat is absorbed on going down through the boundary into the denser more compact phase (while unusual in most systems, this is also true of the solid–liquid transition in water). This also means that the dependence of the transition temperature on pressure has a negative slope; thus cold material impinging on the phase boundary causes the phase change to deflect in the cold region to greater pressures; this induces



Mantle Convection, Figure 6 Tomographic image of slabs beneath Mexico (W. Spakman via van Keken et al., 2002) extending into the lower mantle (*left*); and plate velocity versus trench length following Forsyth and Uyeda (1975) showing the fastest plates are connected to slabs (*right*). Slabs are not only cold mantle downwellings but effectively drive plate tectonics.

a depression in the boundary that acts to buoyantly rebound upward and oppose the motion of the descending cold material. Thus the endothermic boundary at 660 km depth possibly impedes flow across that boundary. (However, the actual 660 km phase is complex and at higher temperatures might become exothermic; Weidner and Wang (1998).) Computational studies and simulations of mantle flow across this boundary demonstrated that downwellings (i.e., slabs) can indeed be impeded and stalled as they impinge on this boundary, but not permanently or globally; that is, while some are pooling at the boundary others have gathered up enough "weight" to push through the boundary (Christensen, 1995; Tackley et al., 1993). This picture appears to be in keeping with the picture from seismology that while the phase boundary impedes slab and downwelling flow into the mantle, it is not an impermeable boundary and there is in the end significant exchange between the upper and lower mantle and hence whole mantle convection.

However, while the mineralogical, seismological, and geodynamical (fluid mechanical) arguments imply that there is flow between upper and lower mantles, there are numerous data from geochemical analysis of basaltic lavas implying that the mantle is not well stirred on a large scale, that is, it is possibly layered with poor or nonexistent communication between upper and lower mantle.

Structure of mantle convection and mantle mixing

Upwelling mantle reaching the Earth's surface undergoes melting (see above under pressure-release melting) and this melt reaches the surface in two types of volcanic settings: mid-ocean ridges where tectonic plates spread apart and draw mantle up into the opening gap, and ocean-islands or hotspots which are anomalously productive and localized volcanic features not necessarily associated with tectonic activity, Hawaii being the most conspicuous such feature. Melts coming from the mantle in this way are silica poor (relative to more silicic rocks such as granite) and largely basaltic; hence these volcanic regions are said to produce mid-ocean ridge basalts (MORB) and ocean-island basalts (OIB), respectively. These melts are in effect messengers from the mantle, and their petrological composition, bulk chemistry, and trace-element chemistry are extensive areas of research (Hofmann, 1997, 2003; Tackley 2007; van Keken et al., 2002). In the end, while these two basalts nominally come from the same mantle, they have distinct features suggesting they come from parts of the mantle that have been isolated from each other for billions of years. That seismology, mineral physics, and fluid dynamics (geodynamics) argue for a largely well-stirred mantle with whole-layer circulation creates a dichotomy between geophysical and geochemical observations. This paradox has been one of the most fervent areas of debate in mantle dynamics for the last 30–40 years.

Because MORB and OIB are both basalts and thus have similar bulk chemistry, geochemical measurements largely focus on trace elements, in particular incompatible elements, which dissolve more readily in a rock's melt phase than its solid phase; hence during partial melting incompatible elements partition toward the melt. Indeed, the trace-element signature of elements such as uranium, thorium, helium, have demonstrated that MORB and OIB are very distinct. In particular, MORBs appear to be significantly depleted in such trace elements relative to OIB. Since such elements tend to be removed by melting, it implies that MORBs come from a region of the mantle that has already been melted and depleted of trace elements, while OIB come from a region of the mantle that has undergone little previous melting and depletion. This observation implies that MORBs come from an upper mantle that has been cycled repeatedly through the plate tectonic process of mid-ocean ridge melting and separation of crust (and trace elements) from mantle, in essence cleaning the MORB source. In contrast, OIB would appear to come from a part of the mantle that has seen little of this melt processing, and hence would be isolated presumably at depth from the upper mantle and plate-tectonic circulation.

There are other geochemical observations that argue for separated and isolated regions or reservoirs in the mantle. For example the concentration of radioactive daughter isotopes (e.g., ^{206}Pb, which is the final product of the decay of ^{238}U) relative to the abundance of that element's dominant and primordial isotope (e.g., ^{204}Pb) is a metric for reservoir isolation from surface processing. In particular, the relative accumulation of daughter products implies that the rock in which they reside has seen little processing or partial melting that would have cleaned out these elements after they formed. A small relative abundance of daughter isotopes means the sample has been recently processed and cleaned, and thus little time has passed in which to produce new daughter isotopes.

Indeed, many OIBs tend to show distinctly greater relative abundance of daughter products (e.g., the concentration ratio ^{206}Pb/^{204}Pb) than do MORB for many isotopic ratios, implying some isolation of the OIB source. However, OIBs from various islands are also fairly different from each other suggesting that reservoirs isolated from the upper mantle (the presumed MORB source) are possibly also isolated from each other. Moreover, the OIB isotopic ratios have some variation, from low MORB-like values to much higher values, which indicates that there is some mixing of a "young" processed MORB-source-like mantle and a more primitive, isolated one.

These isotopic ratios are also more easily interpreted for refractory daughter products since they do not tend to escape the system (also the reason they are used to radiometrically date rocks). Volatile products, especially isotopes of noble gases such as helium and argon, require different interpretations since they can readily escape the mantle and, for helium, escape the Earth. For example, MORB in fact has a high daughter to primordial isotope ratio, ^{4}He/^{3}He, relative to many OIBs, which is opposite to the refractory ratios involving, for example, lead isotopes. This is often interpreted as resulting from

degassing and loss of primordial helium, ^{3}He, from the upper mantle through plate tectonic and mid-ocean ridge processing, and the subsequent repopulating of helium with its radiogenic isotope ^{4}He (i.e., α-particles from most large element decay sequences); in contrast an isolated lower mantle or OIB source reservoir would have undergone little loss of primordial helium thus maintaining a smaller isotopic ratio ^{4}He/^{3}He.

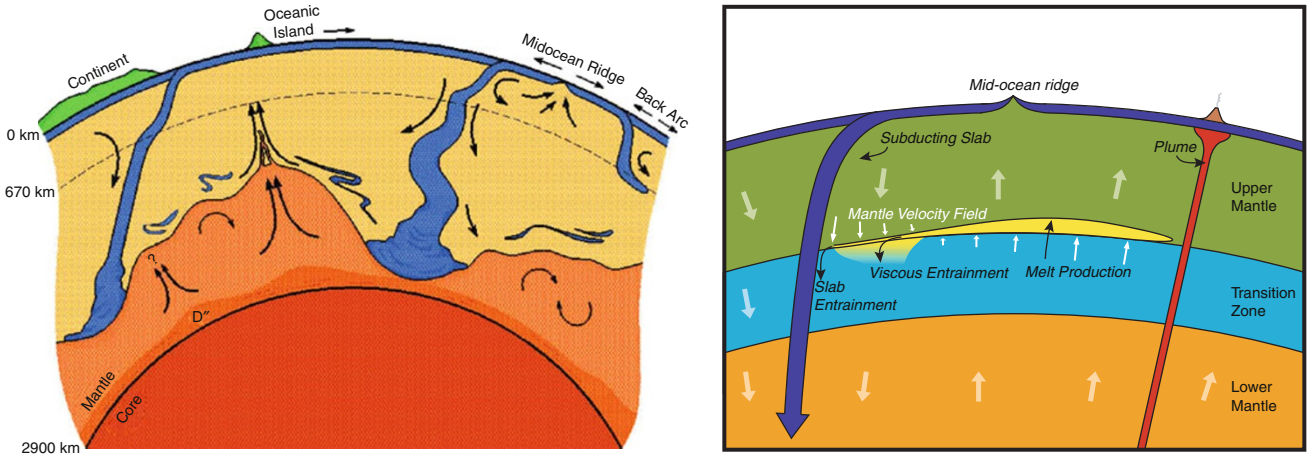
The production of the argon isotope ^{40}Ar from the decay of the potassium isotope ^{40}K has two important arguments relative to mantle layering. First, the total amount of original ^{40}K in the Earth can be roughly estimated from chondritic abundances (and other arguments beyond the scope of this entry). However, the amount of ^{40}Ar it should have produced over the age of the Earth is far in excess (by a factor of 2) of the ^{40}Ar in the atmosphere, implying that much of this argon is still buried and isolated. Moreover, one can also estimate from the trace element composition of MORBs themselves that the MORB source region as it stands now would have been lacking primordial ^{40}K and even if the entire mantle were composed of this MORB source region, it would not have been able to produce even the atmospheric levels of ^{40}Ar; this implies that the bulk of original ^{40}K was buried in a layer different and more enriched than the MORB-source region, which then produced most of the ^{40}Ar, much of which is still buried in this layer.

An often quoted straw-man argument for mantle layering is the heatflow paradox. In this case, it is reasoned that if the entire mantle were composed of MORB source material with its depleted concentration of heat-producing elements (U, Th, K), then it would not be able to produce the total heat output through the top of the mantle (about 30TW). This suggests that the heat-producing elements allowing for the mantle heat output must be buried at depth. However, this makes the assumption that most heat output is from radiogenic heating, whereas possibly less than half of it is; if most heat output is from secular cooling (lost of primordial heat) then the heat-flow paradox argument is questionable. A similar but shakier argument is based on the fact that the volcanic flux of the helium isotope ^{4}He, which is produced from heat producing radioactive decay of U and Th, seems to be very low relative to what would be expected given the heat that is emanating from the mantle, implying that ^{4}He is also buried at depth. Again, if heat output is more than 50% from secular cooling and not all from radiogenic heating, then the low flux of helium is to be expected. Even if that were not the case, mechanisms for the flux of helium are not the same as for flux of heat; that is, heat always escapes to space eventually by convection, conduction even radiation, whereas helium only escapes from the mantle if it passes through narrow melting zones, which is not inevitable, and thus it can be hidden and buried almost anywhere and not only hidden at great depth.

Finally, the production of continental crust also argues for an isolated layer in the mantle. Continental crust represents an accumulated history of mantle melting, segregation of lighter components, and removal of incompatible elements to the surface. If the continental crust were removed uniformly from a whole mantle made of primordial "bulk silicate earth" (i.e., an earth derived from chondritic material and only segregated into mantle and core), then the concentration of incompatible elements would not have been reduced enough to produce a mantle made of MORB source material (i.e., removal of continental crust would not have depleted the whole mantle enough to make MORB source). However, if the crust were removed from 1/3 to 1/2 of the mantle, that resulting depleted portion would very closely match MORB source composition. This argues that the continental crust segregated only from the upper portion of the mantle, not the whole mantle, and thus there remains a deeper unsegregated mantle at depth (see van Keken et al., 2002).

Although there are numerous geochemical arguments for a layered mantle with an isolated and undepleted mantle at depth, they largely conflict with geophysical evidence for whole mantle convection. Mineral physics experiments suggest that the 660 km phase change boundary might provide an impediment to mantle flow but not an impermeable barrier. Seismic tomography consistently shows subducting slabs and apparently cold downwellings extending into the lower mantle (Grand et al., 1997; van der Hilst et al., 1997). Recent high-resolution images of a hot upwelling mantle plume beneath Hawaii (Wolfe et al., 2009) as well as seismic images of other plumes (Montelli et al., 2004) also suggest vertical upward transport across the 660 km boundary. Finally, geodynamical arguments against separated layers suggest that if a lower mantle held most of the mantle's heat-producing elements, it would be implausibly hot, and by heating up the bottom of the upper mantle it would generate much bigger mantle plumes than would be observed (Tackley, 2002).

The contradiction between geochemical and geophysical inference of layered versus whole mantle convection has been and largely remains an unsolved problem. Attempts to reconcile these observations have been numerous. A reasonably popular approach has been to allow that the 660 km boundary is not a barrier to mantle flow, but that the barrier exists at greater depth. There are seismically observable layers at the bottom of the mantle (the D″ layer), which could store enriched material, although these are also so thin they could possibly overheat (depending on the amount of radioactive heat sources stored there). As a compromise, it has been argued that the enriched mantle exists in an approximately 1,000-km-thick layer at the base of the mantle (Kellogg et al., 1999), although this layer has never been seismologically observed (see Figure 7). More recently, chemical heterogeneity gathered into piles on the core–mantle boundary and below upwelling zones (Jellinek and Manga, 2002, 2004) has been suggested by convection models (McNamara and Zhong, 2005) with support from joint seismology-gravity analyses (e.g., Ishii and Tromp, 1999).



Mantle Convection, Figure 7 Two end-member mantle mixing models. In the abyssal layered model (*left*) the source for enriched ocean-island basalt (OIB) is in a deep primordial layer, while the source for depleted mid-ocean ridge basalt (MORB) is in the upper recycled mantle (After Kellogg et al. (1999)). In differential melting models, such as the water-filter model (*right*), the MORB and OIB sources undergo different styles of melting, but the mantle still undergoes whole-mantle circulation; see text for further explanation (After Bercovici and Karato (2003a); Leahy et al. (2007)).

Other mechanisms for reconciling geochemical and geophysical observations, but not invoking layering, have been recently proposed as well, mostly involving differential melting. For example, one model considers the whole mantle as a plum-pudding mix of enriched and volatile (water)-rich plums in a depleted, drier, and harder-to-melt pudding. The pressure-release melting in mantle plumes is stopped at higher pressures at the base of 100 km thick lithosphere, so this could involve mostly melting of easily melted, enriched, and volatile-rich components; melts making it all the way to the surface at ridges would undergo more pressure drop and thus could also melt the depleted mantle component, resulting in MORB that seems depleted relative to OIB (Ito and Mahoney 2005a, b). Another model exploits the fact that transition zone minerals seem to be able to absorb water more readily than material above it (and below it). A transition zone with a little water will be dry relative to its solubility or water storage capacity. However, upwelling material passing through the transition zone would carry this slightly damp material into the upper mantle at the 410 km boundary, and since the upper mantle olivine has poor water solubility, it would be closer to saturation and likely melt. A little melting at 410 km depth of the broad upwelling mantle (forced upward by the downward flux of slabs) passing through and out of the transition zone would cause it to be stripped of incompatible elements as it flows into the upper mantle leaving a depleted MORB source region; because of the high pressures and high compressibility of melt, the partial melt that has cleaned this upwelling mantle would be dense and remain behind, eventually to be entrained by slab-driven downwelling back into the lower mantle. Upwelling mantle plumes on the other hand would go through the transition zone too fast to become hydrated and thus would undergo little melting and filtering at 410 km depths, leaving largely enriched OIB. This model, called the transition zone Water Filter (Bercovici and Karato, 2003a; Karato et al., 2006; Leahy and Bercovici, 2010) (see Figure 7), predicts that the 410 km should be the site of melting, and this has been borne out in various seismological studies (e.g., Revenaugh and Sipkin, 1994; Song et al., 2004; Tauzin et al., 2010); however, the theory is still controversial given poor knowledge of melt properties and their solubilities of incompatible elements at these depths and pressures, so it remains an active subject of investigation.

Mantle convection and the generation of plate tectonics

The oldest problem in mantle convection

The link between plate tectonics and mantle convection is one of the oldest and most challenging problems in the history of geodynamics. The original theories of mantle convection put forward by Holmes (1931) were developed in the context of explaining continental drift as articulated by Wegener (1924). Although the later theory of plate tectonics is considered the grand-unifying principle of geology, it is a kinematic theory in that it describes surface motions but not their cause. Mantle convection is widely accepted to be the engine for plate motions since it is a fundamental mechanism for exploiting the energy sources of the Earth's interior, that is, loss of primordial and radiogenic heat. It is now generally regarded that the plates themselves are a feature of mantle convection in that they are the mobile upper thermal boundary layer of convective cells that thicken as they cool in their migration away from ridges until they are heavy enough to sink along subduction zones.

One of the major accomplishments of mantle dynamics theory is that convective fluid velocities, calculated in any

number of ways, consistently predict the measured scales of plate tectonic velocities, that is, between 1 and 10 cm/year. This was in fact inferred even in the 1930s from both gravity and heat-flow measurements (Hales, 1936; Pekeris, 1935) and is well known by the force balance on sinking slabs (e.g., Davies and Richards, 1992), as well global heat extraction from mid-mantle cooling by slabs (Bercovici, 2003). Moreover, plate motions are well correlated with the presence of slab forcing, in particular that tectonic plates with a significant portion of subduction all have velocities roughly an order of magnitude faster than plates without substantial subduction zone (Forsyth and Uyeda, 1975); see Figure 6. Thus, because cold sinking slabs seem to be the major expression of convection and major drivers of plate motions implies that the plates are convection.

The plate generation problem

While convective and slab-driving forces for plate tectonics are important, understanding how plates self-consistently arise (or why they do or do not arise) from planetary convection has been a major goal in geodynamics. Up until the early 1990s it was believed that since plate-like motion of the lithosphere was essentially discontinuous, it could not be predicted or reproduced by fluid dynamical convection theories. However, in the last 15 years or so there has been major progress with fluid dynamical models yielding very plate-like motion by incorporating more sophisticated rheological weakening mechanisms, such as brittle/plastic yielding or damage mechanics.

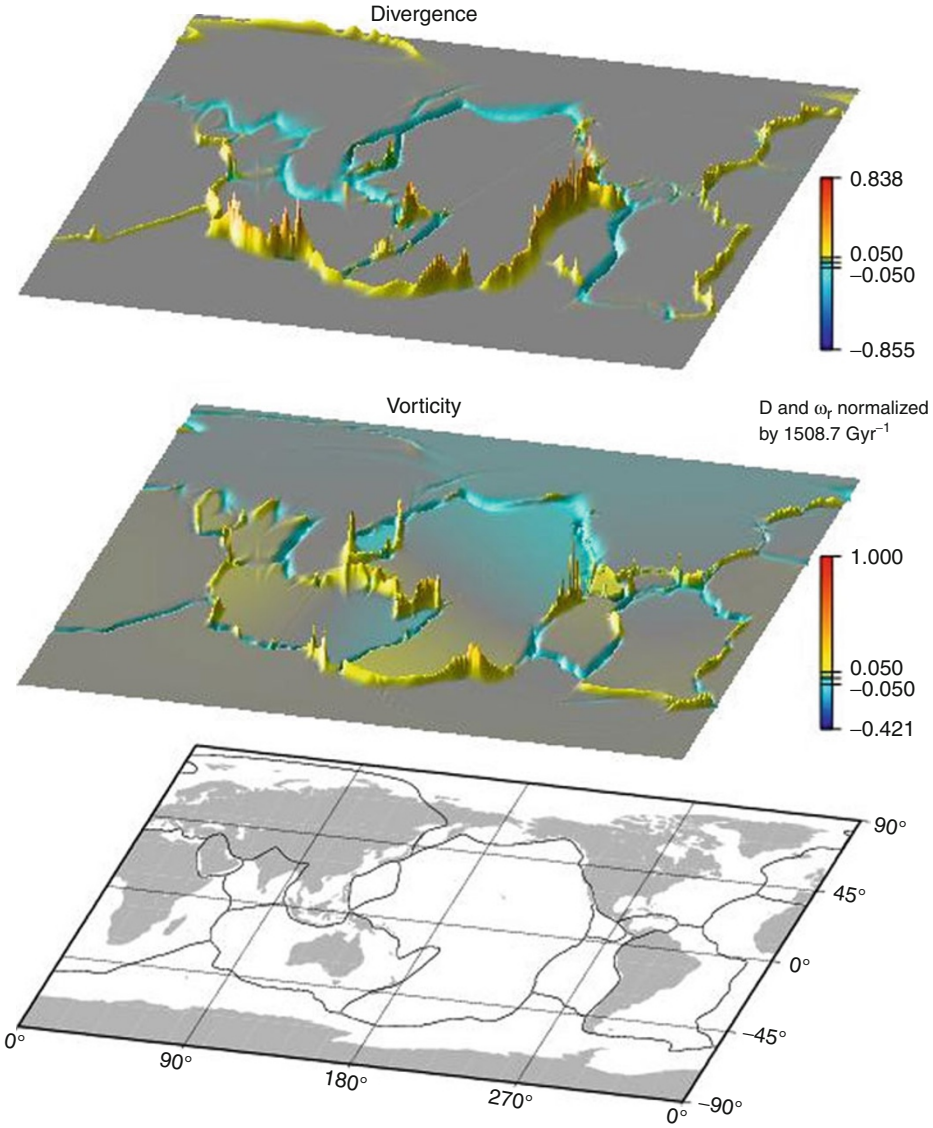
However, even so, there is still no comprehensive theory of how the plates and mantle are related, and in particular how plate tectonics self-consistently arises from a convecting mantle. Both a clue and frustration is that the Earth appears to be the only known terrestrial planet that has plate tectonics in addition to liquid water as well as life, which are all either causative (i.e., necessary conditions for each other) or coincidental. While our planet supports plate tectonics, our ostensible twin, Venus, does not; this remains a leading-order quandary in Earth sciences and to solve it one must understand how and why plate tectonics is generated at all. Of course, geoscience has until recently only sampled the few planets of our own solar system and thus the data is sparse; with the advent of extra-solar planet discovery, this sparseness should be mitigated and perhaps other planets will be discovered with plate-like mantle circulation from which we will learn more about how our own planet works (e.g., Valencia et al., 2007).

There are of course various qualitative and philosophical questions about how and why plate tectonics forms, evolves, and exists, but for these to be predicted or reproduced in a physical theory, one requires quantifiable questions. In short, what are the metrics of plate generation? Two fundamental features of plate-like motion exist for instantaneous motions: *plateness* (Weinstein and Olson, 1992) and toroidal motion (e.g., Bercovici et al., 2000; Hager and O'Connell, 1979; O'Connell et al., 1991). Plateness is essentially the extent to which surface motion and the strength of the lithosphere is like that for nearly rigid blocks separated by narrow weak boundaries. Toroidal flow is characterized by strike-slip motion and plate spin (Figure 8). Toroidal motion has no direct driving force in buoyant convection (which drives only vertical and divergent motion); however, it has as much energy in the present-day plate tectonic velocity field as the buoyantly driven motion (called poloidal motion). The globally-averaged toroidal motion is dependent on the lithosphere's reference frame (e.g., hostpot frame), and the field in general changes through time (Cadek and Ricard, 1992; Lithgow-Bertelloni et al., 1993); however, the toroidal field is a quantifiable and significant feature of global plate motions. Such measurable quantities as plateness and toroidal flow are important for testing the predictions of plate generation theories. Both phenomena rely on reasonably strong nonlinear rheological feedback effects to permit large strength variations for high plateness (i.e., rapidly deforming zones are weak, slowly deforming ones are strong), as well as coupling of buoyantly driven flow to vertical torques that drive toroidal spin and shear (Bercovici, 2003).

Recent progress

In the last decade, plate generation models have become increasingly sophisticated, in concert with further expansion and accessibility of high-performance computing. Instantaneous plate-like behavior has been achieved with convection models employing various forms of plastic yield and self-weakening criteria. Incorporation of these laws has led to the prediction of reasonable toroidal and poloidal flow (Bercovici, 1995) (Figure 9) and by including the rheological effects of melting at ridges have attained localized passive spreading zones (Tackley, 2000) (Figure 10). Most recently, these models have been extended to 3-D spherical models, creating the first global models of plate generation from mantle convection (Foley and Becker, 2009; van Heck and Tackley, 2008) (Figure 11).

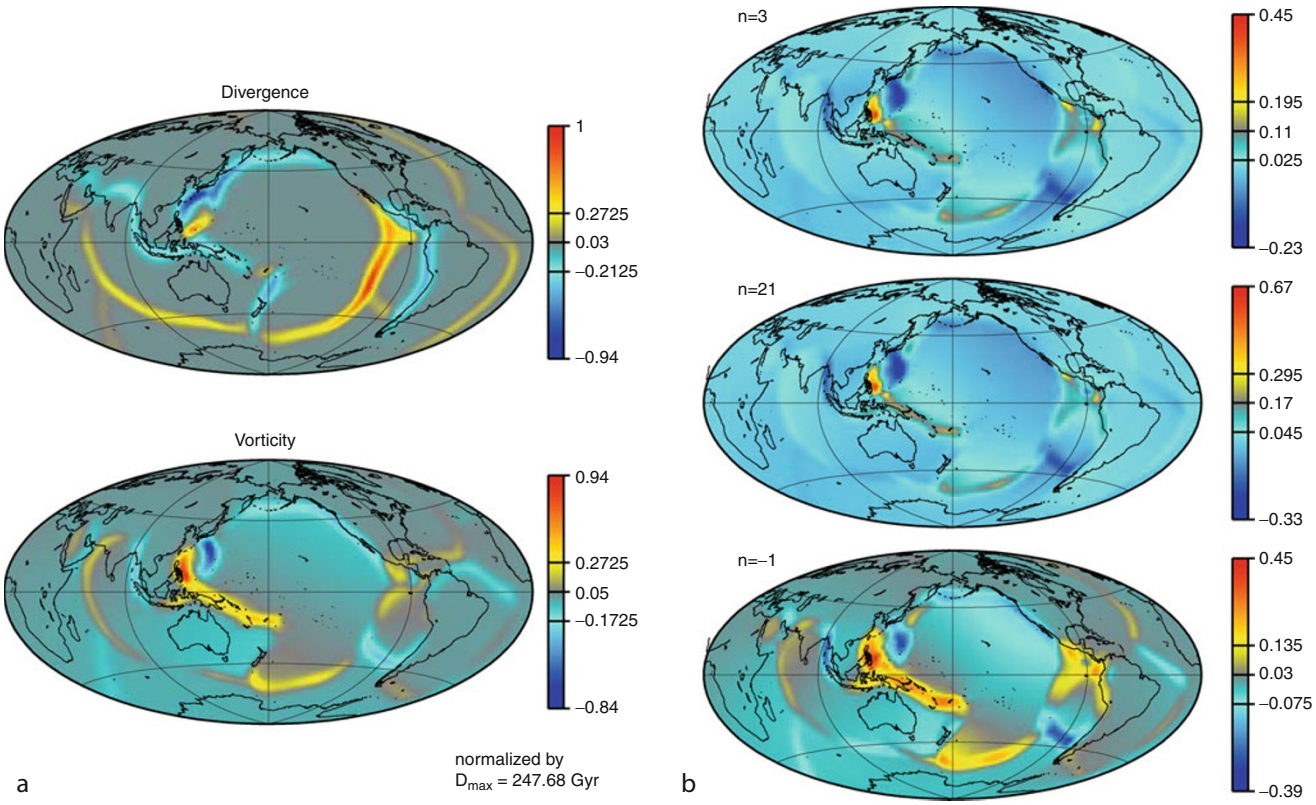
However, models that use plastic or instantaneous self-weakening rheologies only allow weak zones to exist while being deformed, thus cannot correctly model dormant weak zones (e.g., sutures and inactive fracture zones). Rheological mechanisms that allow weakening to persist over time have also been studied. While thermal weakening is a well-understood mechanism, thermal anomalies diffuse too fast, and it is highly unlikely that, for example, sutures are thermal remnants. Weakening by hydration or as a secondary phase (i.e., by providing pore pressure) is a strong candidate as well, although the mechanism for ingesting water to depth is problematic, and requires mechanisms such as cracking enhanced by thermal stresses (Korenaga, 2007). Damage in the form of microcracks (Bercovici, 1998) and grain-size reduction



Mantle Convection, Figure 8 The horizontal divergence (which measures spreading and convergence rates) and vertical vorticity (which measures angular velocity and horizontal shear) of present-day plate motions (from Dumoulin et al. (1998)), which are associated with poloidal and toroidal flow, respectively. Poloidal flow is equivalent to basic convective motion driven directly by thermal buoyancy. Although present in Earth's plate-mantle system, toroidal flow does not arise naturally in basic viscous convection, but requires the coupling of convective motion with nonlinear rheological effects.

(e.g., Braun et al., 1999; Karato et al., 1980; Montési and Hirth, 2003) are also strong candidate-weakening mechanisms because of their longevity and evidence in the form of mylonites (Jin et al., 1998). The need for shear-localization at significant depth makes grain-size weakening particularly appealing and has proven to be successful at creating plate-like mantle flows (Bercovici and Ricard, 2005; Landuyt and Bercovici, 2009b; Landuyt et al., 2008) (Figure 12). However, grain-size reduction and weakening tend to occur in exclusive areas of deformation space (i.e., weakening occurs during diffusion creep while reduction occurs in dislocation creep; see De Bresser et al., 2001) and require mixing of mechanisms in physical or grain-size distribution space (Bercovici and Karato, 2003b; Ricard and Bercovici, 2009).

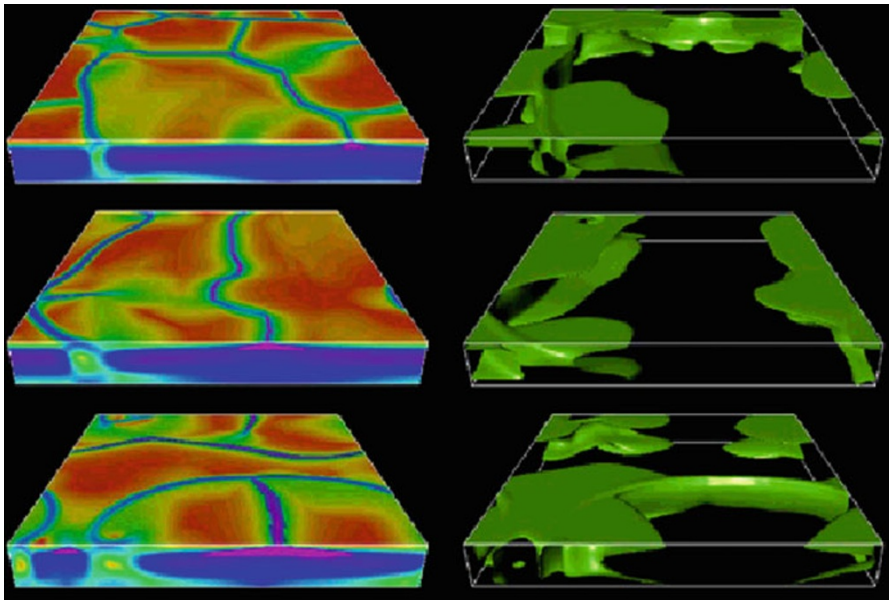
Use of both plastic/brittle yielding and damage theories of plate generation have been used to elucidate the planetary dichotomy between Earth and Venus and the causal link between climate, liquid water, and plate generation. Earth and Venus are ostensible twins but Earth has plate tectonics and Venus does not. This is usually attributed to lack of water on Venus which would otherwise lubricate



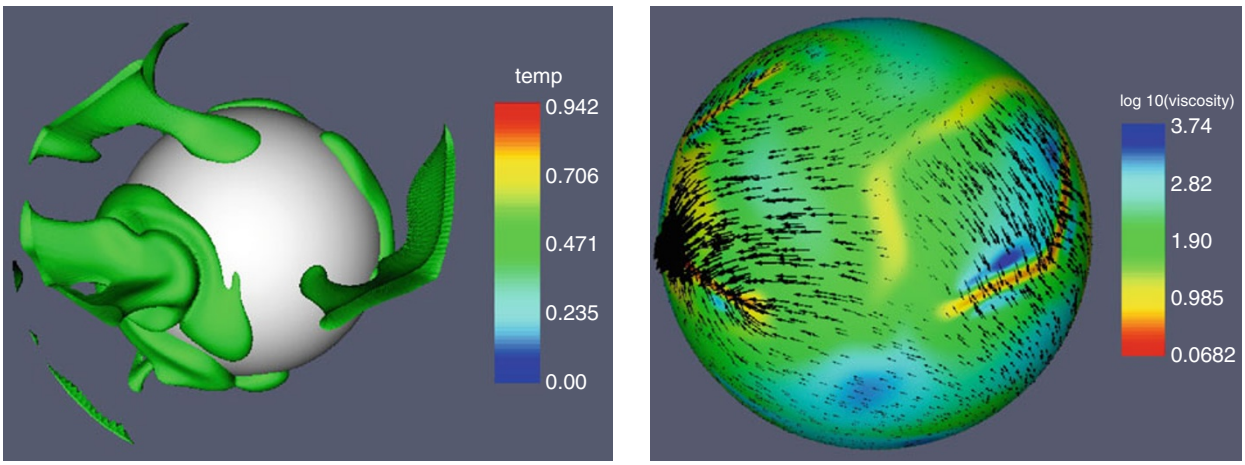
Mantle Convection, Figure 9 The source-sink model of lithospheric motion uses present-day plate motions to drive flow with the divergent field (*upper left*) in order to model and recover the known vorticity field (*lower left*) using non-Newtonian flow calculations (*right column*). Simple power-law or pseudo-plastic rheologies (typical of dislocation creep) do not recover the vorticity field well (*top* two, *upper right*), while velocity-weakening or shear-localizing rheologies (where stress decreases with increased strain rate) recover it very well (*lower right*) (From Bercovici et al. (2000) after Bercovici (1995). American Geophysical Union).

plate motions; however, Earth's lithosphere is likely to be no more hydrated than Venus's because of melting dehydration at ridges (Hirth and Kohlstedt, 1996). Recent studies have hypothesized that the role of water in maintaining plate motion is not to lubricate plates, but to be an agent for the carbon cycle, which thus allows for a temperate climate on Earth. A cool surface on Earth, according to one hypothesis (Lenardic et al., 2008) causes a larger temperature drop across the lithosphere than would occur on Venus (because of temperature-dependent viscosity), and thus lithospheric buoyant stresses are large enough on Earth to induce plastic/brittle failure, but not on Venus. An alternative hypothesis (Landuyt and Bercovici, 2009a) states that plate-like motion depends on the competition between damage and healing (where, e.g., if damage is due to grain-size reduction, healing is due to normal grain growth), where a high damage-to-healing ratio promotes plate-like motion, while a lower ratio yields stagnant-lid behavior. A cooler surface temperature inhibits healing while a hot surface promotes healing, thus leading to plate-like behavior on a planet like Earth with a temperate climate but not on a planet like Venus. Both hypotheses emphasize that water dictates conditions for plate generation by its modulation of climate and not on direct strength reduction.

Finally, subduction initiation continues to be an extremely challenging issue in geodynamics (King, 2007; Stern, 2004). The strength of thick, cold pre-subduction lithosphere is such that it should never go unstable and sink, at least not on geological timescales (or cosmological ones either). Thus, how and why subduction zones form remains enigmatic. Mechanisms range from weakening by rifting (Kemp and Stevenson, 1996; Schubert and Zhang, 1997), sediment loading and water injection (Regenauer-Lieb et al., 2001), and reactivation of preexisting fault-zones (Hall et al., 2003; Toth and Gurnis, 1998), all of which have some observational motivation, although fault reactivation might be the most compelling (e.g., Lebrun et al., 2003). Another unresolved enigma concerns the age of subduction zones. The convective picture of plate motions would have plates subduct when they get old, cold, and heavy. However, the sea-floor age distribution implies that subduction rate is independent of plate age, such that the age of plates at subduction



Mantle Convection, Figure 10 A simulation of plate generation over mantle convection. The plate rheology is visco-plastic and the viscosity reduction associated with melting is parameterized into the model, leading to exceptional plate-like behavior and apparent passive spreading (i.e., narrow spreading centers not associated with any deep upwelling). The *right panels* show surfaces of constant temperature, which here are dominated by cold downwellings; the *left panels* show the viscosity field (red being high viscosity and blue low viscosity). Different rows show different times in the simulation (After Tackley (2000). American Geophysical Union).

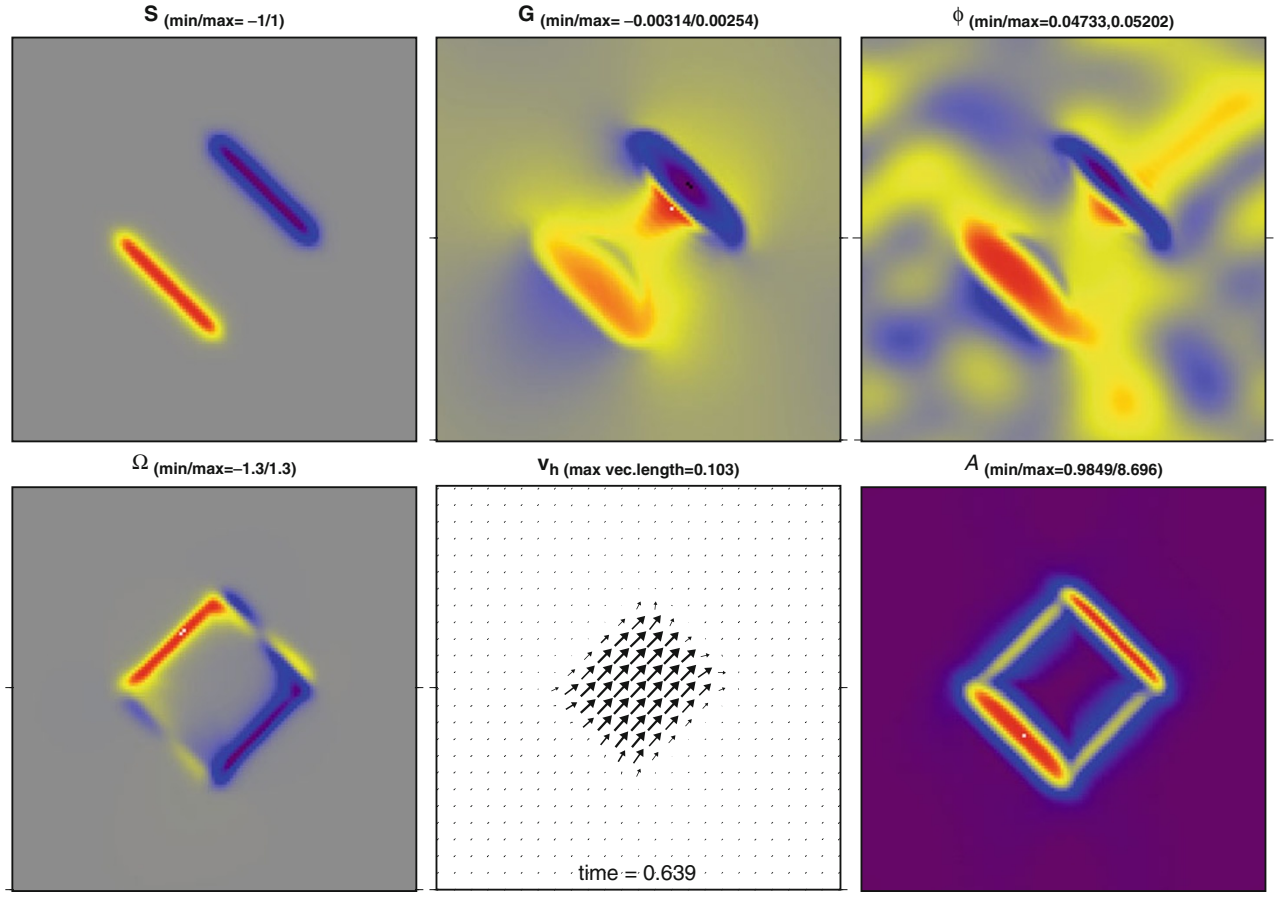


Mantle Convection, Figure 11 Three-dimensional spherical shell convection with a plastic-type lithospheric rheology showing plate-like behavior. *Left panel* shows isothermal surfaces and in particular cold downwellings. The *right panel* shows surface viscosity with velocity vectors superimposed. Note the passive divergent and rheological weak zone forming midway between the two major downwelling regions (Adapted from Foley and Becker (2009). American Geophysical Union).

zones is distributed from age nearly 0 (i.e., subducting ridges), to the oldest ages of roughly 200 Myr (see Becker et al., 2009).

Summary

Mantle convection is the central theory for how the Earth works, that is, what drives geological motions as well as the cooling history and chemical evolution of the planet. The theory is based on the simple notion that the Earth (and other terrestrial planets) must cool to space and in so doing release heat and gravitational potential energy through thermal convection. Thermal convection theory itself is a well-established physical theory rooted in classical mechanics and fluid dynamics. Much of the physics of



Mantle Convection, Figure 12 A simple source-sink model of shallow flow with a two-phase and grain-size-reducing damage mechanism. Damage per se involves transfer of deformational work to the creation of surface energy on interfaces by void and/or grain boundary generation in the continuum. In the case shown, all damage is focused on grainsize reduction. The panel meanings are indicated by symbols where S is the imposed divergence rate (i.e., the source-sink field) that drives flow; G is the dilation rate due to void formation; φ is void volume fraction; Ω is vertical vorticity or rate of strike-slip shear; $\mathbf{v_h}$ is horizontal velocity; and A is the "fineness" or inverse grainsize. This particular calculation shows that fineness-generating, or grain-size-reducing, damage is very effective at creating localized fault-like strike-slip zones in vorticity Ω, and solid-body-like translation in the velocity field $\mathbf{v_h}$ (Adapted from Bercovici et al. (2005). American Geophysical Union).

basic convection goes far in describing circulation and structure in the Earth's mantle, for example, the flow velocities, establishment of thermal boundary layers, and even the prevalence of slab-like downwellings and plume-like upwellings. However, numerous quandaries and paradoxes persist because much remains to be understood regarding the many complexities of the exotic convecting "fluid" in the mantle. How the Earth appears to be stirred by deep subducting slabs, but still appear to be unmixed when it comes up at either mid-ocean ridges or ocean-islands, remains a major mystery, much of which is due to our incomplete understanding of the processes of melting, chemical segregation, and mixing in the mantle. How and why the Earth's mantle convects in the form of plate tectonics at all and unlike other terrestrial planets remains one of the biggest questions in geoscience; this problem is undoubtedly related to the mantle's exotic rheology in which flow depends on multiple properties including temperature, stress, chemistry, and mineral grain size. Even more than 100 years since Kelvin's controversial dating of the Earth's age by cooling, in fact the cooling history of the Earth remains poorly understood because of incomplete knowledge of radiogenic heat sources as well as the complex physics of mantle flow. Thus, while mantle convection remains one of the grand unifying physical theories of how the Earth works, many of the major questions and mysteries about the mantle remain unsolved and are thus ripe for discovery by future generations of Earth scientists.

Bibliography

Becker, T., Conrad, C., Buffett, B., and Müller, R., 2009. Past and present seafloor age distributions and the temporal evolution of plate tectonic heat transport. *Earth and Planetary Science Letters*, **278**, 233–242.

Bénard, H., 1900. Les tourbillons cellulaires dans une nappe liquide, *Revue générale des Sciences pures et appliquées*, **11**, 1261–1271 and 1309–1328.

Bénard, H., 1901. Les tourbillons cellulaires dans une nappe liquide transportant de la chaleur par convection en régime permanent. *Annales de Chimie et de Physique*, **23**, 62–144.

Bercovici, D., 1995. A source-sink model of the generation of plate tectonics from non-Newtonian mantle flow. *Journal of Geophysical Research*, **100**, 2013–2030.

Bercovici, D., 1998. Generation of plate tectonics from lithosphere-mantle flow and void-volatile self-lubrication. *Earth and Planetary Science Letters*, **154**, 139–151.

Bercovici, D., 2003. The generation of plate tectonics from mantle convection. *Earth and Planetary Science Letters*, **205**, 107–121.

Bercovici, D., 2007. Mantle dynamics, past, present and future: An overview. In Bercovici, D., and Schubert, G. (eds.), *Treatise on Geophysics: Mantle Dynamics*. New York: Elsevier, Vol. 7, pp. 1–30.

Bercovici, D., and Karato, S., 2003a. Whole mantle convection and the transition-zone water filter. *Nature*, **425**, 39–44.

Bercovici, D., and Karato, S., 2003b. Theoretical analysis of shear localization in the lithosphere. In Karato, S., and Wenk, H. (eds.), *Reviews in Mineralogy and Geochemistry: Plastic Deformation of Minerals and Rocks.* Washington, DC: Minerological Society of America, Vol. 51, Chap. 13, pp. 387–420.

Bercovici, D., and Ricard, Y., 2005. Tectonic plate generation and two-phase damage: void growth versus grainsize reduction. *Journal of Geophysical Research*, **110**(B03401), doi:10.1029/2004JB003,181.

Bercovici, D., Ricard, Y., and Richards, M., 2000. The relation between mantle dynamics and plate tectonics: A primer. In Richards, M. A., Gordon, R., and van der Hilst, R. (eds.), *History and Dynamics of Global Plate Motions*. Washington, DC: American Geophysical Union. Geophysical Monograph Series, Vol. 121, pp. 5–46.

Braun, J., Chery, J., Poliakov, A., Mainprice, D., Vauchez, A., Tomassi, A., and Daignieres, M., 1999. A simple parameterization of strain localization in the ductile regime due to grain size reduction: A case study for olivine. *Journal of Geophysical Research*, **104**, 25167–25181.

Cadek, O., and Ricard, Y., 1992. Toroidal/poloidal energy partitioning and global lithospheric rotation during cenozoic time. *Earth and Planetary Science Letters*, **109**, 621–632.

Chandrasekhar, S., 1961. *Hydrodynamic and Hydromagnetic Stability*. New York: Oxford University Press.

Christensen, U., 1995. Effects of phase transitions on mantle convection. *Annual Review of Earth and Planetary Sciences*, **23**, 65–88.

Davies, G., and Richards, M., 1992. Mantle convection. *Journal of Geology*, **100**, 151–206.

De Bresser, J., ter Heege, J., and Spiers, C., 2001. Grain size reduction by dynamic recrystallization: can it result in major rheological weakening? *International Journal of Earth Sciences*, **90**, 28–45.

Dumoulin, C., Bercovici, D., and Wessel, P., 1998. A continuous plate-tectonic model using geophysical data to estimate plate margin widths, with a seismicity based example. *Geophysical Journal International*, **133**, 379–389.

Foley, B., and Becker, T., 2009. Generation of plate-like behavior and mantle heterogeneity from a spherical, visco-plastic convection model. *Geochemistry, Geophysics, Geosystems*, **10**, q08001, doi:10.1029/2009GC002378.

Forsyth, D., and Uyeda, S., 1975. On the relative importance of the driving forces of plate motion. *Geophysical Journal of the Royal Astronomical Society*, **43**, 163–200.

Grand, S., van der Hilst, R., and Widiyantoro, S., 1997. Global seismic tomography: A snapshot of convection in the earth. *GSA Today*, **7**, 1–7.

Hager, B., and O'Connell, R., 1979. Kinematic models of large-scale flow in the earth's mantle. *Journal of Geophysical Research*, **84**, 1031–1048.

Hales, A., 1936. Convection currents in the Earth. *Monthly Notices of the Royal Astronomical Society. Geophysical Supplement*, **3**, 372–379.

Hall, C. E., Gurnis, M., Sdrolias, M., Lavier, L. L., and Mueller, R. D., 2003. Catastrophic initiation of subduction following forced convergence across fracture zones. *Earth and Planetary Science Letters*, **212**, 15–30.

Halliday, A., 2001. In the beginning. *Nature*, **409**, 144–145.

Hess, H., 1962. History of ocean basins. In Engeln, A., James, H., and Leonard, B. (eds.), *Petrologic Studies - A Volume in Honor of A.F. Buddington*. New York: Geological Society of America, pp. 599–620.

Hirth, G., and Kohlstedt, D., 1996. Water in the oceanic upper mantle: implications for rheology, melt extraction and the evolution of the lithosphere. *Earth and Planetary Science Letters*, **144**, 93–108.

Hofmann, A., 1997. Mantle geochemistry: the message from oceanic volcanism. *Nature*, **385**, 219–228.

Hofmann, A., 2003. Sampling mantle heterogeneity through oceanic basalts: Isotopes and trace elements. In Holland, R. C. H., and Turekian, K. K. (eds.), *Treatise on Geochemistry: The Mantle and Core*. Amsterdam: Elsevier, pp. 61–101.

Holmes, A., 1931. Radioactivity and earth movements. *Transactions of the Geological Society of Glasgow*, **18**, 559–606.

Howard, L. N., 1966. Convection at high Rayleigh number. In Gortler, H. (ed.), *Proceedings of the Eleventh International Congress of Applied Mechanics*. New York: Springer, pp. 1109–1115.

Ishii, M., and Tromp, J., 1999. Normal-mode and free-air gravity constraints on lateral variations in velocity and density of earth's mantle. *Science*, **285**, 1231–1236.

Ito, G., and Mahoney, J., 2005a. Flow and melting of a heterogeneous mantle: 1. method and importance to the geochemistry of ocean island and mid-ocean ridge basalts. *Earth and Planetary Science Letters*, **230**, 29–46.

Ito, G., and Mahoney, J., 2005b. Flow and melting of a heterogeneous mantle: 2. Implications for a chemically non-layered mantle. *Earth and Planetary Science Letters*, **230**, 47–63.

Jaupart, C., Labrosse, S., and Mareschal, J.-C., 2007. Temperatures, heat and energy in the mantle of the earth. In Bercovici, D., and Schubert, G. (eds.), *Treatise on Geophysics: Mantle Dynamics*. New York: Elsevier, Vol. 7, pp. 253–303.

Jellinek, A., and Manga, M., 2002. The influence of a chemical boundary layer on the fixity and lifetime of mantle plumes. *Nature*, **418**, 760–763.

Jellinek, A., and Manga, M., 2004. Links between long-lived hotspots, mantle plumes, D" and plate tectonics. *Reviews of Geophysics*, **42**, RG3002, doi:10.1029/2003RG000,144.

Jin, D., Karato, S., and Obata, M., 1998. Mechanisms of shear localization in the continental lithosphere: Inference from the deformation microstructures of peridotites from the Ivrea zone, northwestern Italy. *Journal of Structural Geology*, **20**, 195–209.

Karato, S., 2008. *Deformation of Earth Materials: An Introduction to the Rheology of Solid Earth*. Cambridge: Cambridge University Press.

Karato, S., Toriumi, M., and Fujii, T., 1980. Dynamic recrystallization of olivine single crystals during high temperature creep. *Geophysical Research Letters*, **7**, 649–652.

Karato, S., Bercovici, D., Leahy, G., Richard, G., and Jing, Z., 2006. The transition zone water filter model for global material

circulation: where do we stand? In Jacobsen, S., and van der Lee, S. (eds.), *Earth's Deep Water Cycle, AGU Monograph Series*. Washington, DC: American Geophysical Union, Vol. 168, pp. 289–313.

Kellogg, L., Hager, B., and van der Hilst, R., 1999. Compositional stratification in the deep mantle. *Science*, **283**, 1881–1884.

Kemp, D., and Stevenson, D., 1996. A tensile flexural model for the initiation of subduction. *Geophysical Journal International*, **125**, 73–94.

King, S., 2007. Mantle downwellings and the fate of subducting slabs: Constraints from seismology, geoid, topography, geochemistry, and petrology. In Bercovici, D., and Schubert, G. (eds.), *Treatise on Geophysics: Mantle Dynamics*. New York: Elsevier, Vol. 7, pp. 325–370.

Korenaga, J., 2007. Thermal cracking and the deep hydration of oceanic lithosphere: A key to the generation of plate tectonics? *Journal of Geophysical Research*, **112**, 1–20, doi:10.1029/2006JB004502.

Korenaga, J., 2008. Urey ratio and the structure and evolution of earth's mantle. *Reviews of Geophysics*, **46**, rG2007, doi:10.1029/2007RG000241.

Lamb, S., and Sington, D., 1998. *Earth Story*. Princeton: Princeton University Press.

Landuyt, W., and Bercovici, D., 2009a. Variations in planetary convective via the effect of climate on damage. *Earth and Planetary Science Letters*, **277**, 29–37.

Landuyt, W., and Bercovici, D., 2009b. Formation and structure of lithospheric shear zones with damage. *Physics of the Earth and Planetary Interiors*, doi:10.1016/j.pepi.2009.03.005.

Landuyt, W., Bercovici, D., and Ricard, Y., 2008. Plate generation and two-phase damage theory in a model of mantle convection. *Geophysical Journal International*, **174**, 1065–1080.

Leahy, G., and Bercovici, D., 2007. On the dynamics of a hydrous melt layer above the transition zone, *Journal of Geophysical Research*, **112**, doi:10.1029/2006JB004631.

Leahy, G., and Bercovici, D., 2010. Reactive infiltration of hydrous melt above the mantle transition zone. *Journal of Geophysical Research*, **115**, b08406, doi:10.1029/2009JB006757.

Lebrun, J.-F., Lamarche, G., and Collot, J.-Y., 2003. Subduction initiation at a strike-slip plate boundary: The cenozoic pacific-Australian plate boundary, south of New Zealand. *Journal of Geophysical Research*, **108**, 2453, doi:10.1029/2002JB002041.

Lenardic, A., Jellinek, M., and Moresi, L.-N., 2008. A climate change induced transition in the tectonic style of a terrestrial planet. *Earth and Planetary Science Letters*, **271**, 3442.

Lithgow-Bertelloni, C., Richards, M., Ricard, Y., O'Connell, R., and Engebretson, D., 1993. Toroidal-poloidal partitioning of plate motions since 120 Ma. *Geophysical Research Letters*, **20**, 375–378.

McNamara, A., and Zhong, S., 2005. Thermochemical structures beneath Africa and the pacific ocean. *Nature*, **437**, 1136–1139.

Montelli, R., Nolet, G., Dahlen, F. A., Masters, G., Engdahl, E. R., and Hung, S.-H., 2004. Finite-frequency tomography reveals a variety of plumes in the mantle. *Science*, **303**, 338–343.

Montési, L., and Hirth, G., 2003. Grain size evolution and the rheology of ductile shear zones: From laboratory experiments to postseismic creep. *Earth and Planetary Science Letters*, **211**, 97–110.

Morgan, W., 1971. Convection plumes in the lower mantle. *Nature*, **230**, 42–43.

Murakami, M., Hirose, K., Kawamura, K., Sata, N., and Ohishi, Y., 2004. Post-perovskite phase transition in $MgSiO_3$. *Science*, **304**, 855–858.

O'Connell, R., Gable, C., and Hager, B. 1991. Toroidal-poloidal partitioning of lithospheric plate motion. In Sabadini, R. et al. (ed.), *Glacial Isostasy, Sea Level and Mantle Rheology*. Norwell: Kluwer Academic, pp. 535–551.

Pearson, J., 1958. On convection cells induced by surface tension. *Journal of Fluid Mechanics*, **4**, 489–500.

Pekeris, C., 1935. Thermal convection in the interior of the earth. *Monthly Notices of the Royal Astronomical Society. Geophysical Supplement*, **3**, 343–367.

Ranalli, G., 1995. *Rheology of the Earth*. London: Chapman and Hall Publishers.

Regenauer-Lieb, K., Yuen, D., and Branlund, J., 2001. The initiation of subduction: Criticality by addition of water? *Science*, **294**, 578–580.

Revenaugh, J., and Sipkin, S., 1994. Seismic evidence for silicate melt atop the 410-km mantle discontinuity. *Nature*, **369**, 474–476.

Ricard, Y., and Bercovici, D., 2009. A continuum theory of grain size evolution and damage, *Journal of Geophysical Research*, **114**(B01204), doi:10.1029/2007JB005,491.

Runcorn, S., 1962a. Towards a theory of continental drift. *Nature*, **193**, 311–314.

Runcorn, S., 1962b. Convection currents in the earth's mantle. *Nature*, **195**, 1248–1249.

Schubert, G., and Zhang, K., 1997. Foundering of the lithosphere at the onset of subduction. *Geophysical Research Letters*, **24**, 1527–1529.

Schubert, G., Turcotte, D., and Olson, P., 2001. *Mantle Convection in the Earth and Planets*. Cambridge: Cambridge University Press.

Solomatov, V., 1995. Scaling of temperature dependent and stress dependent viscosity convection. *Physics of Fluids*, **7**, 266–274.

Song, T., Helmberger, D. V., and Grand, S. P., 2004. Low velocity zone atop the 410 seismic discontinuity in the northwestern United States. *Nature*, **427**, 530–533.

Stern, R. J., 2004. Subduction initiation: spontaneous and induced. *Earth and Planetary Science Letters*, **226**, 275–292.

Strutt, J. W. (Lord Rayleigh), 1916. On convective currents in a horizontal layer of fluid when the higher temperature is on the underside. *Philosophical Magazine*, **32**, 529–546.

Tackley, P., 2002. Strong heterogeneity caused by deep mantle layering. *Geochemistry, Geophysics, Geosystems (G3)*, **3**, 1024, doi:10.1029/2001GC000,167.

Tackley, P., 2007. Mantle geochemical dynamics. In Bercovici, D., and Schubert, G. (eds.), *Treatise on Geophysics: Mantle Dynamics*. New York: Elsevier, Vol. 7, pp. 437–505.

Tackley, P., Stevenson, D., Glatzmaier, G. A., and Schubert, G., 1993. Effects of an endothermic phase transition at 670 km depth in a spherical model of convection in the earth's mantle. *Nature*, **361**, 699–704.

Tackley, P., 2000. Self-consistent generation of tectonic plates in time-dependent, three-dimensional mantle convection simulations, 2. strain weakening and asthenosphere. *Geochemistry Geophysics Geosystems (G^3)*, **1**, 2000GC000,043.

Tauzin, B., Debayle, E., and Wittlinger, G., 2010. Seismic evidence for a global low-velocity layer within the earth's upper mantle. *Nature Geoscience*, **3**, 718721, doi:10.1038/ngeo969.

Toth, G., and Gurnis, M., 1998. Dynamics of subduction initiation at preexisting fault zones. *Journal of Geophysical Research*, **103**, 18053–18067.

Turcotte, D., and Schubert, G., 1982. *Geodynamics*. New York: Wiley.

Valencia, D., O'Connell, R., and Sasselov, D., 2007. Inevitability of plate tectonics on super-earths. *Astrophysics Journal*, **670**, L45–L48.

van der Hilst, R., Widiyantoro, S., and Engdahl, E., 1997. Evidence for deep mantle circulation from global tomography. *Nature*, **386**, 578–584.

van Heck, H., and Tackley, P., 2008. Planforms of self-consistently generated plates in 3d spherical geometry. *Geophysics Research Letters*, 35, L19,312, doi:10.1029/2008GL035,190.

van Keken, P., Hauri, E., and Ballentine, C., 2002. Mantle mixing: The generation, preservation, and destruction of chemical heterogeneity. *Annual Review of Earth and Planetary Sciences*, **30**, 493–525.
Vine, F., and Matthews, D., 1963. Magnetic anomalies over oceanic ridges. *Nature*, **199**, 947949.
Wegener, A., 1924. *The Origin of Continents and Oceans*. London: Methuen.
Weidner, D., and Wang, Y., 1998. Chemical- and clapeyron-induced buoyancy at the 660 km discontinuity. *Journal of Geophysical Research*, **103**, 7341–7441.
Weinstein, S., and Olson, P., 1992. Thermal convection with non-newtonian plates. *Geophysical Journal International*, **111**, 515–530.
White, D., 1988. The planforms and onset of convection with temperature-dependent viscosity. *Journal of Fluid Mechanics*, **191**, 247–286.
Wolfe, C., Solomon, S., Laske, G., Collins, J., Detrick, R., Orcutt, J., Bercovici, D., and Hauri, E., 2009. Mantle shear-wave velocity structure beneath the Hawaiian hotspot. *Science*, **326**, 1388–1390.
Zhong, S., Zuber, M., Moresi, L., and Gurnis, M., 2000. Role of temperature-dependent viscosity and surface plates in spherical shell models of mantle convection. *Journal of Geophysical Research*, **105**, 11063–11082.

Cross-references

Core Dynamo
Core-Mantle Coupling
Mantle D″ Layer
Mantle Viscosity

MANTLE D″ LAYER

Thorne Lay
Earth and Planetary Sciences Department, University of California, Santa Cruz, CA, USA

Definition and introduction

Earth's lower mantle extends from the base of the transition zone (commonly assigned a depth of 660 km, corresponding to the average depth of a global seismic velocity discontinuity) to the core-mantle boundary (CMB) at a depth of ~2,900 km below the surface. This thick layer of rock comprised of silicate and oxide minerals has gradual increases with depth of P- and S-wave seismic velocities and density that are generally consistent with adiabatic self-compression of a uniform composition material over most of the depth range (see *Earth's Structure, Lower Mantle*). Large-scale lateral heterogeneities in the seismic velocities of about ±1% are apparent in the central lower mantle, and these likely involve thermal and chemical signatures of upwellings and downwellings in the slowly convecting deep mantle system (see *Mantle Convection*). The deepest few hundred kilometers of the lower mantle have generally reduced seismic velocity gradients, localized seismic velocity discontinuities, and strong large-scale seismic velocity heterogeneities of ±2–4%. This portion of the lower mantle is called the D″ layer (or, more commonly, the D″ region, because the onset of anomalous properties varies significantly in height above the CMB from place to place) in recognition of its distinctive properties relative to the shallower lower mantle (D′ layer). Localized regions of extreme (10–30%) seismic velocity reductions in thin layers or lumps only tens of kilometers thick are found at the base of D″. The structural heterogeneities in the D″ region are usually interpreted as manifestations of complex thermal and chemical boundary layers at the CMB, to some extent mirroring the structural heterogeneities found in the near-surface lithospheric boundary layers. The structure and dynamics of the D″ region are believed to be of fundamental importance to the mantle and core dynamical systems, prompting many seismological, geodynamical, and mineral physics investigations of D″ properties.

The D″ region may have distinct bulk composition from the rest of the lower mantle, involving residues from deep mantle melting, core formation, chemical reactions between the core and mantle, or dynamical segregation of materials over time, but this possibility has to be assessed in the context of expected behavior of major lower-mantle minerals for the extreme pressure-temperature (P-T) conditions near the CMB. The primary minerals in the lower mantle are thought to be $(Mg_x,Fe_{1-x})SiO_3$ perovskite, $(Mg_x,Fe_{1-x})O$ ferropericlase, and smaller amounts of $CaSiO_3$ perovskite and high-pressure forms of SiO_2 (see *Earth's Structure, Lower Mantle*). The relative amount of Mg versus Fe, given by x, is ~0.9, but the properties of the minor Fe component are being intensely investigated because transitions from high-spin to low-spin in Fe^{3+} and Fe^{2+} are expected in lower mantle perovskite and ferropericlase as pressure increases, with strong effects on chemical and transport properties of the material in the deep mantle. Some Al may substitute in both the Mg and Si sites in the Mg-perovskite. Mg-perovskite comprises as much as 70% of the lower mantle, making it the most abundant mineral in the Earth. A phase change in Mg-perovskite, discovered in 2004, occurs under P-T conditions existing a few hundred kilometers above the CMB, likely giving rise to some of the complexities of the D″ region (Hirose and Lay, 2008). Mineralogical complexity of D″ is expected to be enhanced relative to the bulk of the lower mantle due to proximity to huge contrasts in composition and physical properties across the CMB.

The CMB separates the solid silicate- and oxide-mineral mantle from the molten metallic-alloy outer core. The density increases from about 5,500 kg/m^3 to 9,900 kg/m^3 across the CMB, and the viscosity decreases by many orders of magnitude. These contrasts maintain a very sharp boundary and result in thermal and chemical boundary layers in both D″ and the outermost core, although any core structures may be very thin and hard to detect (see *Earth's Structure, Core*). With any mass flux across the CMB being constrained to involve chemical diffusion, heat transports across the CMB by conduction from the hot core into the lower-temperature

mantle. The temperature drop across the resulting CMB thermal boundary layer is estimated to be in the range 1,000–2,000°C, and there is a heat flow of 5–15 TW across the boundary (the heat flow through Earth's surface is ~46 TW) (Lay et al., 2008). Over time, relatively dense materials in the mantle (for example, any basaltic component in subducted slabs that penetrate into the lower mantle) are likely to have concentrated in D″, while light dross, expelled from the core may have accumulated on the underside of the CMB. Thermo-mechanical and electromagnetic interactions across the CMB couple the mantle and core dynamical systems, and this influences the geodynamo (see *Core-Mantle Coupling*). The D″ region is relatively accessible to seismic imaging, and its inferred structural and dynamical properties are discussed here. Trønnes (2009), Garnero and McNamara (2008), Lay (2007), and many papers in Hirose et al. (2007) provide technical reviews of ongoing research topics and conceptual models for structures and processes occurring in D″.

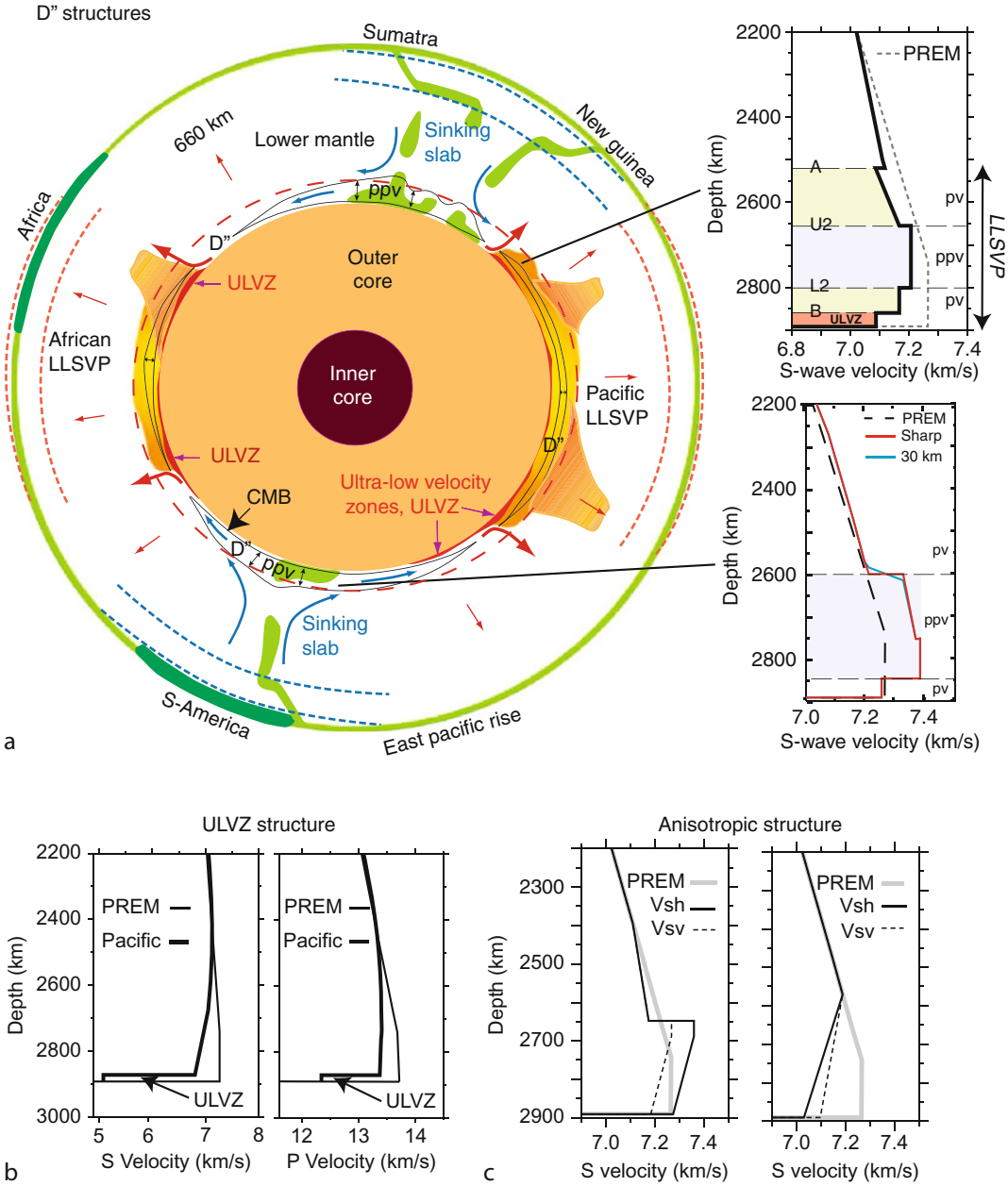
Seismic velocity models for D″

Seismology provides the highest resolution of D″ elastic material properties because it is possible to decipher seismic wave interactions with the lowermost mantle and CMB in recordings of ground shaking at seismic stations around the world. Seismic wave observations can resolve D″ region P- and S-wave velocities, reflections from sharp velocity jumps, shear-wave splitting caused by anisotropy of the elastic structure, and scattering properties. One-dimensional (1D) depth-varying seismic velocity models, such as the Preliminary Reference Earth Model (PREM) of Dziewonski and Anderson (1981), usually have reduced velocity gradients in the deepest ~150 km of the mantle, representing the global departure of D″ velocity structure from that of the overlying lower mantle. Efforts were initially made to interpret the low (or even negative) velocity gradients in such laterally averaged models of D″ structure as the result of a superadiabatic temperature increase across the CMB thermal boundary layer; strong temperature increases are required to account for seismic velocity reductions due to the low thermal expansion coefficient of deep mantle rock under high pressures. However, it is now clear that there is no meaningful "average" 1D structure for the D″ region useful for a robust interpretation of the boundary layer. It is more useful to discuss the seismic properties of D″ by emphasizing the lateral variations in structure, as is also the case for the lithosphere.

Seismic tomography has been applied to develop three-dimensional (3D) seismic velocity models for the entire mantle for over 30 years. This method extracts 3D P-wave and S-wave velocity fluctuations relative to 1D reference models by inversion of massive data sets with crossing ray paths for which seismic wave travel times are measured. 3D velocity inversions indicate that seismic velocity heterogeneity in the mid-lower mantle is relatively weak, but that in the lowermost 300–500 km of the mantle heterogeneity increases with depth, becoming stronger in D″ than anywhere else in the Earth except for the uppermost mantle where major thermal, chemical and partially molten heterogeneities are known to exist.

An unanticipated aspect of the D″ seismic velocity heterogeneity revealed by seismic tomography is that it is dominated by large-scale patterns, with huge continuous volumes of high or low-seismic-velocity material having scale-lengths of hundreds to thousands of kilometers. This predominance of large-scale variations is observed in both P-wave and S-wave velocities, with relatively high-velocity regions of D″ underlying the circum-Pacific margins and two nearly antipodal regions of pronounced low S-wave velocity beneath the southern central Pacific and southern Africa/southern Atlantic/southern Indian Ocean (Figure 1a). High-seismic-velocity regions are likely to have lower temperature than low-seismic-velocity regions, although chemical differences and phase changes may contribute to the variations. Recent global tomographic models have ±4% S-wave velocity variations in D″, two or three times stronger than those found in the mid-lower mantle, and comparable to the ±5–8% variations at depths near 150 km in the upper mantle. While mid-mantle variations tend to have more spatially concentrated heterogeneities, there is a significant degree of radial continuity of high- and low-velocity anomaly patterns throughout the lower mantle, possibly linking subduction zones at the surface to high-velocity regions in D″ and hot-spot locations at the surface to low-velocity regions in D″ (Garnero and McNamara, 2008).

P-wave velocity variations in current tomographic models have ±1.0–1.5% fluctuations, with low-velocity areas beneath the southern Pacific and south Atlantic/Africa and high velocities under eastern Eurasia. There is not as strong of an increase in P-wave velocity heterogeneity in D″ relative to that in the overlying mantle as there is for S-wave velocity, which can be partially attributed to the expected greater sensitivity of S-wave velocity to temperature variations in a thermal boundary layer. There is a large-scale spatial correlation between S- and P-wave velocity patterns, as expected for temperature-induced variations, but in some regions, such as beneath the northern Pacific, this correlation breaks down. In other regions, such as beneath the central Pacific, the S-wave variations are much stronger than the P-wave variations even though both have the same sign. The decorrelation of the seismic velocities and the variation in their relative strengths provide strong evidence that thermal variations alone cannot explain the large-scale patterns of seismic velocity heterogeneity; so there is likely to be chemical heterogeneity present in D″ (Trampert et al., 2004). It is important to recognize that tomography patterns are relative to the global average velocities at D″ depths, and that these averages do not necessarily define "normal" mantle structure. This results in uncertainty in interpretations of the velocity fluctuations (for example, high velocity regions may be anomalously low temperature, perhaps due to cool downwellings that have disrupted an overall hot boundary



Mantle D″ Layer, Figure 1 (**a**) Schematic cross-section through the Earth indicating large structures in the D″ region (below the red dashed line) above the core-mantle boundary (CMB) (adapted from Trønnes, 2009). The two large low shear velocity provinces (LLSVPs) under Africa and the Pacific are indicated, with thin ultra-low velocity zones (ULVZs) at their base and margins. Areas of upwelling warm mantle surround and overlie the LLSVPs. Areas of relatively cool mantle downwellings are underlain by D″ material with post-perovskite (ppv) occurrence. Detailed seismic S-wave velocity models from both an LLSVP margin and a down-welling region are shown on the right, with the D″ velocity discontinuity seen at depths of 2,600–2,650 km. Both regions have a deeper velocity decrease that may represent conversion from post-perovskite back to perovskite (pv) in the steep thermal gradient above the CMB at 2,891 km depth. (**b**) S-wave and P-wave velocity models from areas of intense ULVZ under the central Pacific, showing strong velocity reductions below low velocity (LLSVP) regions. (**c**) Representative anisotropic S-wave velocity models, indicting high Vsh velocities below circum-Pacific regions with strong D″ discontinuities and high Vsv velocities in LLSVP environments.

layer, or low-velocity regions may be anomalously hot regions, perhaps partially melted, within a chemically distinct, high velocity layer). Resolution of finer scale seismic properties in D″ is pursued in order to overcome this ambiguity in interpreting global seismic tomography images.

Seismic velocity discontinuities in D″

Rapid P-wave and S-wave velocity increases and/or decreases are observed several hundred kilometers above the CMB in many regions (Lay and Garnero, 2007), and are sometimes used to define the top of D″. The most prominent structure is a rapid shear velocity increase of 1.5–3% 150–350 km above the CMB found over intermediate-scale (500–1,000 km) regions beneath circum-Pacific margins and some other locations (Figure 1a). The increase in velocity may be distributed over a few tens of kilometers or it may be very abrupt (a "discontinuity") (Lay, 2008), and in some regions it can vary in depth by as much as 100 km over lateral scales of just 200 km. The velocity increase is required to account for observed S-wave reflections that arrive before reflections from the CMB. Similar structures have been found in localized regions for P-waves, usually with a smaller velocity increase of 0.5–1.0%, but in some regions there is a P-wave velocity decrease at the same depth as an S-wave increase (Hutko et al., 2008). Lay and Garnero (2007) review many observations and models for this D″ seismic velocity discontinuity.

The D″ S-wave velocity discontinuity is commonly observed in regions with large volumes of high S-wave velocity material in D″, indicating that it is associated with relatively low temperature environments. An abrupt increase in velocity with depth is not expected for a thermal boundary layer structure, so most interpretations of this correlation invoke the notion of localized ponding of cool subducted lithospheric slabs that have sunk to the lowermost mantle, retaining enough thermal and chemical anomaly to account for the high seismic velocity. However, it is not clear that slab thermal and compositional anomalies can account for a sharp reflecting structure, so additional presence of a phase change may be required. Alternatively, chemically distinct high velocity material may be present in these regions, perhaps involving ancient accumulated material concentrated during core formation or segregated oceanic crustal materials that have accumulated over Earth history. Observations of an S-wave velocity increase beneath the central Pacific, a relatively low velocity region far from any historical subduction zone, complicates any attempt to interpret the velocity discontinuity solely as the result of recent slab thermal anomalies (Lay and Garnero, 2007).

Post-perovskite in D″

The major seismic velocity discontinuities in the transition zone are generally attributed to phase changes in upper mantle minerals that cause abrupt changes in density and elastic wave velocities. The experimental discovery of a phase transition in $MgSiO_3$ perovskite (Murakami et al., 2004) for P-T conditions near the top of D″ may provide a corresponding explanation for the D″ S-wave velocity discontinuity. The high-pressure polymorph is called post-perovskite, and laboratory and theoretical predictions of its properties indicate that it should have 1–2% higher S-wave velocity, little, or no change in P-wave velocity, and 1–1.5% higher density than perovskite, which can account for some of the seismological complexities of D″ (Hirose et al., 2007). Presence of Fe and Al in the perovskite cause the phase transition to occur over a pressure range rather than at a single pressure for a given temperature, and this may be difficult to reconcile with an efficient seismic wave reflector. The perovskite-to-post-perovskite transition occurs at lower pressure for lower temperatures, thus lateral temperature variations in a thermal boundary layer are expected to modulate the depth of the transition. If the phase transition occurs in a relatively cool region of accumulated downwelling slab material, the discontinuity should be higher above the CMB than in warmer areas of D″ with the same composition. Some calculations indicate that the pressure of the transition may decrease if Fe is present in the perovskite, thus chemical heterogeneity could also modulate the depth of the phase transition.

The P-T behavior of the post-perovskite phase change tends to enhance thermal instabilities within the thermal boundary layer, but its thermal transport properties are still uncertain due to the possibility of Fe high-spin to low-spin transition and the lack of constraint on Fe partitioning coefficients between post-perovskite and ferropericlase. One of the most important attributes of the phase transition is that it can provide an absolute temperature tie-point for D″, if an observed S-wave reflector at a specific depth (hence, pressure) is correctly attributed to the phase change and if the P-T-X behavior (where X indicates precise mineralogical composition) of the phase change is experimentally and theoretically constrained. This is important because estimates of temperature structure in the lowermost mantle and outermost core are largely based on extrapolations over tremendous depth ranges of laboratory-estimated-temperature tie-points for the olivine-spinel phase transition near 410-km depth and the iron alloy solidus at the 5,150-km deep inner core-outer core boundary.

Laboratory measurements indicate that the temperature at a depth of about 2,600 km is close to 2,500 K, if the D″ discontinuity is caused by transition to post-perovskite. With an estimated outermost core temperature of 3,500–4,000 K, this favors a 1,000–1,500° increase in temperature across the D″ thermal boundary layer. Such a strong temperature increase could cause a reversion from post-perovskite to perovskite to take place in the hottest region right above the CMB, and some evidence for such a second crossing of the phase boundary has been inferred from seismic observations of a S-wave velocity decrease 50–100 km above the CMB (Figure 1a; Lay et al., 2006; van der Hilst et al., 2007). The effects of the phase

transition on seismic velocities must be considered when inferring lateral temperature structure in D″ from seismic velocity variations, given that the volume of the high S-wave velocity post-perovskite material should vary laterally in either a depth-modulated layer or a "lens," depending on stability of post-perovskite right at the CMB.

Large low shear velocity provinces in D″

The two large antipodal regions with low S-wave velocity appear to be chemically distinct regions of D″ (Figure 1a). The margins of these large low shear velocity provinces (LLSVPs) have abrupt steep-walled lateral gradients, over scales of tens of kilometers, which indicates a chemical change relative to surrounding D″ material rather than just a thermal change. The region beneath the southern Atlantic and Africa has been modeled as having margins with an abrupt −1 to −3% S-wave velocity decrease 250–300 km above the CMB (shallowing to ∼800 km above the CMB under Africa), with average velocities in D″ that are 3–5% lower than for PREM (e.g., Wang and Wen, 2006). The sub-Pacific LLSVP has comparable velocity contrasts and varying vertical extent, and may be two separate mounds of material that extend upward into the central lower mantle. P-wave velocity tends to be slightly low in the LLSVPs, but less than what would be expected if the S-wave velocity reductions were entirely caused by high temperatures. Thus, there is an anomalously high incompressibility in the LLSVPs, which also favors distinct chemistry. While limited in resolution, free oscillation measurements suggest that the LLSVPs are relatively high density, adding to the evidence for chemical inhomogeneity, and suggesting that these are not buoyant 'superplumes' as some researchers have suggested. Internal structure of LLSVPs has been detected in the form of localized velocity discontinuities (Figure 1a) and the presence of underlying ultra-low velocity zones (see below). The velocity discontinuities may represent chemical heterogeneity or possibly post-perovskite phase transition in material with distinct Fe and Al content from surrounding high seismic velocity areas of D″.

The current position of the LLSVPs can be attributed to accumulation of chemically distinct, relatively dense material in D″ that has been displaced away from circum-Pacific areas of downwelling slab materials over the past several hundred million years (Garnero and McNamara, 2008). Thermal calculations indicate that any dense chemical piles in a convecting mantle will be relatively hot as a result of inefficient heat loss, so LLSVPs would likely have combined effects of high incompressibility, high temperature, and high density. This predicts complex dynamical behavior, but the presence of the structures in D″ at this point in Earth's evolution suggests that the features are long-lived and are either remnants of much larger chemical anomalies that have slowly been entrained by mantle flow or they are being regenerated by ongoing chemical differentiation. Aggregation of basaltic (ocean crust) components of subducted slabs is one possible mechanism for accumulation of chemically distinct material that may be sustaining the LLSVPs. The total volume of LLSVP material is several percent of the mantle, so these are significant chemical reservoirs. The possibility that the LLSVPs have been close to their present locations for hundreds of millions of years is supported by the fact that reconstructed emplacement locations of many Large Igneous Provinces (LIPs) overlie margins of the LLSVPs (Torsvik et al., 2006). If plume upwellings rise from the lateral margins of the dense piles, as suggested by numerous dynamical models, this association may be understood as a first order impact of D″ structure on geological processes at Earth's surface.

Ultra-low velocity zones in D″

Thin layers or mounds of material with very strong seismic velocity reductions have been detected in D″ just above the CMB (Figure 1b), and these are called ultra-low velocity zones (ULVZs). P- and S-wave velocity models for ULVZ may have a thin layer or mound of 10–40 km thick and from hundreds to thousands of kilometers across with P-wave velocity reductions of −4 to −10% and S-wave velocity reductions of −8 to −30% (Thorne and Garnero, 2004). These structures are commonly found near the margins of LLSVPs (Figure 1a), but also occur in localized regions elsewhere. They are detected in seismic waves that reflect from or graze along the CMB.

The magnitude of the velocity reductions in the ULVZs and the factor of 2–3 ratio of S-wave velocity/P-wave velocity decrement require either the presence of a melt component or very strong chemical contrast. The temperature in the D″ thermal boundary layer will reach a peak right at the CMB (which is nearly isothermal due to the rapid convective flow occurring in the core); so ULVZs are intrinsically the hottest regions in the mantle. However, they are not globally detectable (a very thin layer $< \sim 1$ km could be present everywhere without being resolved by seismic data); so a combination of partial melting and chemical heterogeneity is implied by the patchy nature of the thicker regions of ULVZ. Seismic data indicate that the ULVZ material may be ∼10% denser than surrounding D″ material, favoring high Fe content, which would contribute to the strong seismic wave velocity reductions. ULVZ affiliation with LLSVP margins may be associated with thermal convection in the LLSVPs, along with interactions with flow in the surrounding mantle (Garnero and McNamara, 2008). It is not yet clear how to account for ULVZ chemical evolution, but one possibility is that they are the residue of a much more extensive lower mantle magma ocean which has largely solidified (Labrosse et al., 2007).

Seismic velocity anisotropy in D″

The seismic velocity structure in D″ is more anisotropic than the shallower lower mantle, with seismic velocities

being dependent on the direction of propagation and the polarization of ground shaking. This results in shear-wave splitting, involving an S-wave separating into two-components with orthogonal polarization, one traveling slightly faster than the other while in the D″ region. By measuring the polarizations and travel time difference between the fast and slow S-waves, it is possible to determine the anisotropic characteristics of the medium. For most S-wave phases with ray paths grazing horizontally through the D″ region, the data can be explained by models in which horizontally polarized (SH) vibrations travel with 1–3% higher velocities (Vsh) than vertically polarized (SV) vibrations (Figure 1c). This behavior is consistent with the medium having vertical transverse isotropy (VTI), which can result from hexagonally symmetric minerals with vertically oriented symmetry axes or from stacks of thin horizontal layers with periodic velocity fluctuations. The regions with the best documented cases for strong VTI in D″ tend to have higher than average S-wave velocities and strong D″ discontinuities (Figure 1c). There are observations favoring slightly tilted (non-vertical) transverse isotropy (Maupin et al., 2005), which results in weak coupling of the SH and SV signals (the fast wave is still close to the SH polarization), as well as limited regions where SV signals are found to propagate with higher velocities (Vsv) than SH signals (such as in the central Pacific LLSVP, Figure 1c).

Anisotropy in D″ is likely to be caused by either lattice-preferred orientation (LPO) or shape-preferred orientation (SPO). LPO can arise when minerals systematically orient within a flowing medium. This may exist in D″ if dislocation creep occurs, which requires relatively high stresses and low temperatures as might be found near slab downwellings. Ferropericlase has been shown to be a viable candidate for development of LPO in horizontal shear flows above the CMB, if suitable temperatures and stresses are present. It is also possible that certain slip systems can be activated such that post-perovskite can account for shear-wave splitting observed in some regions. If a region of D″ is too warm for dislocation creep, deformation will occur by diffusion processes and seismic anisotropy is only viable if heterogeneities (either chemical blobs or pockets of partial melt) are systematically aligned with the flow. Transitions from horizontal to vertical flows may account for SPO transitions from VTI to horizontal transverse isotropy, which can have Vsv higher than Vsh. Efforts continue to improve the characterization of D″ anisotropy because it has the potential to reveal ongoing deformation processes occurring in the boundary layer.

Summary

The diverse seismic properties of the D″ region appear to demonstrate the presence of heterogeneity at the base of the mantle associated with thermal and chemical boundary layers of a complex dynamical system. There is a predominance of large-scale structures in D″, and the strength of heterogeneities at intermediate and large-scales appears to be significantly greater than in the overlying lower mantle. Large low shear velocity provinces (LLSVPs), ultra-low velocity zones (ULVZs), multiple seismic discontinuities, a major phase change, and seismic anisotropy are all fundamental attributes of the D″ region. The present day configuration of D″ structures represents a “snap-shot” of an evolving system, with some aspects reflecting very long time-scale processes (LLSVPs may have been in place for at least hundreds of millions of years; ULVZs may be the last remnant of an extensive magma ocean that dates back to core formation) and much shorter time scales (post-perovskite lenses may be found in recently downwelled slab materials and seismic anisotropy may be sustained by present day dynamic shear flows). Together with the ongoing role of D″ as a thermal boundary layer that regulates cooling of the core and as the site of electromagnetic-mechanical coupling of rotation between the core and mantle, the significance of this region for Earth dynamics appears more evident than ever. While debate continues regarding the extent to which mantle material is fluxing between the shallow and deep mantle, evidence has grown to support a significant feedback/control of D″ structures on geological processes at Earth’s surface.

Bibliography

Dziewonski, A. M., and Anderson, D. L., 1981. Preliminary reference earth model. *Physics of the Earth and Planetary Interiors*, **25**, 297–356.

Garnero, E. J., and McNamara, A. K., 2008. Structure and dynamics of Earth’s lower mantle. *Science*, **3230**, 626–628.

Hirose, K., and Lay, T., 2008. Discovery of post-perovskite and new views of the core-mantle boundary region. *Elements*, **4**, 183–190.

Hirose, K., Brodholt, J., Lay, T., and Yuen, D. A. (eds.), 2007. *Post-Perovskite: The Last Mantle Phase Transition*, American Geophysical Union, Monograph, 174, 287 pp.

Hutko, A. R., Lay, T., Revenaugh, J., and Garnero, E. J., 2008. Anti-correlated seismic velocity anomalies from post-perovskite in the lowermost mantle. *Science*, **320**, 1070–1074.

Labrosse, S., Hernlund, J. W., and Coltice, N., 2007. A crystallizing dense magma ocean at the base of the Earth’s mantle. *Nature*, **450**, 866–869.

Lay, T., 2007. 1.18 Deep Earth Structure - Lower Mantle and D″. In Romanowicz, B., and Dziewonski, A. (eds.), *Treatise on Geophysics*. Amsterdam: Elsevier. Seismology and Structure of the Earth, Vol. 1, pp. 619–654.

Lay, T., 2008. Sharpness of the D″ discontinuity beneath the cocos plate: implications for the perovskite to post-perovskite phase transition. *Geophysical Research Letters*, L03304, doi:10.1029/2007GL032465.

Lay, T., and Garnero, E. J., 2007. Reconciling the post-perovskite phase with seismological observations of lowermost mantle structure. In Hirose, K., Brodholt, J., Lay, T., and Yuen, D., (eds.), *Post-Perovskite*: *The Last Mantle Phase Transition*, American Geophysical Union, Monograph, pp. 129–153.

Lay, T., Hernlund, J., Garnero, E. J., and Thorne, M. S., 2006. A post-perovskite lens and D″ heat flux beneath the Central Pacific. *Science*, **314**, 1272–1276.

Lay, T., Hernlund, J., and Buffett, B. A., 2008. Core-mantle boundary heat flow. *Nature Geoscience*, **1**, 25–32.

Maupin, V., Garnero, E. J., Lay, T., and Fouch, M. J., 2005. Azimuthal anisotropy in the D″ layer beneath the Caribbean. *Journal of Geophysical Research*, **110**, doi:10.1029/2004JB003506.
Murakami, M., Hirose, K., Kawamura, K., Sata, N., and Ohishi, Y., 2004. Post-perovskite phase transition in $MgSiO_3$. *Science*, **304**, 855–858.
Thorne, M. S., and Garnero, E. J., 2004. Inferences on ultralow-velocity zone structure from a global analysis of SPdKS waves. *Journal of Geophysical Research*, **109**, B083201, doi:10.1029/2004JB003010.
Torsvik, T. H., Smethurst, M. A., Burke, K., and Steinberger, B., 2006. Large Igneous Provinces generated from the margins of the large low velocity provinces in the deep mantle. *Geophysical Journal International*, **167**, 1447–1460.
Trampert, J., Deschamps, F., Resovsky, J., and Yuen, D., 2004. Probabilistic tomography maps chemical heterogeneities throughout the lower mantle. *Science*, **306**, 853–856.
Trønnes, R. G., 2009. Structure, mineralogy and dynamics of the lowermost mantle. *Minerlogy and Petrology*, doi:10.1007/s00710-009-0068-z.
van der Hilst, R. D., de Hoop, M. V., Wang, P., Shim, S.-H., Ma, P., and Tenorio, L., 2007. Seismic-stratigraphy and thermal structure of Earth's core-mantle boundary region. *Science*, **315**, 1813–1817.
Wang, Y., and Wen, L., 2006. Geometry and P- and S-velocity structures of the 'African Anomaly'. *Journal of Geophysical Research*, doi:10.1029/2006JB004483.

Cross-references

Core-Mantle Coupling
Earth's Structure, Core
Earth's Structure, Lower Mantle
Mantle Convection
Mantle Plumes

MANTLE PLUMES

Cinzia G. Farnetani[1], Albrecht W. Hofmann[2,3]
[1]Institut de Physique du Globe de Paris, Paris, France
[2]Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, USA
[3]Max-Planck-Institut für Chemie, Mainz, Germany

Definition

Mantle plumes. localized upwelling currents of solid rock that are hotter, thus less dense, than the surrounding mantle.

Hotspots

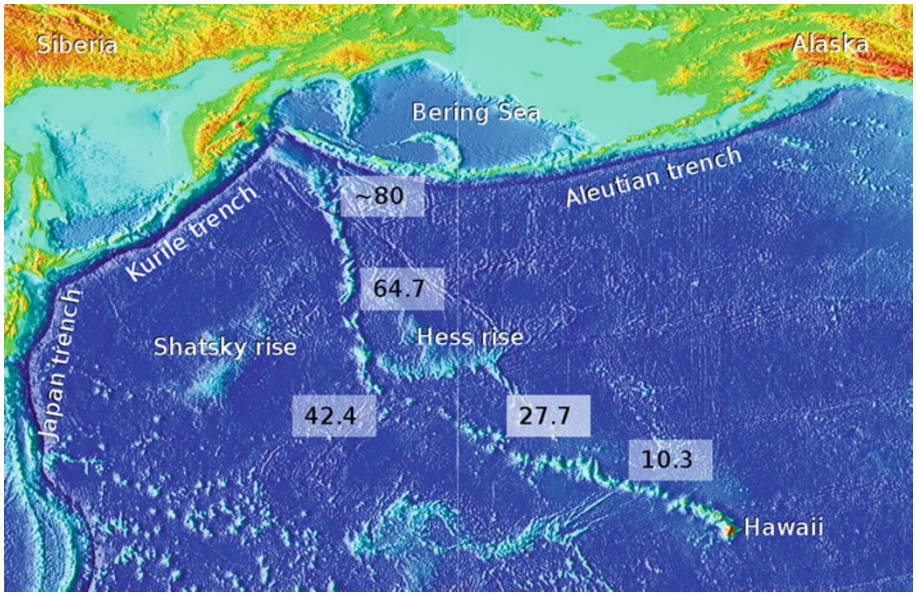
Plate tectonics provides a framework for interpreting volcanism at plate boundaries, namely, along spreading ridges (divergent boundaries) and subduction zones (convergent boundaries). However, it does not explain intraplate volcanism, such as Hawaii, nor the excess volcanism along certain sectors of a spreading ridge, as observed in Iceland. In these regions, called hot spots, the volcanic activity can last more than a hundred million years, as indicated by the resulting age-progressive volcanic chain formed during the hotspot's lifetime. For example, the 6,000 km-long Hawaiian-Emperor volcanic chain (Figure 1) is an alignment of roughly 100 volcanoes of progressively increasing ages. The volcanic chain was created during the last 80 Ma, as the Pacific plate drifted over the Hawaiian hot spot. The vigorous and long-lasting Hawaiian volcanism readily captured the interest of geophysicists: Tuzo Wilson (1963) proposed that the Hawaiian chain was not caused by lithospheric fissures, but by convection currents in the mantle, and Jason Morgan (1971) suggested that hotspots are the surface expression of mantle plumes upwelling from the Earth's lowermost mantle. Mantle plumes can be defined as localized upwelling currents of solid rocks that are hotter, and thus less dense, than the surrounding mantle.

Hotspots are traditionally characterized by some or all of the following features, although exceptions do occur: (1) An age-progressive volcanic chain whose linear trend is consistent with the direction of plate motion. (2) The onset of hotspot magmatism is often marked by a Large Igneous Province (LIP), a term including "continental flood basalts" (CFB) and "oceanic plateaus." Estimated extruded volumes of CFBs are $1–2 \times 10^6$ km^3, with perhaps similar (but unknown) volumes of intrusive magmatism, whereas oceanic plateaus can be one order of magnitude more voluminous (Coffin and Eldholm, 1994). According to the mantle plume initiation model (Richards et al., 1989), the transient and episodic LIP magmatism corresponds to melting of a large plume "head," whereas the subsequent hotspot activity is associated with the long-lasting and narrow plume "tail." (3) The topographic swell is a region of anomalously high topography with a lateral width of ~1,000 km in the direction normal to the volcanic chain (Wessel, 1993) and with an elevation of ~1 km, which decreases along the chain. (4) Hotspot basalts are geochemically distinct and more diverse than mid-ocean ridge basalts.

Some hotspots, however, do not show all of the above features. For example, many volcanic chains lack a clear age progression and, according to Ito and van Keken (2007), there are only 13 long-lived (>50 Ma) and 8 short-lived (<20 Ma) age-progressive volcanic chains. Some hotspots are not associated with a LIP, whereas others such as the Hawaii and Louisville chains terminate in a subduction zone, so that the complete time record of the volcanic activity is lost. A number of LIPs do not have any volcanic track (e.g., the Shatsky rise and the Hess rise in the Pacific Ocean) and their origin is still a matter of debate. In summary, plume magmatism is often, but not always, explained by the classical thermal plume model with a voluminous spherical head followed by a narrow columnar conduit.

Global hotspot distribution and hotspot fixity

Over the years, the estimated number of hotspots has varied from 20 (Morgan, 1971), to a maximum of 117 in the 1980s, whereas in recent compilations (e.g., Ito and van Keken, 2007), the number ranges between 45 and 70



Mantle Plumes, Figure 1 Seafloor topography from Smith and Sandwell (1997). Numbers indicate some ages (in Ma) along the Hawaiian-Emperor volcanic chain.

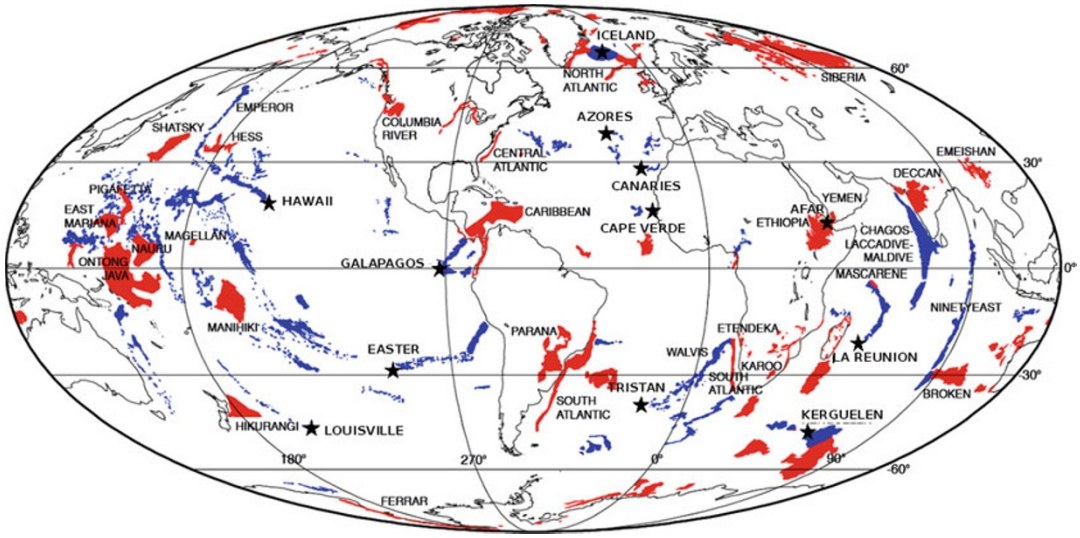
(see Schubert et al., 2001 and references therein). Hotspots younger than 100 Ma are generally active, although their vigor may vary considerably; older hotspots are either waning (100<age<140 Ma) or inactive (age>150 Ma). The best-defined hotspots appear to be relatively stationary over time and are used as a reference frame to determine absolute plate motions. Most hotspots are situated in the oceans, one exception being Yellowstone (USA), whose volcanic activity can be traced back to the 16 Ma-old Columbia River flood basalt. The discrepancy in the number of continental vs. oceanic hotspots has three probable reasons: First, the arrival of a mantle plume can weaken the lithosphere and enhance continental breakup (Courtillot et al., 1999). For example, the magmatic activity of the Iceland plume started with the eruption of the North Atlantic Tertiary Igneous Province (62 Ma ago) and was followed by continental rifting and the appearance of an oceanic spreading ridge after a few million years. In this respect, mantle plumes may have been influential in modifying plate boundaries. The second reason concerns the different thickness (100–150 km) between oceanic and continental lithosphere: weak plumes may not melt beneath a thick continental lithosphere, but are likely do so at lower pressure, beneath a thinner oceanic lithosphere. Third, the lack of hotspots in continents could be due to the continents accumulating over downwellings, precisely where plumes are very unlikely to ascend.

The main oceanic hotspots and their related LIPs (Figure 2) in the Atlantic Ocean are Iceland (0–62 Ma, LIP: the North Atlantic Tertiary Igneous Province) and Tristan da Cunha (0–125 Ma, LIP: the Paraná–Etendeka Province), whereas other hotspots like the Azores (0–20 Ma), Canaries (0–68 Ma), and Cape Verdes (0-Miocene) are not clearly associated with a LIP. In the Indian Ocean, the main hotspots are la Réunion (0–65 Ma, LIP: Deccan Traps), Kerguelen (0.1–120 Ma, LIP: the Kerguelen plateau), and Afar (0–30 Ma. LIP: Ethiopian Traps). In the Pacific Ocean, the main hotspots are Hawaii (0–76 Ma, unknown LIP, since the chain ends in a subduction zone), Louisville (1–77 Ma, possibly 120 Ma if the associated LIP is the Ontong-Java Plateau), and Galapagos (0 to possibly 90 Ma if the associated LIP is the Caribbean Plateau). There are also some hotspots (e.g., in French Polynesia) that may have a plume origin, but with an unclear age progression. Finally, the oceanic floors are "littered" by hundreds of thousands of volcanic seamounts that are certainly not created by plumes (e.g., Hieronymus and Bercovici, 2000; Clouard and Gerbault, 2008).

An interesting aspect of hotspots is that they are relatively stationary with respect to each other. Their respective motions of 1–2 cm/year are much less than the plate velocities, thereby approximating a fixed hotspot frame of reference. However, Tarduno et al. (2009) calculated the age and the paleolatitude of volcanoes belonging to Hawaiian-Emperor chain and concluded that the Hawaiian hotspot moved southward at 4–5 cm/year during the period 80–47 Ma ago, whereas it remained relatively fixed (<2 cm/year) afterward. This complex behavior is probably due to the dynamical interaction between upwelling plumes and the "mantle wind" induced by large-scale mantle convection (Steinberger et al., 2004).

Evidence for mantle plumes

There is now widespread agreement on the existence of mantle plumes, although contrasting views do exist



Mantle Plumes, Figure 2 Distribution of hotspots (*stars*) and Phanerozoic LIPs. In *red*: LIPs (or portions thereof) generated by a transient "plume head." In *blue*: LIPs (or portions thereof) generated by a persistent "plume tail." (Modified from Coffin and Eldholm 1994.)

(see Foulger and Natland, 2003 and the web site www.mantleplumes.org). Several lines of evidence support the existence of mantle plumes:

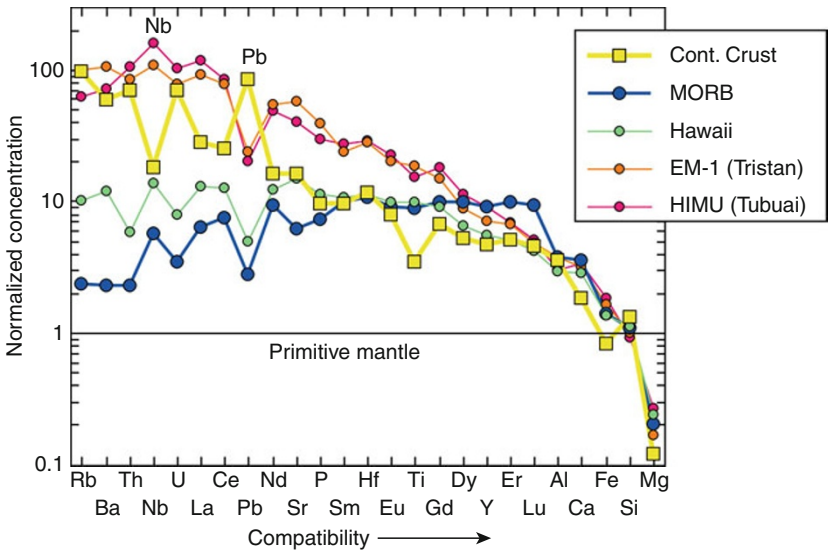
- First, the Earth's Rayleigh number, which governs the vigor of convection, is sufficiently high (10^6–10^8) to insure that mantle convection is time dependent and that its thermal boundary layers (TBLs) become repeatedly unstable. A TBL is a zone characterized by a high-temperature gradient, since heat is transported dominantly by conduction. Fluid dynamical considerations indicate that Rayleigh–Taylor instabilities from a hot TBL generate thermal plumes (Loper and Stacey, 1983). A prominent TBL in the Earth's mantle is possibly the D″ zone, which extends 100–200 km above the core–mantle boundary. The existence of another TBL, for example, at 660 km depth, is still a matter of debate, but it seems unlikely, since the endothermic phase transition does not constitute a complete barrier to mantle convection. Mantle plumes are, therefore, expected to rise from the lowermost mantle, forming cylindrical conduits with a radius of 50–150 km. Such values are based on fluid dynamics, whereas conduit radius estimated by seismology (e.g., Montelli et al., 2004; Wolfe et al., 2009) are much broader.
- Second, seismic detection of narrow plume conduits is challenging, but the first exhaustive study by Montelli et al. (2004) found that at least six plumes (Ascension, Azores, Canary, Easter, Samoa, and Tahiti) extend into the lowermost mantle, whereas others are confined to the upper mantle, and in some cases the model resolution was insufficient. Recently, an extensive ocean-bottom seismological survey of Hawaii (Wolfe et al., 2009) has shown that a low seismic velocity anomaly extends into the lower mantle. Hopefully, future progress in seismic tomography will provide us with further evidence for mantle plumes.
- Third, Oceanic Island Basalts (OIBs) are geochemically distinct and more diverse than Mid-Ocean Ridge Basalts (MORBs), as reviewed by Hofmann (1997). Moreover, isotopic signatures of OIBs indicate an involvement of ancient recycled oceanic crust, as first suggested by Hofmann and White (1982), whereas noble gases indicate that plumes may also carry primordial mantle material, possibly stored in the deep mantle.
- Fourth, Burke and Torsvik (2004) provide another set of evidence supporting the deep origin of mantle plumes. Their plate tectonic reconstruction over the last 200 Ma shows that the paleoposition of 90% of the 25 LIPs considered, were located, at the time of eruption, above lower mantle regions characterized, today, by low S-wave velocities. These broad regions, situated beneath the South-Central Pacific and Africa (see Romanowicz and Gung 2002 and references therein), are likely hotter and possibly compositionally denser than the surrounding mantle and may indeed represent a long-lived source zone of plumes.

Geochemistry of mantle plumes

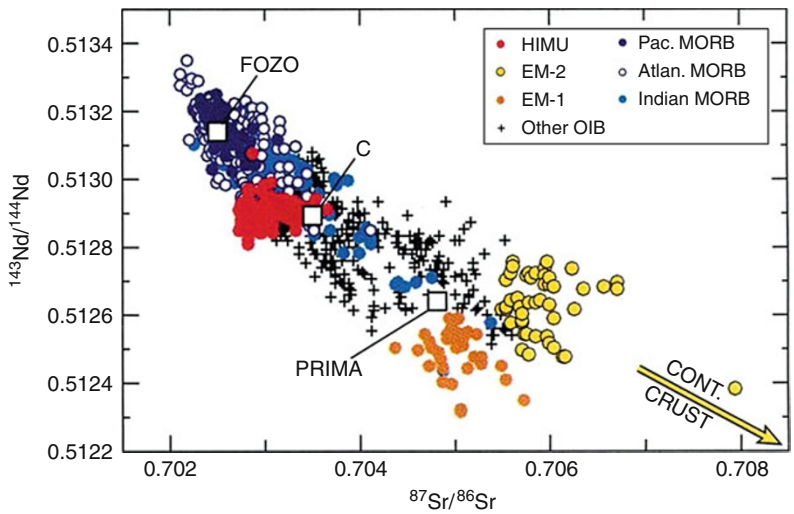
Geochemists became interested in mantle plumes when they discovered that ocean island basalts (OIBs), thought to be derived from plumes, tend to have different chemical (Schilling, 1973) and isotopic (Hart et al., 1973) compositions from mid-ocean-ridge basalts (MORBs). These differences are consistent with the plume model if the deep mantle, the inferred source of plumes, is compositionally different from the upper mantle, source of MORBs. These systematic differences were borne out by many subsequent studies of samples from virtually all available ocean

islands (Hofmann, 2003), as illustrated by Figure 3 for chemical compositions and Figure 4 for isotopic compositions. Both the trace elements and the isotopic compositions indicate that the MORB source has been depleted in "mantle-incompatible" trace elements, that is, those that are scavenged from the mantle by melts that ultimately form the crust, leaving behind a "depleted" mantle. In contrast, plume-type basalts tend to be derived from less depleted mantle sources.

The above observations and conjectures led to the two-layer mantle model, in which "normal" melting producing MORBs and subduction-related magmas have extracted most of the incompatible elements from the upper mantle and sequestered them in the continental crust. By contrast,



Mantle Plumes, Figure 3 Concentrations of selected trace and major elements, arranged in the order of ascending compatibility and normalized to primitive-mantle concentrations, for average continental crust, average MORB, average Mauna Loa, Hawaii, and OIB type representing EM-1 (Enriched mantle 1) and HIMU (high-μ, where $\mu = {}^{238}U/{}^{204}Pb$). The patterns for MORB and OIB differ by their enrichments, but show similar Nb and Pb anomalies, opposite to those of the continental crust. (From Hofmann 1997.)



Mantle Plumes, Figure 4 Nd and Sr isotopic compositions(*) of MORBs and OIBs. EM-1 (Enriched mantle 1), EM-2 (Enriched mantle 2), and HIMU (high μ, where $\mu = {}^{238}U/{}^{204}Pb$). The *squares* indicate compositions of primitive mantle (PRIMA) and of mantle components FOZO and C (see Hofmann, 1997 and references therein). (*)Some basic notions: ^{147}Sm decays to ^{143}Nd. Since Sm is more compatible than Nd, progressive melting enriches the residual rock in Sm, thus MORBs have high $^{143}Nd/^{144}Nd$. ^{87}Rb decays to ^{87}Sr. Since Rb is less compatible than Sr, progressive melting depletes the residual rock in Rb, thus MORBs have low $^{87}Sr/^{86}Sr$.

the lower mantle was thought to be largely undepleted or more "primitive." Thus, plumes, rising from the deep mantle, would sample the lower, relatively primitive reservoir, but they might entrain more depleted mantle rocks on their way up, thus producing a mixing array between primitive and depleted reservoirs (Jacobsen and Wasserburg, 1979). However, mass balance considerations for the incompatible element and isotope budget of the continental crust, the depleted upper mantle, and an undepleted lower mantle, demanded a size of the depleted reservoir of at least 50% of the mantle, which is significantly greater than the roughly 30% mass fraction of the upper mantle. This three-reservoir model (often simply called the "layered mantle" model) was reinforced by the observation that many plume-derived basalts had much higher $^{3}He/^{4}He$ ratios than MORB (Farley and Neroda, 1998), where ^{3}He is a remnant of primordial noble gases from the primitive Earth, whereas ^{4}He is the product of subsequent decay of uranium and thorium.

In spite of its apparent geochemical persuasiveness, the layered mantle model suffered two essentially fatal blows. During the 1980s, evidence began to accumulate that showed isotopically "enriched" (i.e., crustal-like) mantle sources in several ocean islands including Hawaii, Pitcairn, and Tristan da Cunha. This means that the endpoint of the mixing array of the type shown in Figure 4 cannot be an undepleted or "primitive" mantle reservoir but must represent a source component that was crust-like in that it was actually enriched in incompatible elements. And although such a mixing array may pass through the isotopic locus of a primitive mantle reservoir, it clearly does not require the involvement of a primitive reservoir. This is because, when a primitive reservoir is differentiated into enriched and depleted components, any remixing of these differentiated components will produce a compositional array passing approximately through the starting point, the locus of primitive compositions. Therefore, Hofmann and White (1982) proposed that the enriched source components in mantle plumes actually come from subducted oceanic crust rather than a "primitive reservoir."

The second blow to the conventional layered mantle model came from seismic evidence showing tomographic images of high seismic velocities characterizing subduction slabs that penetrate the base of the upper mantle and can be traced into the lowermost mantle (e.g., van der Hilst et al., 1997). If such deep subduction occurred during major portions of Earth's history, convective mixing will have destroyed the chemical separation between upper and lower mantle. The first attempt to integrate numerical convection modeling with chemical differentiation at crustal levels and recycling of ocean crust to generate enriched plume sources at the base of the mantle was made by Christensen and Hofmann (1994), and this has been followed up by more elaborate simulations.

More recently, a variety of new models have sprung up. Some invoke irregularly distributed, relatively primitive plume source regions, which have been protected from convective stirring throughout Earth's history and are thus able to preserve primitive geochemistry. Tolstikhin and Hofmann (2005) and Boyet and Carlson (2005) have proposed a new form of a two-layer mantle, one in which a relatively small, compositionally dense reservoir formed at the base of the mantle, in effect constituting the D″ layer. This irregular layer, on average about 200 km thick, has otherwise been interpreted as a "slab graveyard." The new geochemical model stipulates that these subducted slabs are very ancient, perhaps only a few tens of millions of years younger than the accretion of the Earth. This, in turn, requires stabilization by high intrinsic density, either because they are derived from a primordial iron-rich mafic crust, or because they were generated by downward segregation of dense partial melts in the lowermost mantle (Labrosse et al., 2007) in effect also creating an Fe-rich, dense bulk composition. Such a "new two-layer" mantle model can account for the geochemical differences between plumes and mid-ocean ridge basalts, and it resolves at least two awkward problems with the old model: (1) it does not require that plumes have particularly high helium concentrations, which are never actually observed in plume-derived basalts; (2) it explains trace element characteristics of plume-derived basalts, which are in fact inconsistent with primitive sources. The new two-layer model, by its nature, eliminates the need for any undifferentiated, primitive reservoir, but it creates a repository of primordial noble gases in the permanently sequestered D″-layer at the base of the mantle. From there, these noble gases can easily diffuse into the overlying actual (silicate) source reservoirs of mantle plumes.

It should be emphasized that, by its very nature, geochemical evidence is extremely unlikely to either prove or disprove the existence of mantle plumes or mantle layering, because the mixing and extraction processes that deliver plume-derived or non-plume-derived melts to the surface are not known a priori. Unfortunately, in the past, it has too often been argued that some ocean island must be derived from a deep-mantle plume because it contains high $^{3}He/^{4}He$ ratios, or near-chondritic $^{143}Nd/^{144}Nd$ ratios. Indeed, the inherent weakness of such arguments, which are clearly nonunique, has helped to discredit plume theory in some quarters. On the other hand, the plume model in combination with geochemical evidence can help to elucidate many fundamental aspects of Earth evolution, including the timescale of early Earth evolution and of crustal recycling, as well as the intensity of convective stirring in the mantle.

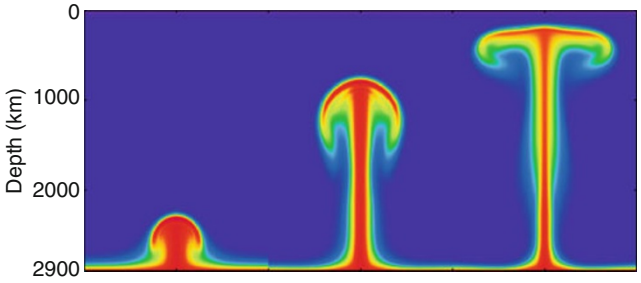
Overall, the "enriched" nature (i.e., enriched in highly incompatible elements relative to more compatible elements), which is evidenced by both the observed trace element and isotope abundances of the decay products of long-lived radioactive decay systems such as ^{87}Rb-^{87}Sr, ^{147}Sm-^{143}Nd, ^{176}Lu-^{176}Hf in plume-derived rocks, is most easily explained by the subduction and recycling of enriched crustal rocks, rather than the involvement of any "primitive" mantle material (Hofmann, 1997). Subduction of crustal materials, including ordinary ocean

crust, enriched ocean crust, ocean islands and seamounts, as well as some sediments and other continental material, is based on geological and geophysical observations. Thus, there is no shortage of suitable source materials in the mantle without the need for alternative enrichment processes, such as "metasomatic" infiltration, which might nevertheless play some role. The recycled enriched materials appear to be more prevalent in mantle plumes than in mid-ocean ridge basalts, because basaltic crust, once subducted, is somewhat denser than ordinary peridotitic mantle. Because of this, it may be segregated and "stored" at the base of the mantle for geologically longer periods of time, and ultimately contribute significant portions of plume source materials (Christensen and Hofmann, 1994).

Genesis and dynamics of plumes and superplumes

On timescales of millions of years, solid mantle rocks behave as highly viscous fluids with a viscosity of $\sim 10^{18}$–10^{22} Pa s (in comparison, the viscosity of a glacier is $\sim 10^{13}$ Pa s, and of a basaltic lava is ~ 10–10^{4} Pa s). Thus, mantle dynamics is governed by a set of equations (conservation of mass, momentum, and energy) for a viscous fluid where inertial effects can be neglected. The growth rate of a Rayleigh–Taylor instability from a thermal boundary layer heated from below is controlled by the Rayleigh number: $Ra = \rho\, \alpha\, \Delta T\, g\, d^3/k\, \eta$ where ΔT is the temperature contrast across a layer of thickness d, density ρ, thermal expansion coefficient α, thermal diffusivity k, and viscosity η. Instabilities are enhanced by the thermal buoyancy and inhibited by viscosity and thermal diffusivity. Using reasonable values for the above physical parameters, growing instabilities of the boundary layer form diapirs. Due to the high mantle viscosity, such a diapir will separate from the TBL only when its volume becomes sufficiently large (e.g., Whitehead and Luther, 1975). The morphology of a thermal plume is controlled by the viscosity contrast between the hot fluid and the mantle above it: If the viscosity contrast is weak, plumes will have a "spout" shape, with little difference between the radius of the leading diapir (the plume head) and the following conduit (the plume tail), whereas increasing viscosity contrast leads to a larger head and a narrower conduit. This "mushroom" shape (Figure 5) is favored, because a hot plume is likely to be 100 times less viscous than the surrounding mantle, owing to the strong temperature dependence of viscosity.

Although fluid dynamics laboratory experiments (e.g., Whitehead and Luther, 1975; Griffiths and Campbell, 1990) and numerical simulations (e.g., Parmenteir et al., 1975; Olson et al., 1993) on purely thermal plumes enabled us to gain a quantitative understanding of plume dynamics, in the 1990s it became progressively clear that the lowermost mantle is compositionally heterogeneous. D″ is a region of preferential segregation and accumulation of denser subducted crust (Christensen and Hofmann, 1994), and larger scale regions of the lower mantle may be chemically



Mantle Plumes, Figure 5 Numerical simulation of a thermal plume.

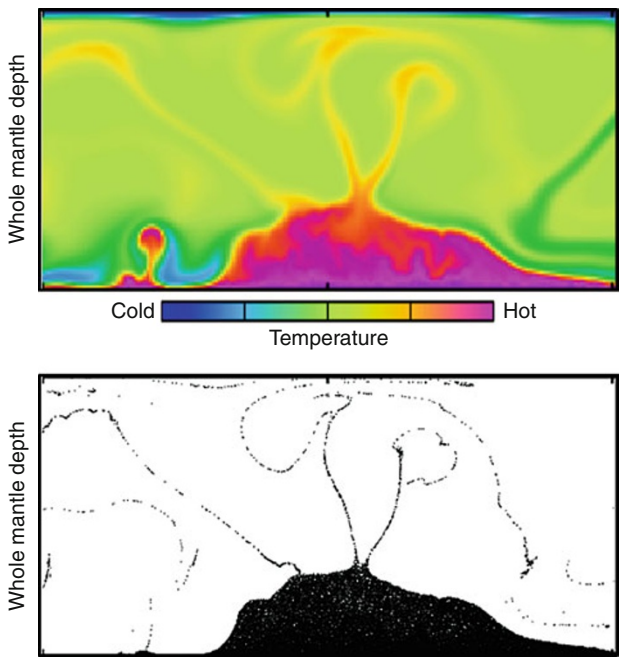
heterogeneous (Kellogg et al., 1999). This offered new and exciting avenues to explore the complex dynamics of thermochemical plumes, which can be defined as hot (positively buoyant) plumes that carry compositionally denser (negatively buoyant) material. Laboratory experiments by Davaille (1999) investigated a variety of regimes and found that instabilities may form dome-like structures with an oscillatory behavior (i.e., they rise and sink in response to a subtle balance between thermal and chemical buoyancies). Her experiments, together with numerical simulations of thermochemical convection (e.g., Tackley, 1998) provide a fluid dynamically consistent framework to interpret observations that are otherwise unexplained by purely thermal convection. For example, the commonly referred "superplumes," situated beneath the South-Central Pacific and Africa are broad (thousands of kilometers large) lower mantle zones of low seismic velocity most likely associated with active upwelling. Although there is a debate on the thermal and/or compositional origin of "superplumes," several lines of evidence do support their distinct composition (Ishii and Tromp, 1999; Masters et al., 2000). Another issue that can be readily explained by thermochemical plumes is the discrepancy between the petrologically constrained excess temperature of plumes (100–250°C) and the estimated temperature difference across the D″ region (~1,000°C). Numerical simulations (Farnetani, 1997) show that the deepest, hottest part of the thermal boundary layer does not upwell in mantle plumes if the compositional density contrast exceeds 2%. A denser zone in the boundary layer also tends to "anchor" the base of plume conduit, so that the flow pattern remains stable over timescales longer than the plume rise time (Jellinek and Manga, 2002). Finally, thermochemical plumes present a variety of shapes and surface manifestations (Farnetani and Samuel, 2005; Kumagai et al., 2007) that expand the classical plume head-tail model predictions (Figure 6).

Plume melting and plume strength

At spreading ridges, mantle rocks rise to shallow depths in response to plate spreading, and melt by adiabatic decompression (McKenzie and Bickle, 1988). This is not the case for most plumes impinging at the base of a preexisting and unrifted continental or oceanic lithosphere. In order to melt at relatively high pressure

(4–5 GPa), plumes must be either hotter than normal mantle, or compositionally more "fertile" than peridotitic mantle (Condie, 2001, Chap. 4). Plumes are probably both: their excess temperature is estimated between 100 and 250°C (Putirka, 2005), and they can carry 20%, or more, of recycled crust in the form of eclogite (Sobolev et al., 2007). Eclogite is more fertile than peridotite because it has a lower solidus and a greater melt productivity. According to Sobolev et al. (2005), reaction between eclogite-derived liquids and solid peridotite forms a pyroxenite, whose subsequent melting may explain some compositional characteristics of Hawaiian (and many other plume-derived) lavas. Another efficient way of lowering the solidus temperature is to have fluids (H_2O, CO_2) in the upwelling plume rocks; however, the evidence for "fluid-rich" hotspot lavas is scant. Although it is well known that melting is the most efficient way to create chemical differentiation in the mantle and that it is the unavoidable process to generate surface lavas from mantle rocks, many aspects of partial melting remain elusive (e.g., the pressure–temperature conditions during melting; the source heterogeneous composition; the mechanism of melting, of transport and of storage of the magma). This limits our quantitative understanding of the link between petrological and geochemical observations of surface lavas and the underlying mantle plume composition and dynamics.

The strength of a mantle plume can be calculated on the production rate of volcanic rocks. The method utilizes the volcano's volume and growth time, and the estimated melt fraction. For example, Mauna Loa has a volume of ~70,000 km^3 and it grew in ~1 Ma, suggesting that the Hawaiian plume melt production rate M is about 0.1–0.2 km^3/year. Since only a fraction of the upwelling plume melts, the total plume volume flux $Q_v \gg M$. Another method, based on the rate of swell formation (Sleep, 1990), enables one to calculate the plume buoyancy flux $B = Q_v \Delta\rho$ where $\Delta\rho = \rho\,\alpha\,(T_{PLUME} - T_{MANTLE})$. For Hawaii, the most vigorous hotspot, B ~ 8,000 kg/s, for Iceland B ~ 1,400 kg/s, for Galapagos B ~ 1,000 kg/s. Finally, the plume heat flow is $Q_h = \rho_{MANTLE} C_p Q_v (T_{PLUME} - T_{MANTLE})$, where the specific heat at constant pressure is $C_p \sim 1{,}200$ J/kg K. For Hawaii, the heat flow is $Q_h \sim 360 \times 10^9$ W, and it represents ~16% of the global hotspot heat flow of 2.3×10^{12} W, which, in turn is a small fraction of the Earth's total heat flow of 44×10^{12} W (for further reading, see Schubert et al., 2001, Chap. 11).



Mantle Plumes, Figure 6 Numerical simulation of a thermochemical plume. *Top*: Temperature field. *Bottom*: Chemically denser material, represented with tracers. Animations are available.

Continental flood basalts and continental breakup

The eruption of a large igneous province (LIP) represents a major geologic event head. Over the Earth's history, episodic LIP magmatism contributed to continental growth through the emplacement of CFB and through the accretion/obduction of fragments of oceanic plateaus onto continental crust (Ben-Avraham et al., 1981). Accretion may be due to the difficulty of subducting the 20–40 km thick crust of an oceanic plateau. Three notable examples of accreted LIP fragments are found in the Solomon Island arc (from the Ontong-Java plateau), in Central America (from the Caribbean plateau), and the Wrangellia terrane outcropping in British Columbia and in SE Alaska (Figure 7). This allochthonous terrane, locally attaining a thickness of 6 km, is likely to be a fragment of a 230 Ma-old oceanic plateau (Ben-Avraham et al., 1981; Richards et al., 1991; Greene et al., 2008).

Distinctive characteristics of LIP magmatism are the short duration (1–2 Ma) and the huge volumes ($1–10 \times 10^6$ km^3) of eruption. Because of the high eruption rates (>1–2 km^3/year; for comparison Kilauea grows at 0.1 km^3/year) and the enormous surface extent ($1–2 \times 10^6$ km^2; for comparison France's surface is $\sim 0.5 \times 10^6$ km^2), flood basalt eruptions are extraordinary volcanic events. CFBs consist of sub-horizontal flows of mafic (Fe- and Mg-rich) rocks, mainly tholeiitic basalts. Individual flows can extend for hundreds of kilometers, be tens to hundreds of meters thick, and have volumes of more than 10^3 km^3. CFBs are more accessible than submerged oceanic plateaus; however they can be extensively eroded, fragmented, and dispersed on different continents by plate tectonic processes. Clearly, uncertainties regarding the original volume and surface extent increase with their age, for example, for pre-Cambrian CFB only the giant swarm of mafic dikes, feeding surface volcanism, may be left (Ernst and Buchan, 2001).

The Phanerozoic CFBs are, in chronological order: The 258 Ma-old Emeishan Traps (SW China), which

Mantle Plumes, Figure 7 Photograph of 1,000 m of continuous subaerial flood basalt stratigraphy in the Wrangell Mountains, Alaska. The *yellow* line marks the contact between Nikolai basalts and the overlying Chitistone Limestone. (From Greene et al., 2008. With permission.)

presently cover only 0.3×10^6 km^2, probably a tenth of the estimated original surface; the 250 Ma-old Siberian Traps, which might have had a surface extent of $3–4 \times 10^6$ km^2 and extrusive volumes of 10^6 km^3; and the 200 Ma-old lavas and dike swarms of the Central Atlantic Magmatic Province outcrop in once-contiguous parts of North America (e.g., the Palisades sill, NY), West Africa, and Brazil. This flood volcanism, with an estimated total volume of 3×10^6 to 5×10^6 km^3, preceded the opening of the central Atlantic Ocean. The 184–182 Ma-old Karoo (Southern Africa)–Ferrar (Antarctica) traps preceded the breakup of Gondwana by 10–15 Ma. Similarly, the ~135–130 Ma-old magmatism of the Paraná (Brazil)–Etendeka (Namibia) province led to the opening of the South Atlantic Ocean. The Deccan Traps (India) erupted 66–65 Ma ago over an area of $\sim 1.5 \times 10^6$ km^2; the subsequent rifting split apart the Seychelles from India. The North Atlantic Tertiary Igneous Province covers a surface greater than 10^6 km^2 and has an estimated volume of 6×10^6 km^3. The earliest volcanism occurred as flood basalts in Baffin Island, West- and East-Greenland (62 Ma), and later (56–54 Ma) extended to the continental margins of Greenland, the British Isles, and Norway. These rifted edges of continents constitute the so-called volcanic passive margins of up to 8 km-thick seaward dipping basalt layers. The complete opening of the North Atlantic and the appearance of truly oceanic crust occurred about 53 Ma ago. Finally, the Ethiopian–Yemen traps, ~30 Ma old, have been associated with the rifting in the Red Sea and the Gulf of Aden (Courtillot et al., 1999).

Issues concerning the timing of rifting have been hotly debated: One side claims that rifting preceded, and enhanced, flood volcanism (e.g., White and McKenzie, 1989); the other side argues that rifting is not a prerequisite to flood volcanism, since it occurs only during and after volcanism (e.g., Richards et al., 1989). More geological observations are needed, as well as numerical simulations and/or laboratory experiments investigating how anomalously warm mantle can erode the lithosphere and create (or reactivate) weak zones susceptible to rifting.

LIP magmatism and environmental effects

The remarkable temporal correlation between LIP magmatism and mass extinctions of terrestrial and marine organisms suggests a cause-and-effect connection (Courtillot and Renne, 2003). The most notable example is the Siberian Traps at the Permian–Triassic boundary (Renne et al., 1995; Svensen et al., 2009). The emplacement of the Deccan Traps, spanning the Cretaceous–Tertiary boundary, certainly had an environmental effect (Self et al., 2006); however, the extinction event coincided with a large asteroid impact at Chicxulub, Mexico (Schulte et al., 2010). The peak of volcanism in the North Atlantic Volcanic Province corresponds to the Paleocene-Eocene Thermal Maximum (PETM) event 55 Ma ago. Svensen et al. (2004) propose that voluminous magmatic intrusions in carbon-rich sediments induced massive release of the methane buried in marine sediments, thus suggesting that volcanic and metamorphic processes associated with the opening of the North Atlantic may explain global climate events. Self et al. (2006), Ganino and Arndt (2009), and Svensen et al. (2009) attempt to quantify the release of volatiles (CO_2, CH_4, SO_2) during flood volcanism and to estimate the consequences on the environment. In particular, the release of volcanic ashes and sulfuric acid aerosols can lead to darkening and cooling, SO_2 and HCl cause acid rains, whereas greenhouse gases like CO_2 and CH_4 cause global warming. Carbon has two origins: magmatic and sediment-derived. Interestingly, Ganino and Arndt (2009) show that the mass of sediment-derived CO_2 can be four to eight times larger than the mass of magmatic CO_2, if contact metamorphism reactions occur in sedimentary rocks such as dolomite, evaporite, coal, or organic-rich shale. In other words, the environmental effect of flood basalt magmatism depends, among other factors, on the rock type in contact with magmatic sills and lava flows (Ganino and Arndt, 2009; Svensen et al., 2009).

Marine eruptions can also have an environmental effect, because they can cause variations in water chemistry and oceanic circulation. The emplacement of the (Alaska-sized) Ontong-Java plateau 120 Ma ago and of the Caribbean plateau correlate temporally with some of the major Oceanic Anoxic Events (OAEs) which occurred ~183, 120, 111, and 93 Ma ago. OAEs are associated with an abrupt rise in temperature, induced by rapid influx of CO_2. The consequent increase of organic productivity causes an increase of oxygen demand in the water column, eventually leading to oxygen depletion in the oceans (for a thorough review see

Jenkyns 2010 and references therein). Furthermore, the formation of black shale (marine carbon-rich sediments) may have been favored by massive hydrothermal release of trace metals, poisonous to marine life.

Plume–lithosphere interaction

A quantitative understanding of plume–lithosphere interaction enables us to relate geophysical observations (e.g., the spatiotemporal evolution of surface topography, variations in lithospheric thickness, etc.) to the physical parameters and dynamics of the plume. Central issues concerning plume–lithosphere interactions include, but are not limited to the following:

- First, estimating the dynamic topography (Definition of dynamic topography, extracted from Lithgow-Bertelloni and Silver, (1998). "Dynamic topography is a deformation of the surface of the Earth, supported by the vertical stresses at the base of the lithosphere, which are generated by flow in the mantle below. This is in contrast to the more familiar mechanism of isostatically supported topography which is in equilibrium at the Earth's surface and would exist even in a static mantle.") induced by the arrival of a mantle plume head. Pioneering laboratory experiments by Olson and Nam (1986) found that a thermal diapir upwelling and spreading beneath the lithosphere induces a rapid surface uplift followed by a slower subsidence. According to Farnetani and Richards (1994), uplift should precede flood volcanism by 1–2 Ma, a prediction that is sometimes validated by geological observations (e.g., for the Emeishan Traps), but not always (e.g., for the Siberian Traps). A present-day example of dynamic topography is provided by Lithgow-Bertelloni and Silver (1998), who suggest that the anomalous elevation (~1 km) of South-Central Africa has a dynamic origin, being induced by mantle upwelling from the African "superplume," a broad lower mantle zone of low seismic velocity anomalies.
- Second, understanding the origin of the topographic hotspot swell and the thermomechanical response of the lithosphere drifting over the plume. Two main mechanisms have been proposed for the Hawaiian swell: "thermal rejuvenation" invokes heating and thinning of the lithosphere above the hotspot, whereas "dynamic support" invokes stresses applied to the base of the lithosphere by the buoyant plume. Seismic studies by Li et al. (2004) around Hawaii found a thinning of the lithosphere from the expected 100 km (beneath the Big Island) to only 50–60 km (beneath Oahu and Kauai, 400 km downstream from the Big Island), thus suggesting a hybrid scenario, whereby "dynamic support" prevails above the plume conduit and "thermal rejuvenation" prevails downstream. Far from the conduit, thermal thinning of the lithosphere may be enhanced by small-scale convective instabilities, which take the form of rolls aligned with the plate motion, as observed in numerical simulations with strongly temperature-dependent viscosity (Moore et al., 1998).
- Third, mechanical and chemical interaction between plume magmas and the lithosphere can explain a number of geophysical and geochemical observations. For example, the occurrence of regularly spaced volcanoes in hotspot chains is probably caused by the interaction of magma transport with lithospheric flexure (Hieronymus and Bercovici, 1999, 2000). Under the load of the volcanoes, the lithosphere deforms and the resulting flexural stresses have a key role in determining the locus where magma preferentially extrudes. There are also clear examples of geochemical interaction between the continental lithosphere and plume magmas "en route" to the surface. In particular, CFB lavas erupted at the onset of flood volcanism often have anomalous compositions, reflecting contamination by the lithosphere and crust. It is thus important to quantify the contribution of the continental lithosphere before inferring the plume's geochemical fingerprint.
- Fourth, plume material may preferentially flow – and partially melt – in preexisting zones of thinned lithosphere (Sleep, 1996). In other words, the thickness of the lithosphere and the slope of its base may control the spatial distribution of plume magmatism. According to Ebinger and Sleep (1998), the Cenozoic magmatism in East Africa (including the Afar, the Cameroon volcanic line to the West, and the Comoros Islands to the South) can be explained by a single plume impinging beneath the Ethiopian plateau and its subsequent lateral flow, "channeled" in preexisting rift zones characterized by a relatively shallow base of the lithosphere.

Plume–ridge interaction

At least 21 hotspots are situated near a spreading ridge, the clearest examples being Iceland, a ridge-centered hotspot, Galapagos, and the Azores. At Iceland, the Mid-Atlantic Ridge rises ~4 km above normal and remains anomalously shallow for ~1,000 km to the North and the South. Below Iceland, the oceanic crust is 40 km thick, whereas more normal crustal thicknesses of 8–10 km are attained only ~500 km far from the hotspot. Furthermore, compositional anomalies in trace elements and isotope ratios observed along the ridge clearly indicate plume–ridge interaction (Schilling et al., 1985). The Galapagos hotspot is situated ~200 km off-axis form the Galapagos Spreading Center, and, also in this case, geophysical and geochemical anomalies along the ridge suggest plume–ridge interaction. These observations raise several questions concerning: (a) how plume material flows toward the ridge, opposite to the plate motion, (b) how plume material, once "trapped" beneath the ridge, flows parallel to the ridge axis, and (c) how the plume–ridge flow is affected by physical parameters such as the plume volume flux, the plume–ridge distance, and plate velocity. As reviewed by Ito et al. (2003), such questions have been

addressed with laboratory experiments (e.g., Kincaid et al., 1995; Feighner and Richards, 1995) and theoretical and numerical models (e.g., Ribe et al., 1995; Ribe, 1996; Ribe and Delattre, 1998) that progressively included more realistic conditions, such as the conductive thickening of the oceanic lithosphere away from the ridge and a migrating spreading ridge. The sloping base of the oceanic lithosphere favors the flow of plume material toward the ridge (Sleep, 1996), and the subsequent channeling of plume flow beneath the ridge has been calculated numerically by Albers and Christensen (2001). Such studies concluded that the width over which plume material flows along the ridge increases with higher plume flux and decreases with higher spreading rate and plume–ridge distance (see Ito et al., 2003; Ito and van Keken, 2007). Another important parameter is plume viscosity, in particular, Ito et al. (2003) find that the plume flow is reduced by the process of dehydration strengthening, which increases the viscosity of the plume residuum. For a fluid dynamical analysis of plume–ridge interaction see the comprehensive paper by Ribe et al. (2007) and the review paper by Ito et al. (2003).

Summary and Conclusion

This chapter on mantle plumes has taken us from the deep Earth's mantle where plumes originate, to sublithospheric depths where partial melting occurs, and up to the surface where LIPs have been emplaced in the geologic past, and hotspot lavas are being erupted today. We also explored the environmental effect of massive LIP volcanism on the oceans and the atmosphere. This range of topics shows that a full understanding of mantle plumes requires a multidisciplinary approach. Important progress will be achieved by linking geophysical, geochemical, and geological observations, to models solving for the fluid dynamics of mantle plumes, melt migration, and physical volcanology, just to mention a few.

Future directions

In our view, fundamental questions for the future include the following:

1. Quantifying the effect of heterogeneities on plume dynamics and fertility. The term heterogeneity is used here in a broad sense, from small-scale (1–10 km) lenses of recycled eclogite to the large-scale (100–1,000 km) denser superplumes that may generate short-lived instabilities rising from their top. Future studies should model heterogeneous plumes in a fully convecting mantle, with physical properties provided by mineral physics experiments at appropriate temperature and pressure conditions. Lithologic variability affects the fertility and the geochemical characteristics of lavas (Sobolev et al., 2005), but, at present, numerical simulations are unable to model the complex petrological and geochemical evolution of the solid matrix and melts.
2. Quantifying the role of the lithosphere is needed to understand how lithospheric stresses and variations in lithospheric thickness and composition may affect partial melting and the ascent of magma. Hieronymus and Bercovici (2000) provide a theory to explain the distribution of non-hotspot island and seamount chains in the South Pacific and conclude that volcanism can be activated by tensile stresses. The hypothesis that thermoelastic cracking of the oceanic lithosphere (Sandwell and Fialko, 2004) can trigger magmatism should be explored quantitatively, to assess the volumes and the composition of such magmas.
3. The cause-and-effect relation between the arrival of a mantle plume and continental breakup is often invoked, but we lack a quantitative understanding of the processes involved. In other words, although there is evidence supporting the role of plumes in the creation and modification of plate boundaries, it is unclear how a large-scale system, with continents and preexisting spreading ridges responds to the arrival of a large mantle plume. Moreover, studies of plume–lithosphere interaction should explore the possibility of dynamic instabilities, such as lithospheric delamination, fingering of low-viscosity fluid, and small-scale convection (see Ito and van Keken, 2007 and references therein), and their ability to generate partial melts.
4. Future studies should address in a constructive and quantitative way the current debate (see www.mantleplumes.org) on the "existence of plumes," or better, on the "coexistence of plumes with other processes" that can produce surface volcanism. Progress will be achieved through improved seismic imaging of plumes and their depth of origin, and through a quantitative understanding of lithospheric processes, as explained above. We also note that much of the ongoing debate questioning the existence of plumes is based on a narrow view of what a plume should look like: the head-tail model is valid in a purely thermal, homogeneous mantle, but a variety of plume shapes are found if the Earth's lower mantle is compositionally heterogeneous (Farnetani and Samuel, 2005; Kumagai et al., 2007). Furthermore, when considering plumes in a convecting mantle, rather than in an unperturbed fluid, it is obvious that conduits cannot be fixed (Steinberger et al., 2004), an observation that does not preclude the existence of deep plumes.
5. Although we have restricted our treatment to Mesozoic-Cenozoic plumes (<250 Ma), it is likely that plumes had an important role also at earlier times, although geological evidence becomes increasingly scant with geological age (Ernst and Buchan, 2001). The early Earth, with its vigorous and highly time-dependent convection, probably had an intense plume magmatism. Hopefully, studies of volcanism on Mars and Venus can complement and improve our knowledge of early terrestrial magmatism (see Condie, 2001, Chap. 3). Both Mars and Venus lack plate tectonics, so that old

geologic features can be better preserved. On Mars, the Olympus Mons is an enormous shield volcano (volume 2×10^6 km^3), and the Tharsis rise, which covers 20% of the Martian surface, has been volcanically active for the last 2 Ga, suggesting the presence of huge and long-lived plumes in the Martian mantle (Harderer and Christensen, 1996).

Bibliography

Albers, M., and Christensen, U. R., 2001. Channeling of plume flow beneath mid-ocean ridges. *Earth and Planetary Science Letters*, **187**, 207–220.

Ben-Avraham, Z., Nur, A., Jones, D., and Cox, A., 1981. Continental accretion: from oceanic plateaus to allochthonous terranes. *Science*, **213**, 47–54.

Boyet, M., and Carlson, R. W., 2005. Nd-142 evidence for early (>4.53 Ga) global differentiation of the silicate Earth. *Science*, **309**, 576–581.

Burke, K., and Torsvik, T. H., 2004. Derivation of large igneous provinces of the past 200 million years from long-term heterogeneities in the deep mantle. *Earth and Planetary Science Letters*, **227**, 531–538.

Christensen, U. R., and Hofmann, A. W., 1994. Segregation of subducted oceanic crust in the convecting mantle. *Journal of Geophysical Research*, **99**, 19867–19884.

Clouard, V., and Gerbault, M., 2008. Break-up spots: could the Pacific open as a consequence of plate kinematics? *Earth and Planetary Science Letters*, **265**, 195–208.

Coffin, M. F., and Eldholm, O., 1994. Large igneous provinces: crustal structure, dimensions and external consequences. *Reviews of Geophysics*, **32**, 1–36.

Condie, K. C., 2001. *Mantle Plumes and Their Record in Earth History*. Cambridge: Cambridge University Press.

Courtillot, V. E., and Renne, P., 2003. On the ages of flood basalt events. *Comptes Rendus Geoscience*, **335**, 113–140.

Courtillot, V. E., Jaupart, C., Manighetti, I., Tapponier, P., and Besse, J., 1999. On causal links between flood basalts and continental breakup. *Earth and Planetary Science Letters*, **166**, 177–195.

Courtillot, V. E., Davaille, A., Besse, J., and Stock, J., 2003. Three distinct types of hot spots in the Earth's mantle. *Earth and Planetary Science Letters*, **205**, 295–308.

Davaille, A., 1999. Simultaneous generation of hotspots and superswells by convection in a heterogeneous planetary mantle. *Nature*, **402**, 756–760.

Ebinger, C. J., and Sleep, N. H., 1998. Cenozoic magmatism throughout east Africa resulting from impact of a single plume. *Nature*, **395**, 788–791.

Ernst, R. E., and Buchan, K. L., 2001. The use of mafic dike swarms in identifying and locating mantle plumes. In Ernst, R. E., and Buchan, K. L. (eds.), *Mantle Plumes: Their Identification Through Time*. Colorado: Geological Society of America. Geological Society of America Special Paper, 352.

Farley, K. A., and Neroda, E., 1998. Noble gases in the Earth's mantle. *Annual Review of Earth and Planetary Sciences*, **26**, 189–218.

Farnetani, C. G., 1997. Excess temperature of mantle plumes: the role of chemical stratification across D″. *Geophysical Research Letters*, **24**, 1583–1586.

Farnetani, C. G., and Richards, M. A., 1994. Numerical investigation of the mantle plume initiation model for flood basalt events. *Journal of Geophysical Research*, **99**, 13813–13883.

Farnetani, C. G., and Samuel, H., 2005. Beyond the thermal plume paradigm. *Geophysical Research Letters*, **32**, L07311, doi:10.1029/2005GL022360.

Feighner, M. A., and Richards, M. A., 1995. The fluid dynamics of plume-ridge and plume-plate interactions: an experimental investigation. *Earth and Planetary Science Letters*, **129**, 171–182.

Foulger, G. R., and Natland, J. H., 2003. Is "Hotspot" volcanism a consequence of plate tectonics? *Science*, **300**, 921–922.

Ganino, C., and Arndt, N. T., 2009. Climate changes caused by degassing of sediments during the emplacement of large igneous provinces. *Geology*, **37**, 323–326, doi:10.1130/G25325A.1.

Greene, A. R., Scoates, J. S., and Weis, D., 2008. Wrangellia flood basalts in Alaska: a record of plume-lithosphere interaction in a Late Triassic accreted oceanic plateau. *Geochemistry, Geophysics, Geosystems*, **9**, Q12004, doi:10.1029/2008GC002092.

Griffiths, R. W., and Campbell, I. H., 1990. Stirring and structure in mantle starting plumes. *Earth and Planetary Science Letters*, **99**, 66–78.

Harderer, H., and Christensen, U. R., 1996. A one-plume model of Martian mantle convection. *Nature*, **380**, 507–509.

Hart, S. R., Schilling, J. G., and Powell, J. L., 1973. Basalts from Iceland and along the Reykjanes Ridge: Sr isotope geochemistry. *Nature*, **246**, 104–107.

Hieronymus, C. F., and Bercovici, D., 1999. Discrete alternating hotspot islands formed by interaction of magma transport and lithospheric flexure. *Nature*, **397**, 604–607.

Hieronymus, C. F., and Bercovici, D., 2000. Non-hotspot formation of volcanic chains: control of tectonic and flexural stresses on magma transport. *Earth and Planetary Science Letters*, **181**, 539–554.

Hofmann, A. W., 1997. Mantle geochemistry: the message from oceanic volcanism. *Nature*, **385**, 219–229.

Hofmann, A. W., 2003. Sampling mantle heterogeneity through oceanic basalts: isotopes and trace elements. In Carlson, R. W., Holland, H. D., and Turekian, K. K. (eds.), *Treatise on Geochemistry: The Mantle and Core*. Oxford: Elsevier, pp. 61–101.

Hofmann, A. W., and White, W. M., 1982. Mantle plumes from ancient oceanic crust. *Earth and Planetary Science Letters*, **57**, 421–436.

Ishii, M., and Tromp, J., 1999. Normal-mode and free-air gravity constraints on lateral variations in velocity and density of Earth's mantle. *Science*, **85**, 1231–1236.

Ito, G., and van Keken, P. E., 2007. Hotspots and melting anomalies. In Bercovici, D. (ed.), *Mantle Dynamics*. Amsterdam: Elsevier. Treatise on Geophysics, Vol. 7.

Ito, G., Lin, J., and Graham, D., 2003. Observational and theoretical studies of the dynamics of mantle plume–mid-ocean ridge interaction. *Reviews of Geophysics*, **41**, 1017, doi:10.1029/2002RG000117.

Jacobsen, S. B., and Wasserburg, G. J., 1979. The mean age of mantle and crustal reservoirs. *Journal of Geophysical Research*, **84**, 7411–7427.

Jellinek, A. M., and Manga, M., 2002. The influence of a chemical boundary layer on the fixity, spacing and lifetime of mantle plumes. *Nature*, **418**, 760–763.

Jellinek, A. M., and Manga, M., 2004. Links between long-lived hot spots, mantle plumes, D″ and plate tectonics. *Reviews of Geophysics*, **42**, RG3002, doi:10.1029/2003RG000144.

Jenkyns, H. C., 2010. Geochemistry of oceanic anoxic events. *Geochemistry, Geophysics, Geosystems*, **11**, Q03004, doi:10.1029/2009GC002788.

Kellogg, L. H., Hager, B. H., and van der Hilst, R. D., 1999. Compositional stratification in the deep mantle. *Science*, **283**, 1881–1884.

Kincaid, C., Ito, G., and Gable, C., 1995. Laboratory investigation of the interaction of off-axis mantle plumes and spreading centres. *Nature*, **367**, 758–761.

Kumagai, I., Davaille, A., and Kurita, K., 2007. On the fate of thermally buoyant mantle plumes at density interfaces. *Earth and Planetary Science Letters*, **254**, 180–193.

Labrosse, S., Hernlund, J. W., and Coltice, N., 2007. A crystallizing dense magma ocean at the base of the Earth's mantle. *Nature*, **450**, 866–869.

Li, X., Kind, R., Yuan, X., Wolbern, I., and Hanka, W., 2004. Rejuvenation of the lithosphere by the Hawaiian plume. *Nature*, **427**, 827–829.

Lithgow-Bertelloni, C., and Silver, P. G., 1998. Dynamic topography, plate driving forces and the African superswell. *Nature*, **395**, 269–272.

Loper, D. E., and Stacey, F. D., 1983. The dynamical and thermal structure of deep mantle plumes. *Physics of the Earth and Planetary Interiors*, **33**, 305–317.

Masters, G., Laske, G., Bolton, H., and Dziewonski, A. M., 2000. The relative behavior of shear velocity, bulk sound speed and compressional velocity in the mantle: implication for chemical and thermal structure, in Earth's deep interior. In Karato, S. I., Forte, A., Liebermann, R. C., Masters, G., and Stixrude, L. (eds.), *Mineral Physics and Tomography from the Atomic to the Global Scale*. Washington, DC: American Geophysical Union, Vol. 117, pp. 63–87.

McKenzie, D., and Bickle, M. J., 1988. The volume and composition of melt generated by extension of the lithosphere. *Journal of Petrology*, **29**, 625–679.

Montelli, R., Nolet, G., Masters, G., Engdahl, E. R., and Hung, S.-H., 2004. Finite-frequency tomography reveals a variety of plumes in the mantle. *Science*, **303**, 338–343.

Moore, W. B., Schubert, G., and Tackley, P., 1998. Three-dimensional simulations of plume-lithosphere interaction at the Hawaiian swell. *Science*, **279**, 1008–1011.

Morgan, W. J., 1971. Convection plumes in the lower mantle. *Nature*, **230**, 42–43.

Olson, P., and Nam, I.-S., 1986. Formation of seafloor swells by mantle plumes. *Journal of Geophysical Research*, **91**, 7181–7191.

Olson, P., Schubert, G., and Anderson, C., 1993. Structure of axisymmetric mantle plumes. *Journal of Geophysical Research*, **98**, 6829–6844.

Parmenteir, E. M., Turcotte, D. L., and Torrance, K. E., 1975. Numerical experiments on the structure of mantle plumes. *Journal of Geophysical Research*, **80**, 4417–4424.

Putirka, K. D., 2005. Mantle potential temperatures at Hawaii, Iceland, and the mid-ocean ridge system, as inferred from olivine phenocrysts: evidence for thermally driven mantle plumes. *Geochemistry Geophysics Geosystems*, **6**, Q05L08, doi:10.1029/2005GC000915.

Renne, P. R., Zhang, Z. C., Richards, M. A., Black, M. T., and Basu, A. R., 1995. Synchrony and causal relations between Permian-Triassic boundary crises and Siberian flood volcanism. *Science*, **269**, 1413–1416.

Ribe, N. M., 1996. The dynamics of plume-ridge interaction 2. Off-ridge plumes. *Journal of Geophysical Research*, **101**, 16195–16204.

Ribe, N. M., and Delattre, W. L., 1998. The dynamics of plume-ridge interaction – III. The effects of ridge migration. *Geophysical Journal International*, **133**, 511–518.

Ribe, N. M., Christensen, U. R., and Theissing, J., 1995. The dynamics of plume-ridge interaction, 1: ridge-centered plumes. *Earth and Planetary Science Letters*, **134**, 155–168.

Ribe, N. M., Davaille, A., and Christensen, U. R., 2007. Fluid dynamics of mantle plumes. In Ritter, J. R. R., and Christensen, U. R. (eds.), *Mantle Plumes – A Multidisciplinary Approach*. Berlin: Springer, pp. 1–48.

Richards, M. A., Duncan, R. A., and Courtillot, V. E., 1989. Flood basalts and hotspot tracks: plume heads and tails. *Science*, **246**, 103–107.

Richards, M. A., Jones, D. L., Duncan, R. A., and DePaolo, D. J., 1991. A mantle plume initiation model for the Wrangellia flood basalt and other oceanic plateaus. *Science*, **254**, 263–267.

Romanowicz, B., and Gung, Y. C., 2002. Mega-upwellings from the core–mantle boundary to the lithosphere: implications for heat flux. *Science*, **296**, 513–516.

Sandwell, D. T., and Fialko, Y., 2004. Warping and cracking of the Pacific plate by thermal contraction. *Journal of Geophysical Research*, **109**, B10411, doi:10.1029/2004JB003091.

Schilling, J. G., 1973. Iceland mantle plume: geochemical evidence along Reykjanes Ridge. *Nature*, **242**, 565–571.

Schilling, J. G., Thompson, G., Kingsley, R., and Humphris, S., 1985. Hotspot-migrating ridge interaction in the South Atlantic. *Nature*, **313**, 187–191.

Schubert, G., Turcotte, D. L., and Olson, P., 2001. *Mantle Convection in the Earth and Planets*. Cambridge: Cambridge University Press.

Schulte, P., et al., 2010. The Chicxulub asteroid impact and mass extinction at the Cretaceous-Paleogene boundary. *Science*, **327**, 1214–1218.

Self, S., Widdowson, M., Thordarson, T., and Jay, A. E., 2006. Volatile fluxes during flood basalt eruptions and potential effects on the global environment: a Deccan perspective. *Earth and Planetary Science Letters*, **248**, 518–532.

Sleep, N. H., 1990. Hotspot and mantle plumes: some phenomenology. *Journal of Geophysical Research*, **95**, 6715–6736.

Sleep, N. H., 1996. Lateral flow of hot plume material ponded at sublithospheric depths. *Journal of Geophysical Research*, **101**, 28065–28083.

Smith, W. H. F., and Sandwell, D. T., 1997. Global sea floor topography from satellite altimetry and ship depth soundings. *Science*, **277**, 1956–1962.

Sobolev, A. V., Hofmann, A. W., Sobolev, S. V., and Nikogosian, I. K., 2005. An olivine-free mantle source of Hawaiian shield basalts. *Nature*, **434**, 590–597.

Sobolev, A. V., Hofmann, A. W., et al., 2007. The amount of recycled crust in sources of mantle-derived melts. *Science*, **316**, 412–417.

Steinberger, B., Sutherland, R., and O'Connell, R. J., 2004. Prediction of Emperor-Hawaii seamount locations from a revised model of global plate motion and mantle flow. *Nature*, **430**, 167–173.

Svensen, H., Planke, S., Malthe-Sorenssen, A., Jamtveit, B., Myklebust, R., Eidem, T. R., and Rey, S. S., 2004. Release of methane from a volcanic basin as a mechanism for initial Eocene global warming. *Nature*, **429**, 542–545.

Svensen, H., Planke, S., et al., 2009. Siberian gas venting and the end-Permian environmental crisis. *Earth and Planetary Science Letters*, **277**, 490–500.

Tackley, P. J., 1998. Three-dimensional simulations of mantle convection with a thermo-chemical basal boundary layer: D″? In Gurnis, M., et al. (eds.), *The Core–Mantle Boundary Region*. Washington, DC: American Geophysical Union. Geophysical Monograph Series, Vol. 28, pp. 231–253.

Tarduno, J., Bunge, H.-P., Sleep, N., and Hansen, U., 2009. The bent Hawaiian-Emperor hotspot track: inheriting the mantle wind. *Science*, **324**, 50–53.

Tolstikhin, I., and Hofmann, A. W., 2005. Early crust on top of the Earth's core. *Physics of the Earth and Planetary Interiors*, **148**, 109–130.

van der Hilst, R., Widiyantoro, S., and Engdahl, E., 1997. Evidence for deep mantle circulation from global tomography. *Nature*, **386**, 578–584.

Wessel, P., 1993. Observational constraints on models of the Hawaiian hot spot swell. *Journal of Geophysical Research*, **98**, 16095–16104.

White, R. S., and McKenzie, D. P., 1989. Magmatism at rift zones: the generation of volcanic continental margins and flood basalts. *Journal of Geophysical Research*, **94**, 7685–7729.

Whitehead, J. A., and Luther, D. S., 1975. Dynamics of laboratory diapir and plume models. *Journal of Geophysical Research*, **80**, 705–717.
Wilson, J. T., 1963. A possible origin of the Hawaiian Islands. *Canadian Journal of Physics*, **41**, 863–870.
Wolfe, C. J., Solomon, S. C., Laske, G., Collins, J. A., Detrick, R. S., Orcutt, J. A., Bercovici, D., and Hauri, E. H., 2009. Mantle shear-wave velocity structure beneath the Hawaiian hot spot. *Science*, **326**, 1388–1390.

Cross-references

Core-Mantle Coupling
Mantle Convection
Mantle D″ Layer
Mantle Viscosity

MANTLE VISCOSITY

W. R. Peltier
Department of Physics, University of Toronto, Toronto, ON, Canada

Definition

By the term "mantle rheology" we mean to imply the nature of the relationship between applied stress and realized strain.

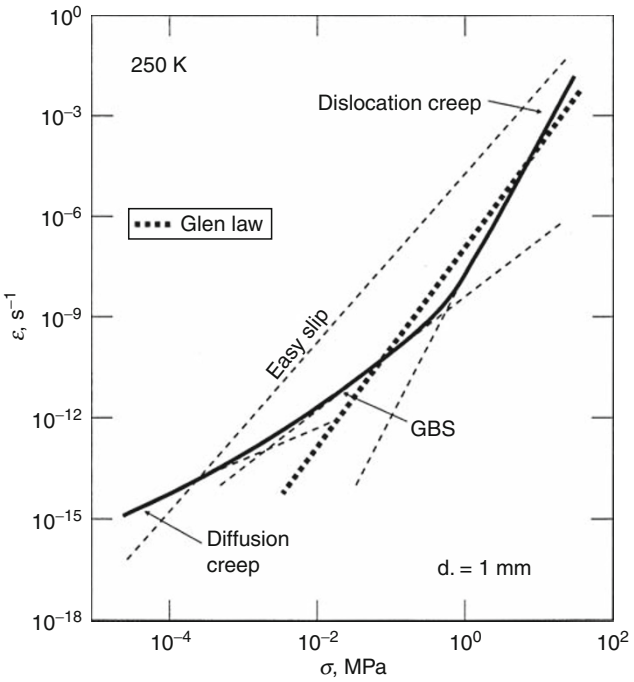
Introduction

The effective viscosity of the Earth's predominantly iron-magnesium silicate mantle is a physical property of fundamental importance. In characterizing this outer shell of the planet by a viscosity at all, one is recognizing that it is capable of flowing as if it were a fluid. This is in spite of the fact that on the timescales characteristic of seismic waves it is known to be accurately describable as a Hookean elastic solid, albeit one which is endowed with a finite quality factor "Q" due to imperfections of elasticity. The idea that viscoelastic materials behave as solids on short timescales but as fluids on sufficiently long timescales is well understood in the material science literature. For Earth sciences, the fact that the mantle behaves viscously on timescales in excess of a few centuries is vital to the understanding of plate tectonics and the origins of continental drift as a consequence of thermal convection in the mantle. This article provides a summary of laboratory and geophysical data that may be brought to bear as a means of measuring mantle viscosity and discusses the compatibility of these measurements with the requirements of the mantle convection process.

Methods of measurement

Laboratory experiments

One might imagine that laboratory measurements of the creep of mantle mineral assemblages at high pressure and temperature would provide all that is required to obtain the data needed to infer effective viscosity. There are several problems with this idea, however, most of which derive from the difficulty of performing the appropriate experiments under the extreme conditions characteristic of the Earth's deep interior. There nevertheless does exist very detailed information on the creep of polycrystalline solids in general which may provide useful insight on the sort of behavior to be expected. Since such polycrystalline aggregates have many characteristics in common insofar as their ability to "flow" is concerned, it is instructive to consider one specific example in order to fix ideas. To this end Figure 1 displays a strain-rate vs. differential stress diagram for fine-grained ice (from Goldsby and Kohlstedt, 1997), an important model system that has been studied in great detail because of its practical importance in glaciology. In this log-log plot, a straight line is indicative of a power law relationship with the slope determining the "stress exponent." Inspection of the figure will show that, as the differential stress is reduced, the flow law undergoes a series of discrete transitions. At the highest values of differential stress, the stress exponent has a value of 4, indicating that the mechanism supporting the "flow" of the polycrystalline aggregate involves the propagation of "dislocations" through the interiors of the individual crystal grains that make up the aggregate. This process is referred to as involving "dislocation climb." As the differential stress diminishes, the stress exponent first drops to a value near 2, indicating that the mechanism supporting flow has come to involve the propagation of dislocations along the boundaries separating individual



Mantle Viscosity, Figure 1 The rheology of fine-grained (d = 1 mm) ice at 250 K temperature, calculated using a particular constitutive equation (Goldsby and Kohlstedt, 1997).

grains. This regime is referred to as involving "grain boundary sliding." With even further reduction of the differential stress, the exponent drops to unity implying that the underlying mechanism involves the diffusion of impurities. In this regime, the stress–strain relation describing the flow of the solid is now equivalent to that which holds for a simple fluid like water which we refer to as being "Newtonian" Investigation of the temperature dependence of the experimentally determined stress-strain relation shows that, as the temperature approaches the melting temperature the strain-rate increases for a fixed value of the differential stress.

Non-Newtonian rheology from seismic anisotropy

It is generally understood that at shallow depths, from the surface through the near-surface "lithosphere" in which the effective viscosity is extremely high, the operative rheology is significantly non-Newtonian. The extent to which this behavior persists to greater depth remains undetermined. That the shallow Earth deforms according to a non-Newtonian stress–strain relation follows from the existence of significant seismic anisotropy. The development of such anisotropy requires nonlinearity of the stress–strain relationship in order that individual crystal grains may be aligned by the ambient shear. Although direct laboratory measurements on the creep of Earth materials at lithospheric depths where the pressures and temperatures are low provide useful confirmation of the seismic observations, they are much less illuminating for the Earth's deep mantle because of the extreme conditions under which measurements must be made.

Natural experiments

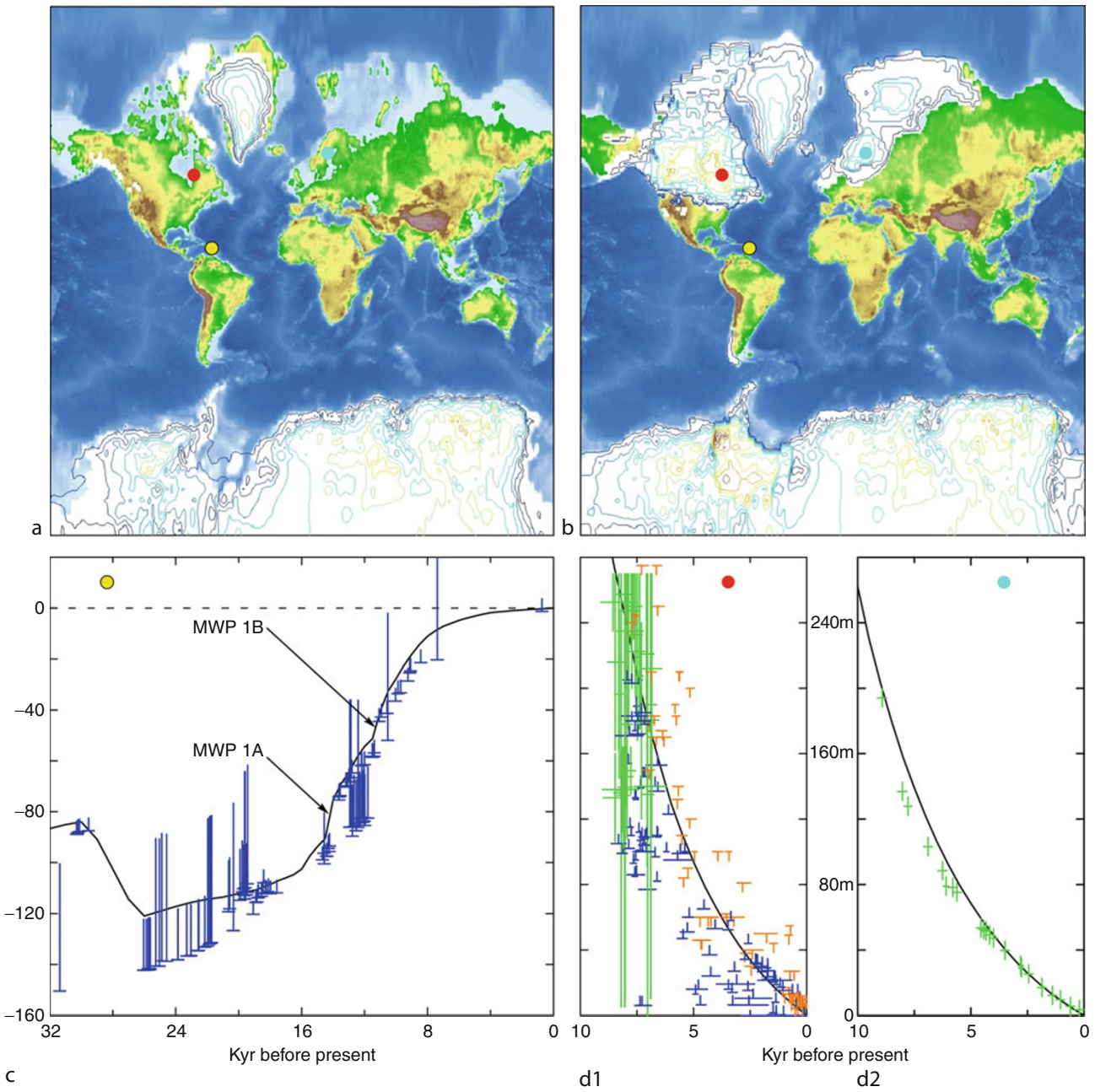
In order to obtain useful estimates of the viscosity at depths below the lithosphere we are obliged to employ the results of "natural experiments" that the Earth itself has performed throughout its history and for which it has also conveniently remembered the results through the fidelity of the geological recording system. The most important such experiment is that associated with the Late Quaternary ice-ages, especially the quasi-periodic sequence of glaciation-deglaciation events that began about 900,000 years before present and during which massive accumulations of land ice began to appear and disappear from the continents, especially at high northern latitudes. Insofar as deep Earth rheology is concerned, these events may be thought of as comprising an experiment during which extreme normal stresses were applied to localized regions of the Earth's surface and thereafter removed. Because the residence time of the land ice on the surface of the continents was sufficiently long, the response to the applied normal stresses was such as to induce a response of the loaded viscoelastic mantle that involved significant viscous flow. This suggests the possibility that the geological data that represent the Earth's record of its response to the time-varying ice load might be interpreted to infer the value of the viscosity that governs the viscous component of the response. Also of interest is the question as to whether the viscosity so inferred is compatible with the requirements of the mantle convection explanation of surface kinematic observations related to the process of plate tectonics.

Mantle viscosity from glacial isostatic adjustment

Observational data sets

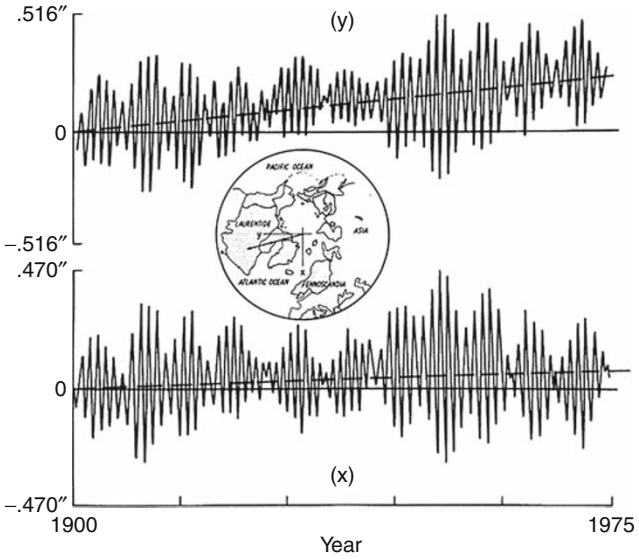
Figure 2 shows examples of critical data related to the Glacial Isostatic Adjustment (GIA) process along with a comparison of maps of the ice-covered regions of the planet under both modern conditions (a), and the conditions that prevailed at the last glacial maximum (LGM) approximately 21,000 years before present (b). Figure 2c shows an extremely important data set for the history of sea level change relative to the surface of the solid Earth at the island of Barbados (Peltier and Fairbanks, 2006). This record covers the interval of time from LGM to the present and has been produced by the uranium/thorium dating of corals, the species Acropora Palmata being the most important as it inevitably dwells within 5 m of the surface of the sea while it remains an open system. These data correlate well with the record of ice-equivalent eustatic sea level history and demonstrate that sea level has risen at Barbados by approximately 120 m since LGM. The Barbados data provide a strong constraint on the total mass of ice that must have been removed from the land in order to cause the rise of sea level observed. In Figure 2 (d1,d2) are shown observations of relative sea level history from sites that were located near the center of the regions in North America and Eurasia respectively that were covered by glacial ice at LGM (Peltier, 1998). Whereas the Barbados data record the extent to which sea level was depressed at LGM, these data from the previously ice-covered regions demonstrate that subsequent to the removal of the ice, the land in these regions has been rising out of the sea. This is the process referred to as postglacial rebound of the crust. The data shown in Figure 2 (d1,d2) may be quantitatively analyzed (Peltier, 1998) to extract information about the amplitude and the relaxation time of the individual sea level cuves and thus the viscosity of the Earth's mantle. The relaxation time data depend almost entirely on the mantle viscosity. Data such as these from hundreds of locations constitute one of the primary data sets that may be invoked to constrain the viscosity model of the Earth's mantle.

The remaining data related to the GIA process that have come to play an increasingly important role in such analyses are those related to the Earth's rotational state. Figure 3 illustrates the first of these additional constraints. It shows the time-dependent position of the Earth's north pole of rotation in a coordinate system with x- and y-axes centered upon the Conventional International Origin (Vincente and Yumi, 1969, 1970). Inspection of the time series in the period between 1900 and 1979, at which time the International Latitude Service, which was responsible for maintaining these measurements, was discontinued,



Mantle Viscosity, Figure 2 (**a**) Global ice cover and surface topography and bathymetry under modern climate conditions, (**b**) same as (**a**) but for Last Glacial Maximum conditions according to the new ICE-6G model. (**c**) A fit of the ICE-6G (VM5a) model to the coral based record of relative sea level history from the island of Barbados. The two major melt water pulses (MWP) are marked. (**d**) VM5a model fit to the records of Holocene relative sea level history at sites near the centers of rebound of the Laurentide (**d1**) and Fennoscandian (**d2**) ice sheets.

shows that the signals are dominated by "beating" at a period very close to 7 years. This is associated with the interference between the 12-month periodic annual wobble of the spin-axis forced by the seasonal variations of the atmosphere and oceans and the 14-month periodic Chandler wobble. The component of these observations that is of interest from the perspective of mantle viscosity measurement, however, consists of the slow secular drift of the pole upon which the oscillatory component is superimposed. These data reveal the secular drift of the pole at a rate near 0.95°/m.y. approximately along the 79°W meridian, i.e., in the direction of present day Hudson Bay. The rate of this "true polar wander" provides a further constraint upon the viscosity of the Earth's mantle.



Mantle Viscosity, Figure 3 Time series for x- and y-components of the location of the north pole of rotation of the planet relative to the coordinate system shown in the inset.

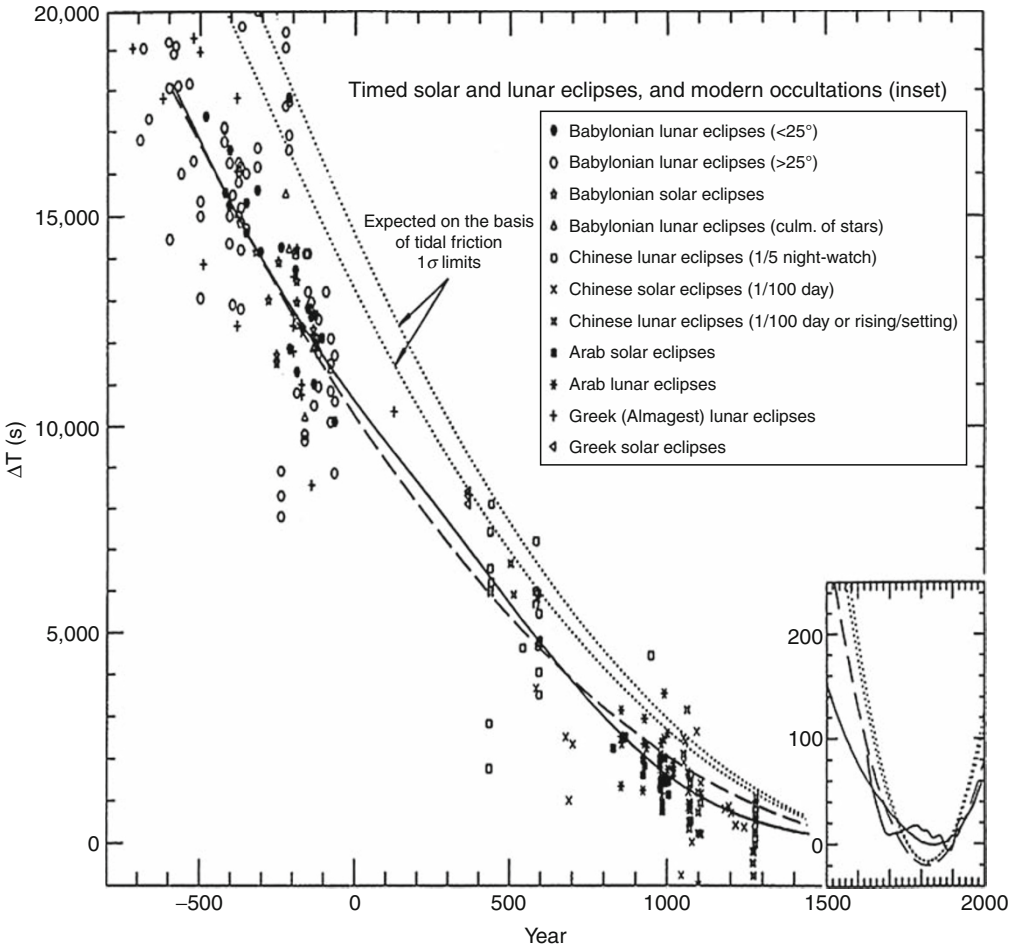
The final observational datum that may contribute to the estimation of mantle viscosity consists of the "nontidal acceleration" of planetary rotation. This datum is illustrated in Figure 4 where specific times of occurrence of total eclipses of the Sun and the Moon are plotted as a function of time spanning the past 2,000 years of the Earth's history covering the periods in both Babylon and China during which naked eye astronomers both recorded, and attempted to make predictions of, the timing of these "epochal" events (from Stephenson and Morrison, 1995). These observations are compared with expectations described by two essentially parabolic curves. The curve that does not go through the data is based on the assumption that only tidal friction has contributed to variations in the Earth's rate of rotation over this period and that this has remained equal to its present day value, a reasonable assumption. Because the predicted times of occurrence of total eclipses of the Sun and the Moon do not agree with expectations based on this assumption, some physical process is missing. The nature of the misfit of this model to the data implies that the action of the missing process is such as to somewhat counteract the influence of tidal friction by decreasing the rate at which angular velocity would otherwise decrease. This is called the nontidal acceleration of rotation and it too is a consequence of ice-age loading and unloading of the continents. The curve in Figure 4 that does go through the data provides a precise measurement of this ice-age influence and this may also be employed to constrain the viscosity of the Earth's mantle.

Results of formal inversion

Models of mantle viscosity may be obtained by formal inversion of the totality of the data related to the Late Quaternary ice-age cycle. It is important to note that each of these data samples the viscosity structure of the deep Earth in a different fashion. The postglacial rebound data from the Fennoscandia region, which was covered at LGM by an ice sheet of moderate horizontal scale, provide information on the viscosity of the upper mantle and transition zone. Data from North America, on the other hand, where the Canadian land mass was entirely covered by the vast Laurentide ice-sheet complex at LGM provide information from depths that extend well into the upper part of the lower mantle. Below a depth of approximately 1,200 km, only Earth rotation data provide any constraint upon the viscosity of this deepest region that extends to the depth of the CMB (Peltier, 1998).

Figure 5 shows a series of the earliest inversion results (Peltier, 1996) based on the application of a formal Bayesian methodology to the simultaneous inversion of all of the data types. For the purpose of these analyses, an initial guess profile (VM1) was assumed and the data were inverted to deduce a best fitting viscosity model. In the model, a 120-km-thick perfectly elastic lithosphere was included on the top of the viscous mantle and so viscosity was inferred as a function of depth only below this level. That such an elastic surface layer should exist is rational on a priori grounds due to the fact that interior temperatures in the near-surface region are so low that viscosity is effectively infinite, thus rendering this layer perfectly elastic. Use of rebound data to provide a direct measurement of lithospheric thickness requires recourse to the incorporation of additional data from the rebound process induced by the removal of ice sheets of sufficiently small lateral extent. Peltier et al. (2002) employed data associated with the rebound of Scotland following removal of the Scottish ice sheet that was in place over this region at LGM for this purpose. Peltier and Drummond (2008) on the other hand employed horizontal motion data from locations outboard of the Laurentide ice sheet for the same purpose. The model labeled VM2 in Figure 5 was obtained from a Bayesian inversion from which the true polar wander data were held back so as to provide the basis for a check on the quality of the results for the deepest mantle that the analysis delivered.

Sensitivity analyses performed by adjusting the magnitude of the observed rate of nontidal acceleration to mimic the influence of a contamination of this observation by the modern melting of land ice from the Polar Regions reveal that the viscosity of the lower mantle is strongly sensitive to this parameter. The model obtained by assuming that polar ice is melting at a rate sufficient to raise global sea level by 1.5 mm/year is the model labeled VM3 in Figure 5. The viscosity in the region shallower than ~1,250 km depth remains "pinned" to the previously determined depth dependence by the relaxation time constraints derived from observations of the rebound of the crust in both North American and Northwestern Eurasia. This clearly implies that below the depth of ~1,250 km, it is only the Earth rotation constraints that deliver information on mantle viscosity.

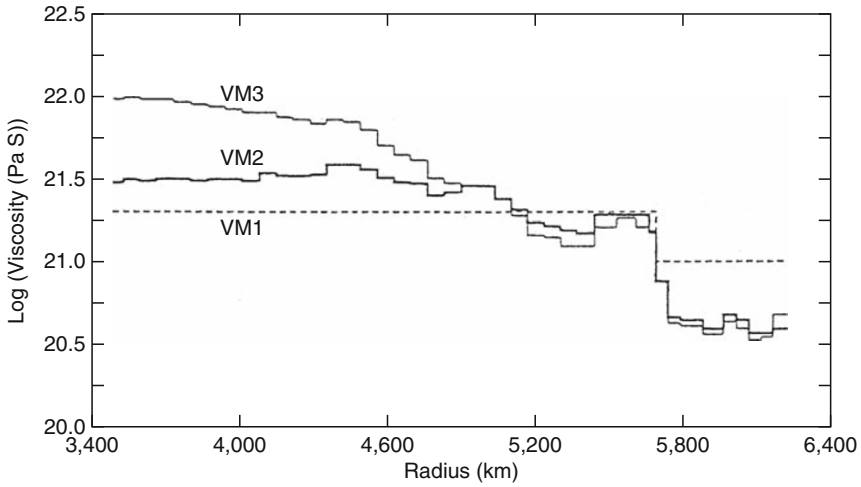


Mantle Viscosity, Figure 4 Plot showing variation of time difference between the observed time of a total eclipse of the Sun or the Moon and the predicted time of the event based on the assumption of constant tidal torque over the past 2,000 years. The inset (on a scale 25 times expanded) includes the continuous curve derived from the lunar occultations in the period 1620–1995.5 AD. The dashed curve is the best fitting parabola and the solid curve is fitted using cubic splines. (From Stephenson and Morrison, 1995.)

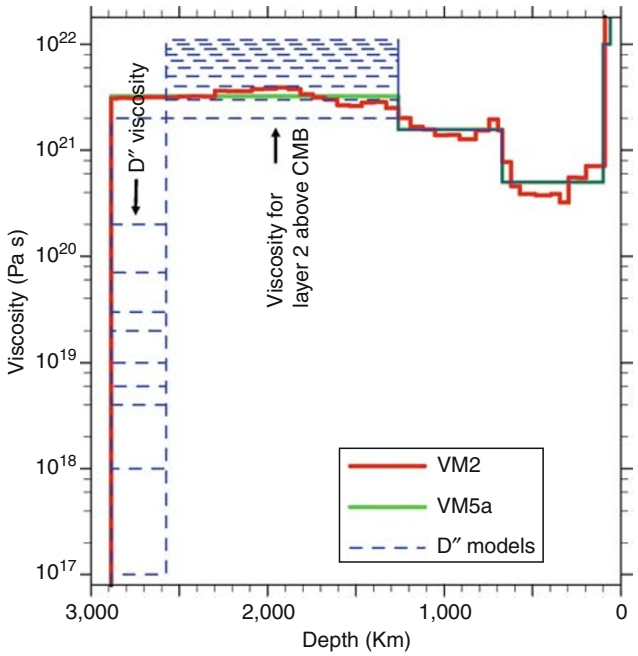
This sets the stage for a focused attempt to measure any variation of viscosity that might exist in the deepest Earth. To this end, Figure 6 illustrates a two-layer parameterization of the mantle below the critical depth of ∼1,250 km in which a multilayer parameterization of VM2, called VM5a, is perturbed by varying the viscosity of the two lowermost layers; the D" layer assumed to be 300-km thick, and the lower mantle layer (layer-2) extending from 1,250 km depth to the top of D". A sequence of such models can be fitted to both the nontidal acceleration and polar wander speed. It is found that for each value assumed for D" viscosity, it is possible to find a corresponding value for the viscosity of layer-2, which enables the model to fit both rotational observables. Figure 7a provides a summary diagram illustrating the manner in which the viscosity of D" trades off against the viscosity of layer-2 (Peltier and Drummond, 2010). As the viscosity of D" decreases, the viscosity of layer-2 first increases as expected in order to recover the fit to the observations. However, once the viscosity of D" is assumed to be sufficiently low, as to be behaving as though it were inviscid, the allowed value of the viscosity of the overlying layer saturates as would be expected on physical grounds. Figure 7b shows the manner in which the prediction of polar wander direction varies for the set of models that provide satisfactory fits to both of the rotational constraints. Since it is the ILS data that correctly capture the direction of true polar wander associated with ancient ice-age influence, it will be clear that the preferred model is one for which the viscosity of D" is only modestly reduced relative to a viscosity of the overlying layer that is itself only modestly enhanced relative to that in the same region of VM5a.

Compatibility of the GIA-based viscosity model with that required by models of the mantle convection process

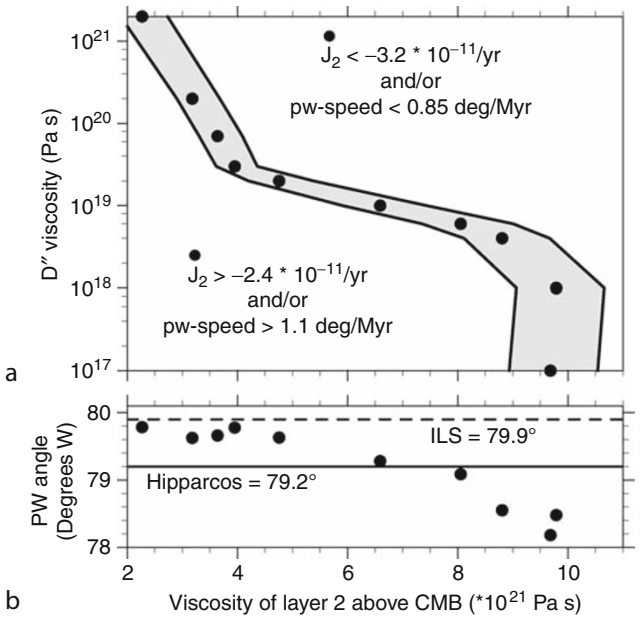
Early efforts to determine the viscosity of the Earth's mantle based on GIA data (e.g., Walcott, 1970) appealed to the



Mantle Viscosity, Figure 5 Models of the mantle viscosity as a function of depth obtained by formal Bayesian inversion of the glacial isostatic adjustment data.

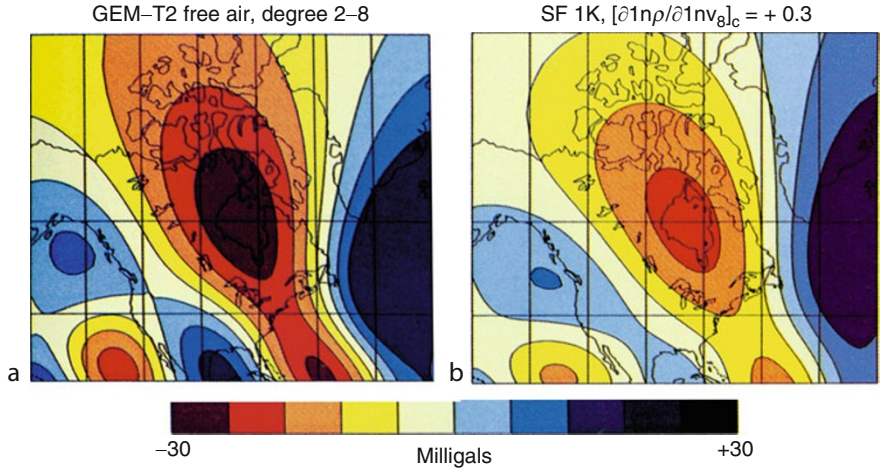


Mantle Viscosity, Figure 6 A family of mantle viscosity models for the D″ layer and the layer-2 above CMB. In these models, the lithosphere consists of two distinct units. The upper 60-km-thick unit is assumed as perfectly elastic whereas the lower 40-km-thick unit has a viscosity of 10^{22} Pa.s.



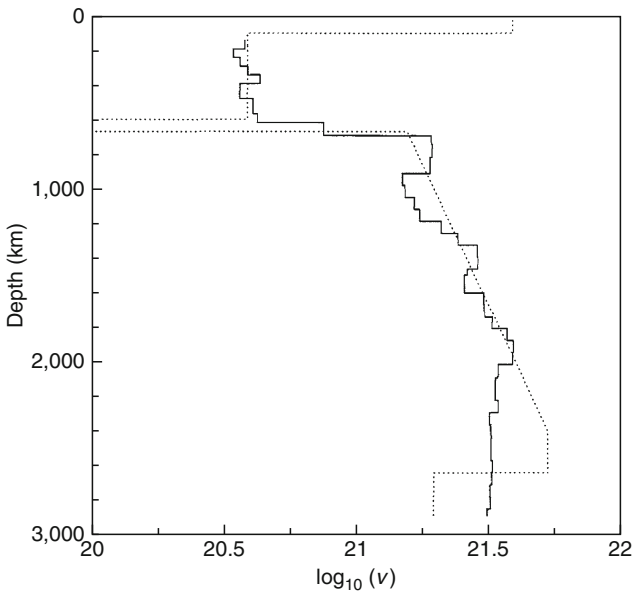
Mantle Viscosity, Figure 7 (**a**) A trade-off diagram describing the relationship between the viscosity assumed for the D″ layer and that of the layer-2 that is required in order that both observables are fit by the same model of the GIA process. (**b**) The predicted direction of polar wander for the above models. Also shown is the observed direction consistent with the International Latitude Service (ILS) data set.

observed magnitude of the free-air gravity anomaly over the Hudson Bay region to argue that the magnitude of this anomaly provided strong evidence that the viscosity of the deep Earth was much higher than implied by the analyses described in the previous section. This anomaly, based on the GEM-T2 model of the Goddard Space Flight Center, is shown in Figure 8a. The anomaly has a magnitude of approximately −30 milligals when the field is restricted to include only spherical harmonic degrees 2 through 8. When the VM2 viscosity model derived by GIA analysis was used in conjunction with an acceptably accurate model of ice thickness history, the predicted free-air



Mantle Viscosity, Figure 8 (**a**) Observed GEM-T2 degree 2–8 non-hydrostatic free-air gravity anomaly over the Hudson Bay region of Canada, and (**b**) the calculated anomaly based on the predictions of the internal loading theory for the density conversion factor for the cratonic root of +0.3.

gravity anomaly had peak amplitude of only −3 milligals. This issue concerning the origin of the free-air gravity anomaly, which is well correlated in its location with that of the LGM Laurentide ice sheet, has been analyzed in detail by Pari and Peltier (1995, 1996, 1998, 2000). Their work focused on the viability of the alternative hypothesis that the anomaly was supported by the mantle convection process. Models of mantle convection incorporating the results of seismic tomography (Forte et al., 1994) and based on internal loading theory were developed and applied in analyses of the non-hydrostatic geoid and free-air gravity anomalies that would be supported by this mantle dynamic process. In the context of internal loading theory for an Earth model with spherically symmetric viscosity distribution, there are only two distinct parameters that need to be specified. The first one is the depth-dependent mapping factor through which the heterogeneity in seismic shear wave velocity is converted into lateral heterogeneity of density and the second is the depth dependence of mantle viscosity. Results were obtained for three different values of the S-wave velocity to density conversion factor in the subcontinental "tectosphere"; −0.3 (positive buoyancy, upwelling flow), 0.0 (neutral buoyancy), and +0.3 (negative buoyancy, downwelling flow). Of the three scenarios, the last model has provided an excellent fit to the free-air gravity anomaly (Figure 8b) and 24 milligals of the 30 milligal free-air anomaly over Hudson Bay could be explained by the mantle convection hypothesis. It needs to be understood that there has been no attempt to tune the model to achieve a best fit locally. Rather the model has been optimized to achieve a best fit to the entire global gravity field. It seems clear on the basis of these results that the free-air gravity anomaly over the Hudson Bay region is essentially entirely due to the mantle convection process. Therefore, the original argument of Walcott (1970) that this anomaly must be due to remaining glacial isostatic disequilibrium and to imply that the viscosity of the mantle of the Earth must be extremely high, is not supported. It seems rather to be explicable as a consequence of thermal convection in a mantle through which the variation of viscosity is rather modest.



Mantle Viscosity, Figure 9 Comparison of the VM2 viscosity profile (*solid line*) to the viscosity profile (*dotted line*) determined on the basis of an optimal fit to the non-hydrostatic free-air gravity anomaly.

The viscosity profile that delivers the high-quality fit to the free-air anomaly of the North American continent is shown in Figure 9 by the dotted line along where it is compared to the GIA-derived model VM2. There are two

primary differences between these profiles which are otherwise extremely close to one another. The first is the structure near the base of the mantle. Evident is the presence of the low viscosity region adjacent to the CMB above which viscosity is enhanced above the level in the VM2 model. The second difference appears in the form of a very soft thin layer coincident with the 660 km discontinuity corresponding to the endothermic phase transformation from Spinel to a mixture of Perovskite and Magnesiowustite. This phase transition acts as a significant, though imperfect, impediment to convective mixing. When mixing does occur across the boundary, however, a significant change in grain size is expected, which leads to the transformational super-plastic softening of the mineral assemblage in the near vicinity of the boundary. An outstanding issue in mantle dynamics concerns the dynamical impact of this feature on the mixing process.

Summary

This article has reviewed the main sources of information on the viscosity of the deep Earth. These sources consist of data pertaining to the phenomena of postglacial rebound and mantle convection. In spite of modest differences between the models of mantle viscosity depth dependence derived in distinct ways, the overwhelming evidence is that the depth dependence of viscosity derived on the basis of GIA observations is in close accord with that required by the mantle convection hypothesis explanation of surface plate tectonics and continental drift.

Bibliography

Forte, A. M., Woodward, R. L., and Dziewonski, A. M., 1994. Joint inversions of seismic and geodynamic data for models of three-dimensional mantle heterogeneity. *Journal of Geophysical Research*, **99**, 21857–21877.

Goldsby, D. L., and Kohlstedt, D. L., 1997. Grain boundary sliding in fine grained ice I. *Scripta Materialia*, **37**, 1399–1406.

Pari, G., and Peltier, W. R., 1995. The heat flow constraint on mantle tomography based convection models: Towards a geodynamically self-consistent inference of mantle viscosity. *Journal of Geophysical Research*, **100**, 12731–12751.

Pari, G., and Peltier, W. R., 1996. The free-air gravity constraint on sub-continental mantle dynamics. *Journal of Geophysical Research*, **101**, 28105–28132.

Pari, G., and Peltier, W. R., 1998. Global surface heat flux anomalies frm seismic tomography based models of mantle flow: Implications for mantle convection. *Journal of Geophysical Research*, **103**, 23743–23780.

Pari, G., and Peltier, W. R., 2000. Subcontinental mantle dynamics: A further analysis based upon the joint constraints of dynamic surface topography and free-air gravity. *Journal of Geophysical Research*, **105**, 5635–5662.

Peltier, W. R., 1996. Mantle viscosity and ice-age ice-sheet topography. *Science*, **273**, 1359–1364.

Peltier, W. R., 1998. Postglacial variations in the level of the sea: Implications for climate dynamics and solid earth geophysics. *Reviews of Geophysics*, **36**, 603–689.

Peltier, W. R., and Drummond, R., 2008. The rheological stratification of the lithosphere: A direct inference based upon the geodetically observed pattern of the glacial isostatic adjustment of the North American continent. *Geophysical Research Letters*, **35**, L16314. 10.1029/2008GL034586.

Peltier, W. R., and Drummond, R., 2010. Deepest mantle viscosity: Constraints from earth rotation anomalies. Geophysysical Research Letters, **37**, L12304,20190.

Peltier, W. R., and Fairbanks, R. G., 2006. Global glacial ice volume and Last glacial maximum duration from an extended Barbados sea level record. *Quaternary Science Reviews*, **25**, 3322–3337.

Peltier, W. R., Shennan, I., Drummond, R., and Horton, B., 2002. On the postglacial isostatic adjustment of the British Isles and the shallow viscoelastic structure of the earth. *Geophysical Journal International*, **148**, 443–475.

Stephenson, E. R., and Morrison, L. V., 1995. Long term fluctuations in the earth's rotation: 700 BC to AD 1990. *Philosophical Transactions of the Royal Society of London. Series A*, **351**, 165–202.

Vincente, R. O., and Yumi, S., 1969. Co-ordinates of the pole (1899–1968) returned to the conventional international origin. *Publications of the International Latitude Observatory of Mizusawa*, **7**, 41–50.

Vincente, R. O., and Yumi, S., 1970. Revised values (1941-1961) of the co-ordinates of the pole referred to the CIO. *Publications of the International Latitude Observatory of Mizusawa*, **7**, 109–112.

Walcott, R. I., 1970. Isostatic response to loading of the crust in Canada. *Canadian Journal of Earth Sciences*, **7**, 716–727.

Cross-references

Mantle Convection
Mantle D″ Layer
Seismic Anisotropy
Seismic Tomography

N

NUMERICAL METHODS, BOUNDARY ELEMENT

Michele Cooke
Geosciences Department, UMass – Amherst, Amherst, MA, USA

Definition

The boundary element method is a numerical technique for investigating deformation within the Earth. The method is particularly adept at investigations of discontinuities in the crust, such as cracks, faults, and dikes. The numerical method is able to analyze complex problems by discretizing the surfaces of the discontinuities into small elements and solving the problem piecewise.

Introduction

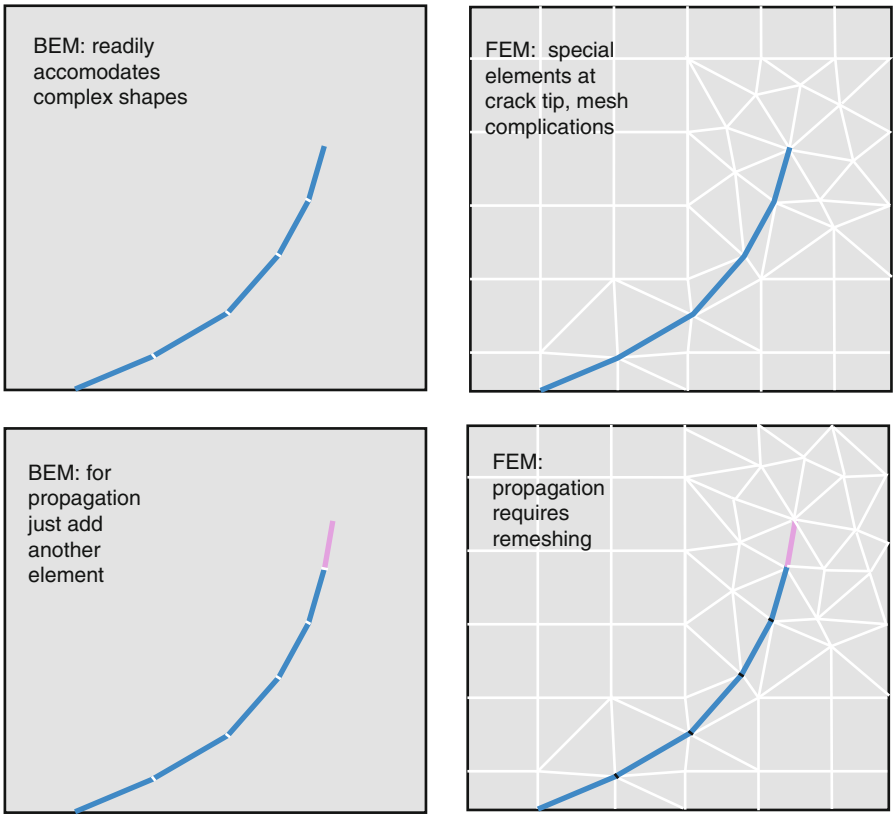
The boundary element method (BEM) is one of several numerical methods used to investigate deformation within the Earth. The principal difference between boundary element method and most other numerical methods of continuum mechanics, such as the finite element method (FEM), is that only the boundaries of the deforming body need to be meshed (Figure 1). While other methods require meshing the entire volume of the deforming body into two-dimensional (2D) or three-dimensional (3D) elements, the BEM only requires discretization of the boundaries (e.g., fractures, bedding planes, or external boundaries) into one-dimensional (1D) or 2D elements. Minimizing discretization decreases model building and run time as well as errors due to discretization (Crouch and Starfield, 1990). The greatest advantages of BEM modeling in geophysics are that (1) very complicated surface geometries, such as undulating fault surfaces, can be incorporated and (2) the method can readily accommodate fracture propagation by addition of new elements (Figure 1). The disadvantages of BEM are (1) that the stiffness matrix relating stresses to strains on the elements is non-sparse so that inverting the matrix to find stresses can be very CPU intensive and (2) heterogeneous material properties are more complex to incorporate in BEM than methods that utilize volume discretization, such as FEM (Becker, 1992).

How does BEM work?

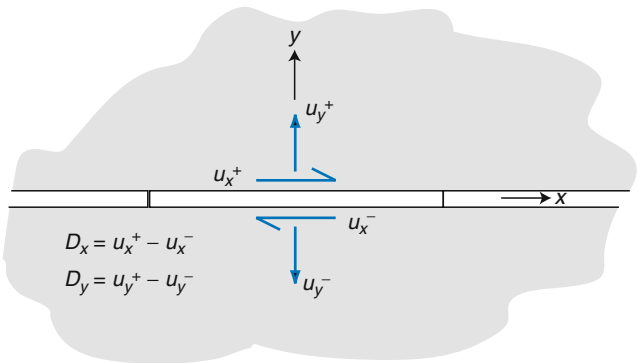
BEM models are based on the solution to Kelvin's problem for the deformation due to a point load within a body. The solution provides the stresses and displacements everywhere due to the point load. We can calculate the effects of distributed loads by integrating the point force. When the loads are applied over discrete elements, the relationships between the displacements at one element due to loads at another are used to build an influence coefficient matrix. When either the displacements or tractions are prescribed to all elements that comprise the BEM model, the method will solve for the unknown element tractions or displacements via the influence coefficient matrix. Once all the displacements and tractions along the boundary elements are known, the stresses and strains at any point within the body can be calculated using the solution of Kelvin's problem. So unlike FEM, where solutions are calculated at each node within the element mesh, within BEM, problems displacements and stresses can be calculated at any point within the body, once all the tractions and displacements along the elements are known.

Two variations of BEMs are commonly used, the *fictitious stress* and *displacement discontinuity* methods. The fictitious stress method is based on Kelvin's problem and applies uniform tractions along each boundary element. In contrast, the displacement discontinuity method applies displacements along each element such that the difference in displacement form one side of the element to the other is constant along the boundary element

Harsh K. Gupta (ed.), *Encyclopedia of Solid Earth Geophysics*, DOI 10.1007/978-90-481-8702-7,



Numerical Methods, Boundary Element, Figure 1 Contrasting discretization approaches of the boundary element method and finite element method. BEM only discretizes the boundaries and/or internal surfaces (*blue line*) rather than the entire volume. Fracture propagation is facilitated by addition of elements (*red line*).



Numerical Methods, Boundary Element, Figure 2 Formulation of the displacement discontinuity along each element within a boundary element method model. The normal and shear displacement discontinuity (D_x and D_y, respectively) are the difference between the displacements along the top and bottom of the element.

(Figure 2). While fictitious stress method works well for homogeneous bodies under external loads, the displacement discontinuity method is better suited for problems of solid bodies containing cracks (Crouch and Starfield, 1990). For this reason, the displacement discontinuity method is more common for geophysical analyses, which investigate deformation associated with dikes, sills, microcracks, joints, and faults within the Earth.

What can BEM do?

Boundary element method codes using the displacement discontinuity formulation have been developed to solve both two-dimensional (e.g., Crouch and Starfield, 1990 and three-dimensional problems (Okada, 1992; Thomas, 1993; Meade, 2007). Two-dimensional formulations use a linear dislocation for each element while three-dimensional formulations utilize either rectangular dislocations (Okada, 1992) or angular dislocations (Cominou and Dundurs, 1975) that can be assembled to form triangular elements of constant displacement discontinuity (Thomas, 1993; Meade, 2007). The advantage of triangular elements over rectangular elements is that complex nonplanar surfaces can be meshed without creating locations of overlaps and gaps.

Important refinements to the basic displacement discontinuity method for investigation of problems in geophysics include incorporation of frictional slip along

crack elements and prevention of interpenetration of crack walls during compressive loading of cracks. These refinements facilitate investigation of deformation associated with subsurface faults and joints. Algorithms for preventing interpenetration include the penalty method (Crouch and Starfield, 1990; Cooke and Pollard, 1997) and the complementarity method (Maerten et al., 2010). The penalty method employs normal and shear stiffness to each element to prevent interpenetration. The complementarity method uses an iterative solver to solve the inequality that opening should be equal to or greater than zero along each element (Maerten et al., 2010).

Heterogeneous properties can be incorporated within BEM formulations by way of a contact boundary. Stresses and displacements are prescribed to be uniform across the boundary between otherwise homogeneous portions of the model. With this constraint along the boundary, the solution for each homogeneous section is found separately. This approach has not been widely applied because it is difficult to implement; finite element methods are generally believed to be better suited for heterogeneous properties.

Applications of the boundary element method

In most BEM implementations, gravitational body forces are superposed onto the solution to the applied loading. This precludes direct investigation of problems related to topographic loading. However, Martel and Muller (2000) developed a formulation for analyzing topographic stresses using a long stress-free crack to simulate the topography.

Investigations of fluid flow through fracture networks has benefited from the surface discretization approach of the boundary element method. Models of subsurface flow through fracture networks utilize a boundary element method model that incorporates fluid flow along each element (e.g., Dershowitz and Fidelibus, 1999). Such approaches can permit the full coupling of mechanical deformation and fluid flow.

The boundary element method has found widespread use in the subfield of structural geology for solving problems of deformation and propagation of opening mode and sliding mode fractures (e.g., joints and faults). Both forward and inverse modeling techniques have utilized the boundary element method. Forward models apply loading to the crack system to determine the resulting deformation of the system. Such investigations may explore the development of secondary cracks around a fault or the aperture of a hydraulic fracture. In contrast, inverse models start with the observed deformation and invert to find the slip on the fault or opening on the crack that must have produced the observations. For example, deformation observed along the surfaces of the earth via GPS stations velocities or InSAR imagery can be inverted to find the slip distribution on the underlying fault (e.g., Maerten et al., 2005).

Bibliography

Becker, A. A., 1992. *The Boundary Element Method in Engineering: A Complete Course*. New York: McGraw-Hill.

Cominou, M. A., and Dundurs, J., 1975. The angular dislocation in a half space. *Journal of Elasticity*, **5**, 203–216.

Cooke, M. L., and Pollard, D. D., 1997. Bedding plane slip in initial stages of fault-related folding. *Journal of Structural Geology*, **19**, 567–581. Special Issue on Fault-Related Folding.

Crouch, S. L., and Starfield, A. M., 1990. *Boundary Element Method in Solid Mechanics with Applications in Rock Mechanics and Geological Engineering*. Unwin Hyman, Boston, 322 p.

Dershowitz, W., and Fidelibus, C., 1999. Derivation of equivalent pipe network analogues for three-dimensional discrete fracture networks by the boundary element method. *Water Resources Research*, **35**(9), 2685–2691.

Maerten, F., Resor, P., Pollard, D., and Maerten, L., 2005. Inverting for slip on three-dimensional fault surfaces using angular dislocations. *Bulletin of the Seismological Society of America*, **95**(5), 1654–1665.

Maerten, F., Maerten, L., and Cooke, M. L., 2010. Solving 3D boundary element problems using constrained iterative approach. *Computational Geosciences*, doi:10.1007/s10596-009-9170-x [in press].

Martel, S., and Muller, J., 2000. A two-dimensional boundary element method for calculating elastic gravitational stresses in slopes. *Pure and Applied Geophysics*, **157**, 989–1007, doi:10.1007/s000240050014.

Meade, B. J., 2007. Algorithms for the calculation of exact displacements, strains, and stresses for triangular dislocation elements in a uniform elastic half space. *Computers and Geosciences*, **33**, 1064–1075, doi:10.1016/j.cageo.2006.12.003.

Okada, Y., 1992. Internal deformation due to shear and tensile faults in a half-space. *Bulletin of the Seismological Society of America*, **82**, 1018–1040.

Thomas, A. L., 1993. *Poly3D: A Three-Dimensional Polygonal-Element Displacement Discontinuity Boundary Element Computer Program with Applications to Fractures, Faults and Cavities in the Earth's Crust*. MS thesis, California, Stanford University.

Cross-references

Numerical Methods, Finite Difference
Numerical Methods, Finite Element

NUMERICAL METHODS, DOMAIN DECOMPOSITION

Alfio Quarteroni[1,2], Luca Formaggia[1]
[1]MOX, Department of Mathematics, Politecnico di Milano, Milano, Italy
[2]CMCS-MATHICSE, EPFL, Lausanne, Switzerland

Definition

Computational domain. A bounded spatial region where a partial differential equation is solved.

Dirichlet boundary condition. A condition imposed directly on the unknown of the differential problem, also called essential boundary condition.

Neumann boundary condition. A condition imposed on the fluxes associated to the differential problem at hand,

that is, stresses, heat flux, etc. Also called natural boundary condition.

Parallel computer. A computer with more than one processing unit capable of computing operations concurrently.

Parallel computation. A numerical procedure executed on a parallel computer.

Preconditioner. In the context of the solution of a linear system by an iterative method, a preconditioner is an easily invertible matrix spectrally similar to the matrix governing the problem.

Introduction

Domain Decomposition (DD) method is a technique for the solution of partial differential equations which can be instrumental to the development of *parallel computations*. It can be used in the framework of discretization methods, for example, finite elements, finite volumes, finite differences, or spectral element methods.

It is based on the reformulation of the given boundary-value problem on a partition of the computational domain Ω into M subdomains Ω_i, with $i = 1, \ldots, M$. Typically M is also the number of processors at disposal, even if it is possible to have more subdomains per processor.

The DD method also provides a convenient framework for the solution of heterogeneous or multi-physics problems, that is, those that are governed by differential equations of different kinds in different subregions of the computational domain. In this case, the subdomains conform with the subregions. In this work, however, we will address only homogeneous DD methods, yet some of the concepts presented here may be readily adapted for the nonhomogeneous case.

We will refer to a partial differential equation of the form

$$-div(\boldsymbol{T}(\boldsymbol{u})) = \boldsymbol{f} \text{ in } \Omega \tag{1}$$

with suitable boundary conditions on $\partial\Omega$.

For instance, in a problem of elastostatic, $\boldsymbol{u}$ is the displacement and $\boldsymbol{T}(\boldsymbol{u})$ the stress tensor; in a heat conduction problem $\boldsymbol{u}$ is a scalar variable, the temperature, and $\boldsymbol{T}$ the Laplace operator. With a rather similar formalism we may also account for the Stokes problem. The method is in fact applicable also to time-dependent problems, since in most cases the time advancing scheme required by the numerical simulation leads eventually to a sequence of problems of type (1). For instance, if we consider the evolution equation

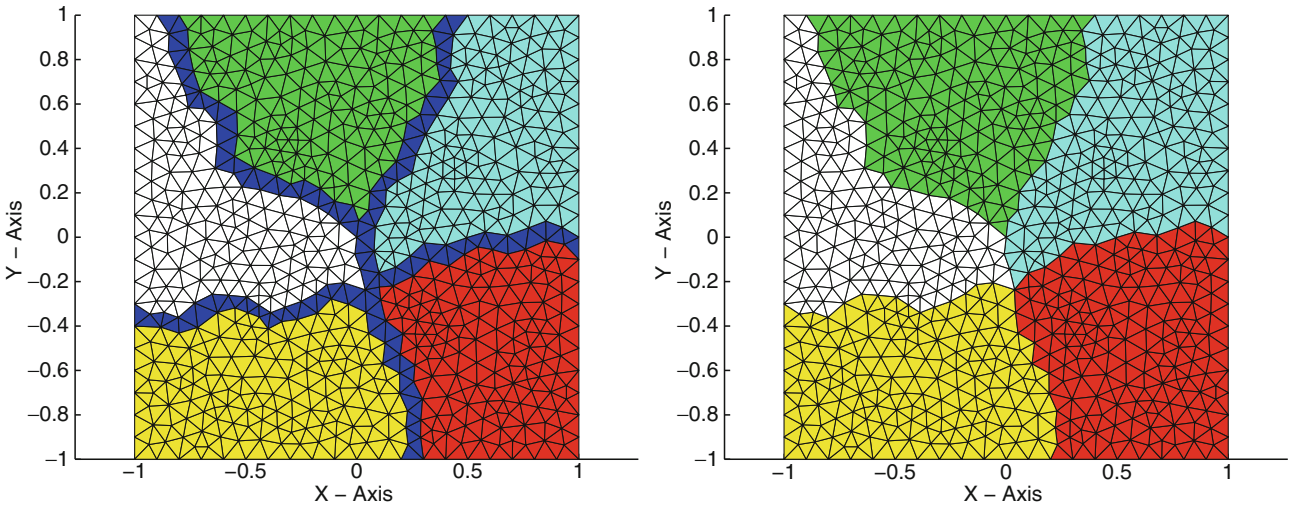
$$\partial_t \boldsymbol{u} - div(\boldsymbol{T}(\boldsymbol{u})) = \boldsymbol{f},$$

a time discretization by the implicit Euler method leads to solve for the unknown $\boldsymbol{u}^n$ at each time step t^n a differential problem of the type

$$\boldsymbol{u}^n - \delta t div(\boldsymbol{T}(\boldsymbol{u}^n)) = \delta t \boldsymbol{f}^n + \boldsymbol{u}^{n-1}.$$

It is still of the form (1) with $div(\boldsymbol{T})$ replaced by $\delta t\ div(\boldsymbol{T}) - \boldsymbol{I}$, δt being the time step and $\boldsymbol{I}$ the identity operator. A similar form is obtained also for the elastodynamic equations, discretized in time, for instance, by a Newmark method (Quarteroni and Valli, 1994).

Typically, and for evident practical reasons, the partition into subdomains is made after having triangulated Ω into a finite element mesh τ_h, each subdomain being in fact formed by a set of elements of the original grid (see Figure 1). The partition is often made automatically, using libraries like METIS (Karypis and Kumar, 1998b) or



Numerical Methods, Domain Decomposition, Figure 1 Partition of a computational domain starting from a given triangulation. *Left*: A partition into subdomains with an overlap equal to one layer of elements. *Right*: Partition of a domain with disjoint subdomains.

PARMETIS (Karypis and Kumar, 1998a), the latter able to exploit a parallel architecture also for this preprocessing stage.

More in particular, there are two ways of subdividing the computational domain into subdomains. The first uses disjoint subdomains, where $\Omega_i \cap \Omega_j = \emptyset$ for $i \neq j$. Here, the interface between subdomains reduces to surfaces (in 3D), and lines (in 2D). The second strategy adopts *overlapping subdomains*, usually built by letting the subdomains of an initial nonoverlapping partition grow of a certain factor δ. In most of the cases δ is dictated by the number of "layers" of grid elements that are added to the original partition. The minimal overlap is of one element, like that shown in Figure 1. For a more complete review of DD methods one may refer to (Smith et al., 1996; Quarteroni and Valli, 1999; Wohlmuth, 2001; Toselli and Widlund, 2005; Mathew, 2008); several examples with emphasis on parallel computations are reported, for instance, in (Bruaset and Tveito, 2006).

Algorithms without overlap

We indicate with Γ_{ij} the interface between subdomains Ω_i and Ω_j and we set $\Gamma = \cup_{ij}\Gamma_{ij}$. We exploit the fact that problem (1) is equivalent to solving the following M coupled problems for $i = 1, \ldots, M$,

$$-div(\boldsymbol{T}(\boldsymbol{u}_i)) = \boldsymbol{f}_i \text{ in } \Omega_i \tag{2}$$

with the same boundary conditions of the original problem applied to $\partial\Omega_i \cap \partial\Omega$, while on each Γ_{ij} we set the continuity of the local solutions $\boldsymbol{u}_i$ and of the fluxes, that is,

$$\boldsymbol{u}_i = \boldsymbol{u}_j, \qquad \boldsymbol{T}(\boldsymbol{u}_i)\cdot\boldsymbol{n}_{ij} + \boldsymbol{T}(\boldsymbol{u}_j)\cdot\boldsymbol{n}_{ji} = 0, \tag{3}$$

where $\boldsymbol{n}_{ij}$ is the normal to Γ_{ij} outward oriented w.r.t. Ω_i. It may be proved that $\boldsymbol{u}_i = \boldsymbol{u}_{|\Omega i}$ (Quarteroni and Valli, 1999).

Problems (2) are coupled because of relations (3). A parallel algorithm may be obtained by an iterative procedure where conditions (3) are enforced in such a way to generate at each iteration decoupled problems that can be run on different processes, with only a small amount of communication needed at the beginning of each iteration. For the sake of space we give an example of just one of these procedures, called *Dirichlet-Neumann* for the case of two subdomains Ω_1 and Ω_2. Starting from a guess for $u_1^{(0)}$ and $u_2^{(0)}$, the algorithm solves for $k = 1, 2, \ldots$ the following sequence of independent problems,

$$\begin{cases} -div(\boldsymbol{T}(\boldsymbol{u}_1^{(k+1)})) = \boldsymbol{f}_1, & \text{in } \Omega_1, \\ \boldsymbol{u}_1^{(k+1)} = \theta u_2^{(k)} + (1-\theta)\boldsymbol{u}_1^{(k)}, & \text{on } \Gamma_{12} \end{cases} \text{ and}$$

$$\begin{cases} -div(\boldsymbol{T}(\boldsymbol{u}_2^{(k+1)})) = \boldsymbol{f}_2, & \text{in } \Omega_2 \\ \boldsymbol{T}(\boldsymbol{u}_2^{(k+1)})\cdot\boldsymbol{n}_{21} = \boldsymbol{T}(\boldsymbol{u}_1^{(k)})\cdot\boldsymbol{n}_{21} & \text{on } \Gamma_{12}. \end{cases}$$

until a measure of the difference $u_1^{(k+1)} - u_2^{(k+1)}$ on Γ_{12} is below a given tolerance; θ here is a convenient relaxation factor. The Dirichlet-Neumann technique is not easily extendable to an arbitrary number of subdomains (unless suitable *coloring techniques* are used) and its convergence characteristics strongly depend on the geometry of the subdomains as well as on the possible jump of characteristic coefficients (for instance, the viscosity of different rocks forming a sedimentary basin). Other more favorable techniques like the Neumann-Neumann, Robin-Robin, and FETI methods are described in (Quarteroni and Valli, 1999; Wohlmuth, 2001; Toselli and Widlund, 2005).

The DD method may be also set from an algebraic viewpoint. Indeed when discretized, for instance, by a finite element method, (2) and (3) reduce to a system of linear equations, which may be written in a block form as

$$\begin{pmatrix} A_{II} & A_{I\Gamma} \\ A_{\Gamma I} & A_{\Gamma\Gamma} \end{pmatrix} \begin{pmatrix} \boldsymbol{u}_I \\ \boldsymbol{u}_\Gamma \end{pmatrix} = \begin{pmatrix} \boldsymbol{f}_I \\ \boldsymbol{f}_\Gamma \end{pmatrix},$$

where A_{II} is a *block diagonal* matrix with M blocks of dimension equal to the number of unknowns internal to each subdomain, the latter being collected in $\boldsymbol{u}_I$, while $\boldsymbol{u}_\Gamma$ is the vector of unknowns on the interface Γ. If A_{II} is invertible (and normally this is the case), we may obtain a problem for the $\boldsymbol{u}_\Gamma$ only (*Schur complement system*), $\Sigma_\Gamma \boldsymbol{u}_\Gamma = \boldsymbol{\chi}_\Gamma$, where $\Sigma_\Gamma = A_{\Gamma\Gamma} - A_{\Gamma I}A_{II}^{-1}A_{I\Gamma}$ is the so-called *Shur complement matrix* w.r.t. $A_{\Gamma\Gamma}$. Having solved for $\boldsymbol{u}_\Gamma$ the computation of $\boldsymbol{u}_I$ can be done in a perfect parallel fashion by solving the block diagonal problem $A_{II}\boldsymbol{u}_I = \boldsymbol{f}_I - A_{I\Gamma}\boldsymbol{u}_\Gamma$.

A DD scheme with no overlap may be interpreted as *preconditioned iterative scheme* (Quarteroni and Valli, 1994; Quarteroni, 2009) for the Schur complement system, where the preconditioner can be efficiently applied in a parallel setting. A crucial issue for parallel computing is that of *scalability*. In the DD setting, an algorithm is said to be scalable if its convergence properties do not depend on the number of subdomains M, and in particular does not degrade if we keep the ratio M/N between number of subdomains and the total number N of unknowns of our problem fixed. Indeed in this case (if we neglect communication overheads), we may solve in the same time a problem twice as large by doubling the number of processors.

A scalable parallel preconditioner cannot be built using only local (i.e., at the subdomain level) approximations of the Schur matrix; we need to add also a *coarse operator* that has the role of transferring information among far away subdomains. The typical form of the preconditioner is (we write directly the inverse operator since it is the one actually required by the iterative procedure (Quarteroni et al., 2007))

$$P_\Sigma^{-1} = \sum_{i=1}^{M} R_i^T \Sigma_i^* R_i + R_0^T \Sigma_0^* R_0, \tag{4}$$

where R_i is a *restriction operator* that selects from all the $\boldsymbol{u}_\Gamma$ those local to the i-th subdomain, and Σ_i^* is a local approximation of the inverse Schur matrix, which

typically can be built using just data related to the i-th subdomain (and thus in parallel). Finally, Σ_0^* is the coarse operator, of small dimension (typically of the order of M), whose role is to guarantee the coupling among all subdomains and it is necessary for scalability. The application of the preconditioner (i.e., the computation of $P_\Sigma^{-1}\boldsymbol{x}$, where $\boldsymbol{x}$ is any vector of the right length), can be done in parallel a part from the coarse operator, which however, being small in size, is irrelevant in the computational cost balance. Several scalable preconditioners are available, see (Toselli and Widlund, 2005) and (Canuto et al., 2007).

Methods with overlap

A typical DD scheme with overlapping subdomains is the Schwarz method. In its basic form, it is an iterative algorithm for the solution of (1) that for $k = 1, 2, \ldots$ solves a series of local problems on each subdomain Ω_i, where on the nonempty interfaces $\Gamma_{ij} = \partial\Omega_i \cap \Omega_j$ we apply Dirichlet boundary conditions using the latest values available from Ω_j, that is,

$$\begin{cases} -div(\boldsymbol{T}(\boldsymbol{u}_i^{(k+1)}) = \boldsymbol{f}_i, & \text{in } \Omega_i, \\ \boldsymbol{u}_i^{(k+1)} = \boldsymbol{u}_j^{(k)}, & \text{on } \Gamma_{ij}. \end{cases}$$

The iteration continues until the difference between two successive iterations is sufficiently small. The convergence analysis of the Schwarz method may be found in (Smith et al., 1996; Quarteroni and Valli, 1999; Toselli and Widlund, 2005), and in (Canuto et al., 2007) in the context of spectral element discretizations.

The method is seldom used in this form, however. Yet, it is probably the most adopted method to build parallel preconditioners for an iterative solver of the global problem, giving rise the so-called *Krylov-Schwarz* (for linear problems) and *Newton-Krylov-Schwarz* methods.

A Schwarz parallel preconditioner P_S may be written similarly to (4), that is,

$$P_S^{-1} = \sum_{i=1}^{M} R_i^T A_i^{-1} R_i + R_0^T A_0^{-1} R_0,$$

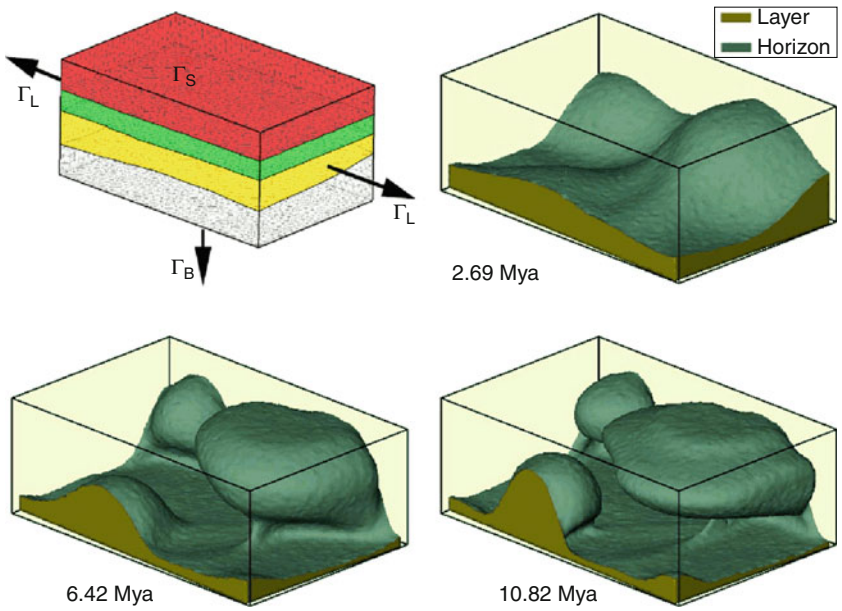
where here the restriction operator R_i extracts from a vector of length N the elements corresponding to the unknowns internal to the extended subdomain Ω_i, $A_i = R_i A R_i^T$ is the local matrix extracted from the matrix A of our problem, and finally A_0 is again a global coarse operator of size of the order M needed for scalability, R_0 being the corresponding restriction matrix.

Again, apart from the coarse operator, the computation of $P_S^{-1}\boldsymbol{x}$ can be done in parallel since the matrices A_i are local, and can be handled by each processor.

Application to geophysical and geological problems

DD methods have been applied successfully in the context of geophysical problems, like acoustic and elastodynamic wave propagation (Faccioli et al., 1997), or full-waveform tomography (Sourbier et al., 2009).

It has been also applied to speed up the simulation of the evolution of sedimentary basins. Here, different types of sediments, such as gravel, sand, rocks, and biological



Numerical Methods, Domain Decomposition, Figure 2 The simulated evolution of a salt diapir. In the first graphic, the computational domain showing the subdivision of the layers. The other pictures show snapshots of the evolution of the salt layer with the formation of a diapir.

remains that have been transported by the wind, the rivers and, sometimes, by the sea, accumulate, are buried deeper and deeper, and are transformed eventually into rocks by a complex process of compaction and diagenesis.

On geological scales, the evolution of a sedimentary basin may be tackled by a fluid approach where each rock layer is modeled as a viscous fluid. This is particularly convenient in the presence of salt tectonics (Massimi et al., 2007).

The computational domain Ω is split in several subdomains Ω_i, which usually correspond to the rock layers, as shown in Figure 2. At each time step, we have to solve in the domain Ω a Stokes problem, possibly with non-Newtonian rheology, to compute the instantaneous velocity field,

$$\begin{cases} -div(\boldsymbol{T}(\boldsymbol{v})) + \nabla p & = \boldsymbol{f}, \\ div(\boldsymbol{v}) & = \phi, \\ \frac{\partial \rho}{\partial t} + (\boldsymbol{v}\nabla)\rho & = 0. \end{cases} \tag{5}$$

Here, $\boldsymbol{v}$ is the velocity, p the pressure, $\boldsymbol{f}$ the external gravitational field, and ρ the density. The stress tensor $\boldsymbol{T}$ depends on characteristics of the rocks, such as viscosity and density, which may be discontinuous across layers. The function ϕ may account for compaction processes, or may be simply set to zero when the hypothesis of isochoric flow is acceptable. The last equation describes the evolution of the density, which is simply advected by the velocity field. The presence of faults is accounted for by appropriately reducing the rock viscosity in the vicinity of the faults.

The movement of the layers has been tracked using a level set technique. A finite element scheme has been adopted for the discretization of the Stokes problem, while a conservative finite volume scheme has been used for the tracking of the layer interfaces. The parallel implementation has been carried out with the help of the TRILINOS library (Heroux et al., 2005), using a Schwarz algorithm with a coarse operator built by aggregation (Sala, 2004).

Figure 2 shows the evolution of a salt dome. Salt is less compressible than the surrounding rock, so during the sedimentation process it ends having a smaller density than the overburden. We are then facing a Rayleigh-Taylor instability and any small perturbation will cause the salt to rise, with large movements that causes the so-called salt diapirism.

Bibliography

Bruaset, A., and Tveito, A. (eds.), 2006. *Numerical Solution of Partial Differential Equations on Parallel Computers*. New York: Springer. Lecture Notes in Computational Science and Engineering, Vol. 51.

Canuto, C., Hussaini, M. Y., Quarteroni, A., and Zang, T. A., 2007. *Spectral Methods. Evolution to Complex Geometries and Application to Fluid Dynamics*. Berlin/Heidelberg: Springer.

Faccioli, E., Maggio, F., Paolucci, R., and Quarteroni, A., 1997. 2D and 3D elastic wave propagation by a pseudo-spectral domain decomposition method. *Journal of Seismology*, **1**(3), 237–251.

Heroux, M., et al., 2005. An overview of the TRILINOS project. *ACM Transactions on Mathematical Software*, **31**(3), 397–423 (TOMS).

Karypis, G., and Kumar, V., 1998a. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal of Scientific Computing*, **20**, 359–392.

Karypis, G., and Kumar, V., 1998b. *METIS: unstructured graph partitining and sparse matrix ordering system*. Techical Report 98-036, University of Minnesota, Department of Computer Science.

Massimi, P., Quarteroni, A., Saleri, F., and Scrofani, G., 2007. Modeling of salt tectonics. *Computer Methods in Applied Mechanics and Engineering*, **197**(1–4), 281–293.

Mathew, T., 2008. *Domain Decomposition Methods for the Numerical Solution of Partial Differential Equations*. New York: Springer. Lecture Notes in Computational Science and Engineering.

Quarteroni, A., 2009. *Numerical Models for Differential Problems*. Berlin: Springer. MS&A, Vol. 2.

Quarteroni, A., and Valli, A., 1994. *Numerical Approximation of Partial Differential Equations*. Berlin/Heidelberg: Springer.

Quarteroni, A., and Valli, A., 1999. *Domain Decomposition Methods for Partial Differential Equations*. Oxford: Oxford Science.

Quarteroni, A., Sacco, R., and Saleri, F., 2007. *Numerical Mathematics*, 2nd edn. Berlin/Heidelberg: Springer.

Sala, M., 2004. Analysis of two-level domain decomposition preconditioners based on aggregation. *Mathematical Modelling and Numerical Analysis*, **38**(5), 765–780.

Smith, B., Bjorstad, P., and Gropp, W., 1996. *Domain Decomposition, Parallel Multilevel Methods for Elliptic Partial Differential Equations*. New York: Cambridge University Press.

Sourbier, F., Operto, S., Virieux, J., Amestoy, P., and L'Excellent, J. Y., 2009. FWT2D: a massively parallel program for frequency-domain full-waveform tomography of wide-aperture seismic data–part 2: numerical examples and scalability analysis. *Computers and Geosciences*, **35**(3), 496–514.

Toselli, A., and Widlund, O., 2005. *Domain Decomposition Methods – Algorithms and Theory*. New York: Springer. Springer Series in Computational Mathematics, Vol. 34.

Wohlmuth, B., 2001. *Discretization Methods and Iterative Solvers Based on Domain Decomposition*. Berlin: Springer. Lecture Notes in Computational Science and Engineering.

Cross-references

Numerical Methods, Finite Element
Numerical Methods, Multigrid
Sedimentary Basins

NUMERICAL METHODS, FINITE DIFFERENCE

Johan O. A. Robertsson[1], Joakim O. Blanch[2]
[1]Schlumberger Cambridge Research, Cambridge, UK
[2]BHP Billiton, Houston, TX, USA

Definition

Finite-difference method. A method to approximate derivatives between neighboring points in a grid. The method can be applied to solve partial-differential equations, such as the wave equation.

Introduction

The finite-difference (FD) method is among the most commonly used methods for simulating wave propagation in a heterogeneous Earth. In this article, we describe the FD method for modeling wave propagation on Cartesian grids in acoustic, elastic isotropic, elastic anisotropic, as well as viscoacoustic/elastic media. The basic equations for wave propagation can be formulated in various formally equivalent ways. We will restrict our description to systems of first-order partial-differential equations with pressure/stress and particle velocities as wavefield variables. Due to its versatility, attractive stability, and dispersive properties, particularly for modeling wave propagation in elastic media, this formulation has been by far the most popular formulation over the last several decades. A more extensive review of the FD method is given by Moczo et al. (2007).

We start with a brief review of wave propagation theory and introduce some fundamental concepts including the basic structure (staggering) of the FD grid. In the following two sections, we discuss the choice of FD approximation and introduce two basic numerical properties of the FD method: numerical dispersion and stability. The last two sections are devoted to different boundary conditions and source implementations.

Theory of wave propagation and fundamental concepts

We focus on wave propagation in 3D Cartesian coordinates with coordinate axes (x, y, z). The special cases of 1D or 2D can easily be generalized from the expressions presented here, for example, in 2D, omitting terms or expressions containing components in the y-direction. The particle velocity wavefield is denoted by $\vec{v} = (v_x, v_y, v_z)$. In acoustic media, the second wavefield variable is the scalar quantity pressure, p, whereas in elastic media it is the (symmetric) stress tensor, S. We will often be representing stresses and particle velocities using stress and strain vectors defined as (the so-called Voigt notation):

$$\vec{\sigma} = \begin{pmatrix} \sigma_{xx} & \sigma_{yy} & \sigma_{zz} & \sigma_{yz} & \sigma_{xz} & \sigma_{xy} \end{pmatrix}^T$$

and

$$\vec{\varepsilon} = \begin{pmatrix} \varepsilon_{xx} & \varepsilon_{yy} & \varepsilon_{zz} & 2\varepsilon_{yz} & 2\varepsilon_{xz} & 2\varepsilon_{xy} \end{pmatrix}^T,$$

where the strain vector is related to the particle velocity vector components through:

$$\partial_t \vec{\varepsilon} = \begin{pmatrix} \partial_x v_x & \partial_y v_y & \partial_z v_z & \partial_z v_y + \partial_y v_z & \partial_z v_x + \partial_x v_z \\ \partial_x v_y + \partial_y v_x \end{pmatrix}^T.$$

The system of equations for wave propagation consists of two (dependent) sets of equations. First, the equation of motion (or Newton's second law):

$$\rho \frac{\partial}{\partial t} \vec{v} = \nabla \cdot S + \vec{f}, \tag{1}$$

where ρ is the density of the medium, $\nabla \cdot S$ denotes divergence of the stress tensor and $\vec{f}$ is a point-force source (source formulations driving the FD simulations will be described in detail later). In an acoustic medium, only the diagonal elements of the stress tensor are non-zero and equal to the negative pressure $-p$.

The second set of equations for wave propagation is the constitutive stress–strain relation:

$$\frac{\partial}{\partial t} \vec{\sigma} = C \frac{\partial}{\partial t} \vec{\varepsilon}, \tag{2}$$

where C is the 6-by-6 stiffness matrix. For an elastic isotropic medium, the stiffness matrix takes the form:

$$C = \begin{pmatrix} \lambda + 2\mu & \lambda & \lambda & 0 & 0 & 0 \\ \lambda & \lambda + 2\mu & \lambda & 0 & 0 & 0 \\ \lambda & \lambda & \lambda + 2\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu \end{pmatrix}, \tag{3}$$

whereas in a general anisotropic medium, all components of the stiffness matrix may be non-zero. As we shall see, this has profound implications on the choice of grid geometry to discretize the constitutive relation. Finally, although redundant, the acoustic constitutive relation can also be written using the notation of the stiffness matrix and the stress–strain constitutive relation. As noted above, instead of a stress vector with three independent components, we obtain:

$$\vec{\sigma} = -\begin{pmatrix} p & p & p & 0 & 0 & 0 \end{pmatrix}^T. \tag{4}$$

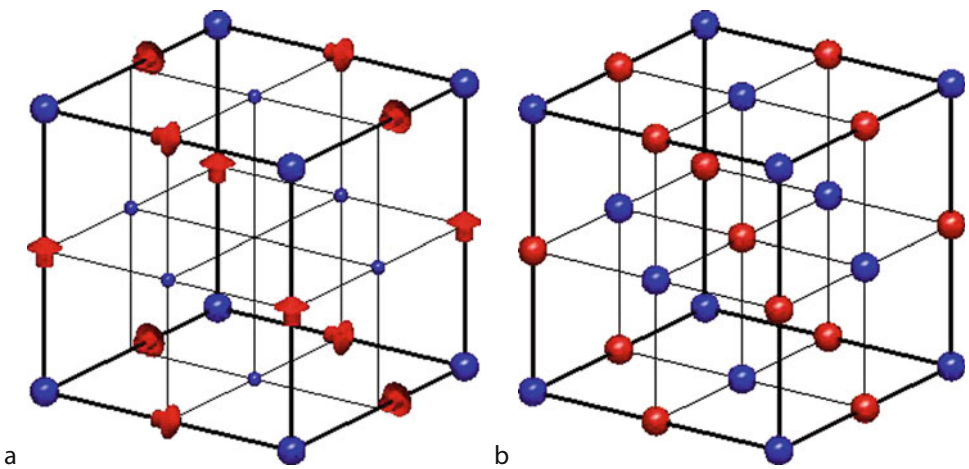
In an acoustic medium, only the (3-by-3) upper left quadrant of the stiffness matrix is non-zero with all elements equal and identical to the bulk modulus of the acoustic medium κ. Clearly this results in three identical equations – the familiar Hooke's law: $\frac{\partial}{\partial t} p = -\kappa \nabla \cdot \vec{v}$.

The FD method comprises solving Equations 1 and 2 by discretizing them in time and space and stepping forward in small incremental time steps. In particular, the exact choice of discretization in space turns out to be of fundamental importance. The so-called staggered grid was introduced for modeling isotropic elastic wave propagation by Virieux (1984, 1986) and Levander (1988, 1989). In the following, we will be referring to this as the Virieux grid. The choice of this grid results in schemes with attractive stability and dispersion properties, which other non-staggered grids do not necessarily possess. FD approximations to first-order derivatives naturally result in outputting the derivative of the wavefield at a location in between the discretized wavefield. By shifting the exact location of individual wavefield components by half a grid step in certain directions within each

grid cell, it is possible to ensure that spatial derivatives can be computed exactly at the locations where they are needed to advance canonical wavefield quantities in time using Equations 1 and 2. The structure of the Virieux staggered grid is illustrated in Figure 1a. The Virieux grid is also a good choice for modeling wave propagation in acoustic media. The resulting sparsely populated staggered-grid cell is equivalent to Figure 1a but without the shear stresses populating the grid cell (small blue balls in Figure 1a). In many situations, it is desirable to have a computational model that may contain both acoustic and elastic regions (e.g., modeling a marine seismic experiment or the seismic response of magma chambers). This is straightforward using a Virieux grid, by simply setting $\lambda = \kappa$ and $\mu = 0$ in Equation 3 in the acoustic regions. Updating the equations will guarantee that the "acoustic stress vector" in Equation 4 is satisfied to within-machine precision.

The Virieux grid is the natural choice for modeling wave propagation in elastic media as long as the stiffness matrix, C, in Equation 2 belongs to a class of anisotropic materials referred to as orthorhombic media and as long as the symmetry axes of the medium are aligned with the Cartesian grid (isotropic media is a special case that belongs in this family of media). For this class of anisotropic materials, all elements that are zero in Equation 3 remain zero. However, as soon as we introduce more complex anisotropic symmetry classes or rotate the symmetry axes with respect to the grid, the elements that are zero in Equation 3 will become non-zero. As a consequence, Igel et al. (1995) noted that spatial derivatives are no longer available at all locations where they are needed. Igel et al. (1995) solved this by interpolating strains between their natural locations. However, the scheme is both expensive and results in fairly complex implementations. Saenger (2000) solved the problem in 2D by observing that a rotation of the staggered grid by 45° results in a natural choice for staggering the wavefield quantities so that no interpolation is necessary. The so-called rotated staggered scheme can also be generalized to 3D, although it can no longer be seen as a simple rotation of the Virieux grid. Another staggered grid that avoids the necessity to interpolate wavefield quantities in anisotropic media is the Lebedev grid (Lebedev, 1964; Lisitsa and Vishnevskiy, 2010), which is illustrated in Figure 1b. Bernth and Chapman (2010) analyzed and compared the different staggered-grid geometries and concluded that whereas the rotated staggered grid and the Lebedev grid are equivalent in 2D, the Lebedev grid is different and a better choice compared to the rotated staggered grid in 3D. The Lebedev grid is both computationally more efficient and also lends itself to a simpler implementation as it can be regarded as a combination of four independent Virieux grids that decouple in isotropic media.

So far, the discussion in this section has concerned lossless media. In order to account for attenuation, a viscoacoustic or viscoelastic model is introduced by using a different constitutive stress–strain relation compared to Equation 2. Equation 2 can be thought of as a model of springs connecting particles in a lattice. When compressing the springs, energy is stored for an infinite time or until the compressed springs are released after which the particles return to their original positions. In a viscoelastic medium, dashpots are introduced in parallel or in series with the springs causing energy to dissipate after compression. The stiffness matrix becomes time-



Numerical Methods, Finite Difference, Figure 1 Staggered-grid geometries. (**a**) (*left*): Virieux staggered grid (Virieux, 1986). Large blue ball: normal stress components; *small blue ball*: shear stress components; red arrows: particle velocity components in the directions indicated. (**b**) (*Right*): Lebedev staggered grid (Lebedev, 1964; Lisitsa and Vishnevskiy, 2010; Bernth and Chapman, 2010). *Large blue ball*: all stress components; large red ball: all particle velocity components. The authors acknowledge Chris Chapman and Henrik Bernth (Schlumberger Cambridge Research) for providing the figure.

dependent and the multiplication with the time derivative of the strains is replaced by a convolution in time. The new viscoelastic constitutive relation becomes:

$$\frac{\partial}{\partial t}\vec{\sigma} = G(t) * \frac{\partial}{\partial t}\vec{\varepsilon}, \tag{5}$$

where *G(t)* is the new relaxation stiffness matrix, which contains elements that are functions of time. Such a particularly useful function corresponds to a spring/dashpot configuration called a standard linear solid (Robertsson et al., 1994a). First, due to the exponential kernel (in time) in the relaxation function of a standard linear solid, the convolution in Equation 5 can be eliminated at the expense of introducing a new set of differential equations to be solved at each time step and spatial location. No additional complications due to various grid geometries arise after introducing the new differential equations solving for the so-called viscoelastic memory variables as these equations are fairly simple ordinary differential equations in time (as opposed to partial-differential equations with additional spatial derivatives). Second, Robertsson et al. (1994a) and Blanch et al. (1995) showed how arrays of standard linear solid elements can advantageously be used to model constant quality factor Q versus frequency and also how to model attenuation of P and S waves in elastic isotropic media separately. However, exactly what attenuation to model for different wave types in an anisotropic medium is unclear and remains a topic for research (there are no pure P and S waves in an anisotropic medium).

Finite-difference approximations

The FD method approximates derivatives by combining neighboring function values on a grid, where the particular combination is commonly derived using the Taylor expansion of the function at the different sample points. The simplest example of a FD approximation is the first derivative of a function p at x_0 using only two samples:

$$\frac{\partial p(x_0)}{\partial x} = \frac{1}{\Delta x}\left(p\left(x_0 + \frac{\Delta x}{2}\right) - p\left(x_0 - \frac{\Delta x}{2}\right)\right) + C_1 \frac{\partial^3 p(x_0)}{\partial x^3}(\Delta x)^2 + O\left((\Delta x)^4\right), \tag{6}$$

where Δx is the sampling interval and C_1 is a constant. The lowest-order error term in this expansion is proportional to the square of the sampling interval. Hence, as we decrease Δx, the FD approximation will become increasingly accurate (the increase in accuracy proportional to the square of Δx) and in the limit be equivalent to the first derivative of p. The approximation is therefore considered second-order accurate since the lowest-order error term depends on the square of the sampling interval. Higher-order approximations of the first derivative can be expressed as a weighted sum of additional adjacent sample points:

$$\frac{\partial p(x_0)}{\partial x} = \frac{1}{\Delta x}\sum_{i=1}^{M}\alpha_i\left(p\left(x_0 + \frac{(2i-1)\Delta x}{2}\right) - p\left(x_0 - \frac{(2i-1)\Delta x}{2}\right)\right) + O\left((\Delta x)^{2M}\right). \tag{7}$$

The coefficients α_i describe what is often called the FD stencil and are chosen such that terms containing lower-order error factors are canceled (Table 1).

The accuracy of the FD approximations for the temporal derivatives and the spatial derivatives are often different and schemes are accordingly named with two accuracy orders. An O(2,8) accurate scheme would be second-order accurate in time and eighth-order accurate in space.

From Equation 6, we note that the derivative estimate of a spatial derivative is computed in between sample points of the wavefield quantity in question. As described in the previous section, this is the reason why the staggered-grid formulation is particularly well suited for FD modeling of the system of first-order differential equations describing wave propagation. The grid staggering is also applied in time, typically utilizing the simple second-order accurate approximation in Equation 6, resulting in a time-stepping method referred to as a "leap-frog" scheme. The leap-frog scheme is in turn referred to as an explicit FD scheme since the differential equations only contain one unknown after discretization in space and time, and which, therefore, can be computed directly without solving further systems of equations. For example, if we are in the process of updating particle velocities at a new time step $t_0 + \Delta t$ using Equation 1, we can compute the particle velocities directly from the known particle velocities at time t_0 and through spatial FD approximations of the known stresses at time $t_0 + \Delta t/2$. We then proceed to update stresses at time $t_0 + 3\Delta t/2$ using Equation 2 explicitly from the stresses at $t_0 + \Delta t/2$ and particle velocities at $t_0 + \Delta t$. In this fashion, we march through the FD simulation until we reach the desired maximum time of the simulation. Clearly, we need to initiate the simulation with a so-called initial conditions for particle velocities at time 0 and stresses at time $\Delta t/2$.

Accuracy and numerical dispersion

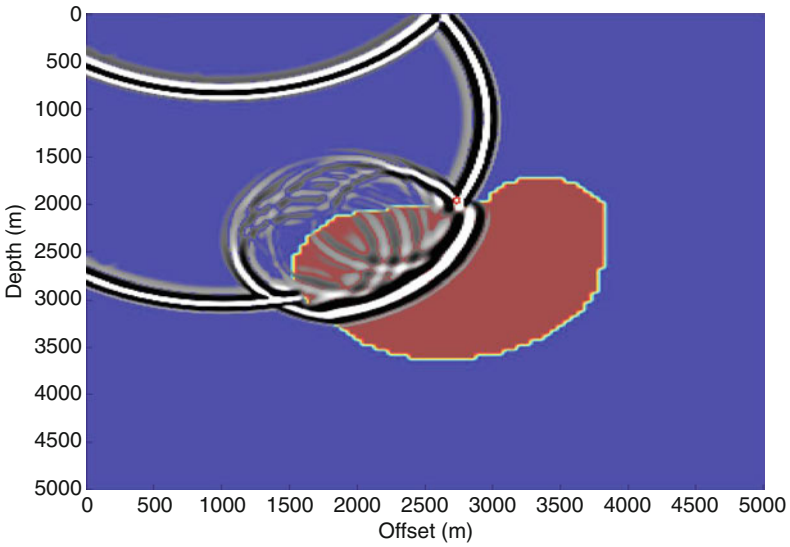
The most visible error introduced through the FD approximation is called numerical dispersion and appears when

Numerical Methods, Finite Difference, Table 1 Examples of α_i for different accuracies

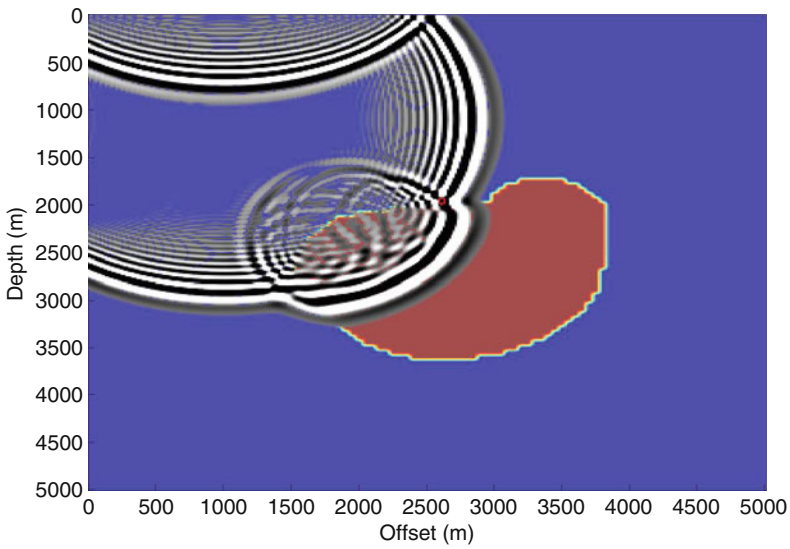
	α_i		
Order	$i=1$	$i=2$	$i=3$
2	1	N/A	N/A
4	9/8	−1/24	N/A
6	75/64	−25/384	3/640

either the sampling interval is too large or the approximation is not of sufficiently high order. Numerical dispersion is defined as the effect when different frequencies of the wavefields propagate slower or faster compared to the correct speed of wave propagation in the medium. Figures 2 and 3 show wave propagation simulations with two different order schemes using the same spatial sampling and clearly demonstrate severe effects of numerical dispersion of the lower-order scheme.

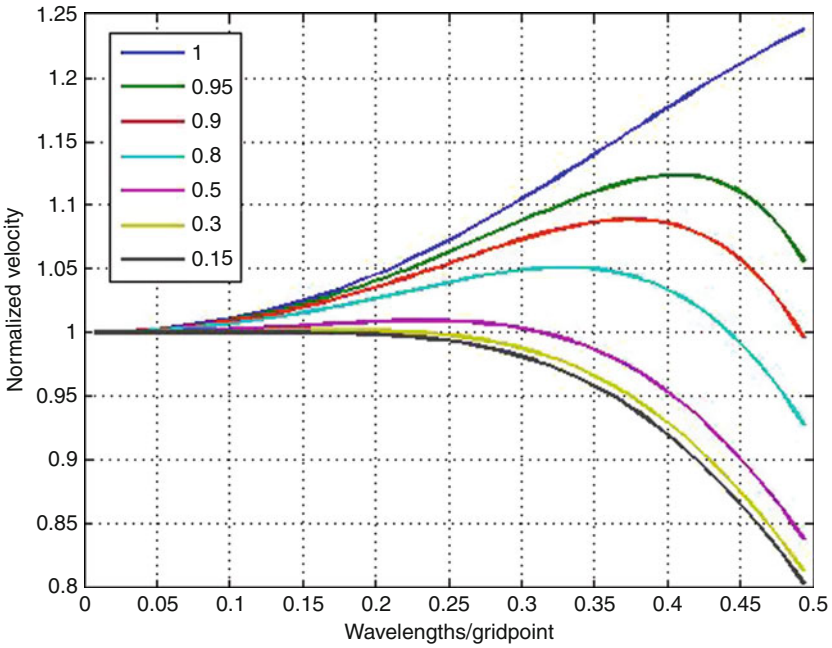
The numerical dispersion can be predicted by Fourier transforming the FD expressions in time and space. Using the Fourier-transformed expressions, we obtain a relationship that shows how the propagation velocity in the medium is affected by numerical dispersion. It is a function of several parameters including the spatial grid step and the so-called Courant number (see stability section below). Figure 4 shows several such dispersion curves (phase velocity error as a function of the number of



Numerical Methods, Finite Difference, Figure 2 Snapshot in time of a simulation using the acoustic wave equation. The black-and-white wave front indicates the location of a travelling wave. The wave has been reflected in the upper free surface and changed polarity. The left side shows the radiating boundary condition, which in this case is a PML. The blue area has slower propagation velocity than the red area. The solution has been computed using an O(2,16) scheme.



Numerical Methods, Finite Difference, Figure 3 Similar snapshot in time to that shown in Figure 2. The spatial sampling and medium is the same as in Figure 2, but in this case the simulation has been performed with an O(2,2) scheme. The trailing ripples of the wave front are typical for numerical dispersion.



Numerical Methods, Finite Difference, Figure 4 Dispersion curves for a O(2,6) FD Scheme. A perfect solution should be equal to one. The dispersion curves vary depending on the ratio of the velocity of the medium and the maximum velocity used to determine the time step, that is, the (normalized) Courant number displayed in the legend. To yield a sufficiently accurate solution, this scheme should probably not be run with less than five to seven gridpoints per minimum wavelength (0.14–0.2 wavelengths/gridpoint).

wavelengths per gridpoint) for different choices of time steps (resulting in different Courant numbers, see below). The numerical dispersion is dominated by the slowest wave speed in the medium, where the wavelength is the shortest. The number of sampling/gridpoints per wavelength necessary for a certain level of accuracy is thus determined by the dispersion curve and the slowest wave speed in a medium.

In practice, it is still difficult to define exactly how fine the wavefields must be sampled to avoid numerical dispersion as the severity of the numerical dispersion depends on the accuracy of the approximation and how far the wave has to propagate. However, as general rules of thumb, in order to roughly limit effects of numerical dispersion within a 2.5% error in propagation velocity, an O(2,4) scheme requires at least eight grid points per minimum wavelength, whereas an O(2,6) scheme requires six grid points per minimum wavelength.

It is obvious from Equation 8 that a higher-order derivative approximation requires more computations than a lower-order approximation. However, in general higher-order approximations (up to some limit) are preferred as the computational requirements increase linearly with accuracy order, whereas a finer grid increases the computational requirement to the fourth power of the refinement in 3D. The fourth power dependence originates from the fact that the temporal step must be proportional to the spatial step (see stability below). Thus, if the spatial step is reduced by a factor n, then a factor n^3 more gridpoints are required in 3D in addition to another factor n more time steps, therefore resulting in an increase of a factor n^4 in the number of computations required.

Lax–Wendroff corrections and optimally accurate FD schemes

The higher-order FD approximations are usually applied to spatial derivatives. To construct a functional scheme that has a higher order of accuracy in time, it is often necessary to use implicit time stepping, which results in a substantial increase in computational requirements. A solution to increase the formal order of accuracy (in a Taylor sense) without resorting to implicit schemes or saving several time levels is to apply a so-called Lax–Wendroff correction (Lax and Wendroff, 1964; Dablain, 1986). The Lax–Wendroff correction works by using the system of equations to express higher-order derivatives in time as spatial derivatives. The approximations of the higher-order time derivatives can be used to construct correction terms to cancel errors of successively higher order. Using the system

$$\begin{cases} \dfrac{\partial p}{\partial t} = \kappa \dfrac{\partial v}{\partial x} \\ \dfrac{\partial v}{\partial t} = \dfrac{1}{\rho} \dfrac{\partial p}{\partial x} \end{cases}, \tag{8}$$

the corresponding Lax–Wendroff correction for a third-order time derivative is:

$$\frac{\partial^3 p}{\partial t^3} = \kappa \frac{\partial^3 v}{\partial x \partial t^2} = \kappa \frac{\partial}{\partial x} \frac{1}{\rho} \frac{\partial^2 p}{\partial x \partial t} = \kappa \frac{\partial}{\partial x} \frac{1}{\rho} \frac{\partial}{\partial x} \kappa \frac{\partial v}{\partial x}. \quad (9)$$

However, the method is somewhat cumbersome in multi-dimensions as it requires mixed derivatives and also derivatives of the material parameters, that is, ρ and κ in Equations 8 and 9 (see for instance Blanch and Robertsson (1997) or Bergmann et al. (1999) for applications of the Lax–Wendroff correction).

The FD stencil can also be adjusted such that it is optimal for a certain frequency/wavelength range and thus achieves a higher accuracy than a standard scheme (Holberg, 1987; Robertsson et al., 1994b; Geller and Takeuchi, 1995; Zingg, 2000). Note that in a strict (and in this case misleading) Taylor sense, these schemes do not necessarily appear to be high-order schemes.

Time stepping and stability

As discussed above, the time stepping is most commonly explicit using a second-order approximation resulting in a leap-frog scheme. The explicit time stepping introduces a so-called CFL (Courant–Friedrich–Levy) stability condition prescribing the maximum size of the time step Δt. The CFL condition can be derived from the same Fourier transformation as described above (in the context of numerical dispersion) and requiring the resulting difference equation to have roots of absolute value less than or equal to one. As a result, the maximum time step is determined by the fastest wave speed in a medium, the spatial step, and a constant depending on the particular FD approximations used:

$$\Delta t \leq \frac{K}{\sqrt{D}} \frac{\Delta x}{c_{\max}}, \quad (10)$$

where the constant K depends on the spatial accuracy and is slowly decreasing with increasing accuracy of the scheme from a value of 1 for an O(2,2) scheme, D is the dimension of the simulation (e.g., D = 2 for 2D) and $c_{\max}$ is the maximum wave propagation speed in the medium. Rapid changes in material parameters may require an even smaller time step, but do not, in general, cause problems (Haney, 2007). Using a temporal step larger than determined by the CFL condition yields an unstable simulation, that is, parasitic solutions will grow at an exponential rate and swamp the physical solution.

The Courant number is often used in the context of discussing stability and numerical dispersion. It is defined as

$$\gamma = c_0 \frac{\Delta t}{\Delta x}, \quad (11)$$

where c_0 is the local velocity of the medium. The Courant number can be interpreted physically as the fraction of the spatial increment that a wave can advance in a time step. From Equation 11, we see that the Courant number controls the stability of the numerical scheme. In a heterogeneous medium, there exists a large range of velocities and it is thus important to study a FD scheme's behavior (i.e., numerical dispersion) for a range of Courant numbers (e.g., see Figure 4). In the literature, the normalized Courant number is often used:

$$\gamma_{norm} = c_0 \frac{\Delta t}{\Delta x} \frac{\sqrt{D}}{K}, \quad (12)$$

which needs to be less than or equal to 1 for a stable simulation.

Boundary conditions

The most important boundary conditions for wave simulations are radiating and free-surface boundary conditions. Radiating boundary conditions are applied at the edge of the finite computational domain to annihilate undesired reflections from the edge of the domain. In this fashion, we are able to emulate an infinite grid without reverberations appearing to occur at or beyond the physical boundary. The radiating boundary conditions can be implemented as "one-way" wave equation propagators using special operators at the boundaries. A first-order simple version is to implement the following equation at the boundary:

$$\frac{\partial p}{\partial t} = \pm c_0 \frac{\partial p}{\partial x}, \quad (13)$$

where the sign depends on which boundary (maximum or minimum x coordinate) and c_0 is the (local) propagation velocity. For a single propagation velocity, Equation 13 is fairly straightforward to implement. However, significant complications arise for elastic or anisotropic media and for higher dimensions than 1D where it is the apparent velocity perpendicular to the boundary that is the relevant parameter (Clayton and Engquist, 1977; Higdon, 1986, 1990).

The radiating boundary conditions are more commonly implemented as absorbing (or attenuating) boundary conditions applied in a finite region just inside the computational domain (Cerjan et al., 1985). The absorbing boundary conditions rapidly attenuate the wavefields such that any reflections from the numerical boundary are so small that they do not affect the result of the simulations. Absorbing boundaries do, however, need to have a transition zone, where the attenuation of the wavefield is gradually increased, since a rapid change in attenuation will cause reflections as well. These gradually increasingly attenuating boundaries are often referred to as sponge layers. The most efficient absorbing boundary condition is called a "Perfectly Matched Layer" (PML) (Bérenger, 1994; Gedney, 1996; Collino and Tsogka, 2001). The PML allows the attenuation to be increased extremely rapidly without causing reflections. The PML boundary conditions are implemented by introducing artificial anisotropy to the boundary and matching the reflection

coefficient in different propagation directions. The most straightforward realization of a PML boundary is through the split-field technique, where the wavefield is split depending on which spatial derivatives are used to update it. For the system in Equation 8 in two dimensions, an implementation for the p variable would be

$$\begin{cases} \dfrac{\partial p_x}{\partial t} + \beta_x p = K \dfrac{\partial v}{\partial x} \\ \dfrac{\partial p_z}{\partial t} + \beta_z p = K \dfrac{\partial v}{\partial z} \\ p = p_x + p_z \end{cases}. \tag{14}$$

The coefficients β_i control the attenuation in either the x or z directions independently.

The straightforward split-field implementation is equivalent to a fully anisotropic implementation (Teixeira et al., 2002) but suffers from a mild instability for waves impinging on the boundary at an incidence angle higher than 45°. A slightly more complex implementation avoids this instability (Bécache et al., 2002).

The free-surface boundary condition is used to approximate the interface between a water/solid material and air. Since the acoustic impedance of air is significantly different from that of water or the Earth, explicit modeling results in a computationally much more expensive simulation. In fact, it is possible to use a (pressure release) free-surface condition instead (water or Earth in contact with vacuum). For a water/air contact it is simply implemented through mirroring,

$$\begin{cases} p(-z) = -p(z) \\ v_z(-z) = v_z(z) \end{cases}, \tag{15}$$

assuming the boundary is located at $z = 0$. The upper boundary in Figure 2 is implemented using a free-surface condition. The free-surface condition for a solid elastic material is more complicated and cannot be achieved solely through mirroring. Care must be taken to ensure that all wavefield components satisfy the wave equation and the free-surface condition in order to accurately model wave phenomena such as Rayleigh waves at the free surface (see Robertsson, 1996, for details). In an anisotropic material the constitutive relationship is more complex and leads to an equation involving a larger amount of elastic constants and components of the strain tensor.

Topography and conformal mapping of grids

Topography can either be implemented explicitly into a FD scheme (Robertsson, 1996) or by stretching the grid through conformal mapping of the regular computational grid onto a grid with continuously varying grid spacings where the top of the grid follows the topography (Fornberg, 1988; Hestholm and Ruud, 1994). The explicit method will lose some of its flexibility to implement rapidly varying topography if a higher-order scheme is used for the solution. If a conformal mapping is used, all derivatives will depend on the mapping function through the chain-rule. For instance,

$$\frac{\partial v_z(f(z))}{\partial z} = \frac{\partial v_z}{\partial f}\frac{\partial f}{\partial z}. \tag{16}$$

This may lead to stricter stability conditions and stricter requirements on spatial sampling.

Source implementations

Seismic sources can be introduced in terms of body forces as indicated in Equation 1 or as moment-tensor sources using an equivalent source term in the constitutive equation (e.g., Equation 2 or 5). A straightforward approach is to simply drive a body force at one point by feeding Equation 1 with the value of a desired source wavelet in a single spatial grid point at each time step. Although this approach tends to work fairly well, researchers have sometimes reported on high-frequency noise being introduced. The origin of this noise has been attributed to the intrinsic difficulty in representing delta functions on a discretely sampled model with continuous wavefields. As a consequence, researchers have attempted to smooth the (delta-like function) source excitation in space by introducing source terms in a vicinity of grid points and scale the grid points accordingly with, for example, a Gaussian taper.

Practitioners in exploration seismology often mimic explosive sources by equal excitation of the diagonal elements of the stress tensor at the source location or alternatively model compressional or shear vibroseis source by either exciting vertical or horizontal body forces. By implication from the body-force equivalence theorem, earthquake seismologists who wish to model a moment-tensor source can do this either through a stress or body-force representation as described by Moczo et al. (2007).

An alternative way to introduce a source wavefield on a FD grid is by means of a boundary condition on an artificial surface interior to the grid. The method was first described by Alterman and Karal (1968) and followed by many authors clarifying the concept (Kelly et al., 1976; Levander, 1989; Robertsson and Chapman, 2000; Moczo et al., 2007). The technique relies on two prerequisites. First, that the spatial FD operators have a limited extent in space and second, that the (isolated) source wavefield can be computed in a region around the source location without reverberations from the rest of the model at the (simulated) time of source insertion. The artificial surface is defined enclosing a volume around the desired source location. Since the source wavefield often is computed in the vicinity of the artificial surface by some different means (e.g., an analytical solution), it is often practical to ensure that the artificial surface encloses a homogeneous medium.

The method of introducing a source wavefield as a boundary condition along an internal surface can also be used in so-called hybrid methods where two different numerical solutions are coupled together (e.g., Robertsson et al., 1996; Zahradník and Moczo, 1996; Robertsson and Chapman, 2000; van Manen et al., 2007). However, the

wavefield on opposite sides of the artificial surface in the technique above is reversed so that the source wavefield is present inside the artificial surface only. Typically, the artificial surface will be designed to be as large as possible just inside the absorbing boundary or free surface of the FD model. If the source wavefield is computed using a different numerical method such as Gaussian beams (Robertsson et al., 1996), some (limited) numerical artifacts will be introduced as the artificial surface crosses sharp discontinuities in medium properties.

Conclusions

We have outlined the basic principles of FD approximations to the systems of first-order partial-differential equations describing acoustic, isotropic elastic, anisotropic elastic, or viscoacoustic/elastic wave propagation. The method is based on explicit leap-frog time-stepping and staggered-grid representations. Most commonly Taylor-series-derived stencils are used to approximate the spatial derivatives, although other means to optimize FD stencils have also been described.

Two types of boundary conditions exist. First, radiating or absorbing boundary conditions (e.g., PML) are used to truncate the model while avoiding boundary reflections. Second, a free-surface boundary condition can be used to simulate the surface of the Earth with or without topography.

Sources are introduced either as point-force or stress sources, or along closed (artificial) surfaces inside the FD grid.

The maximum possible spatial grid size in order to limit numerical error is computed from the slowest propagation velocity in the model (e.g., a shear-wave velocity in the near-surface) using the so-called numerical dispersion relation for the FD scheme. Once the grid spacing has been chosen, the maximum possible time step is computed from the CFL-stability condition. The maximum possible time step is proportional to the spatial grid size and inversely proportional to the maximum propagation velocity in the grid. Since the computational cost of a 3D FD simulation increases to the power of four with a linear reduction of grid size, high-order accurate FD stencils are usually preferred.

Bibliography

Alterman, Z., and Karal, F. C., 1968. Propagation of elastic waves in layered media by finite-difference methods. *Bulletin. Seismological Society of America*, **58**, 367–398.

Bécache, E., Petropoulos, P., and Gedney, S., 2002. On the long-time behavior of unsplit Perfectly Matched Layers. INRIA Report de Recherche, No 4538. ftp://ftp.inria.fr/INRIA/publication/publi-pdf/RR/RR-4538.pdf

Bérenger, J. P., 1994. A Perfectly Matched Layer for the Absorption of Electromagnetic Waves. *Journal of Computational Physics*, **114**, 185–200.

Bergmann, T., Blanch, J. O., Robertsson, J. O. A., and Holliger, K., 1999. A simplified Lax–Wendroff correction for staggered-grid FDTD modeling of electromagnetic wave propagation in frequency dependent media. *Geophysics*, **64**, 1369–1377.

Bernth, H., and Chapman, C. H., 2010. A comparison of Lebedev and rotatated staggered grids for anisotropic finite difference modeling. In Extended abstracts from the 72nd EAGE Conference and Exhibition.

Blanch, J. O., and Robertsson, J. O. A., 1997. A modified Lax–Wendroff correction for wave propagation in media described by Zener elements. *Geophysical Journal International*, **131**, 381–386.

Blanch, J. O., Robertsson, J. O. A., and Symes, W. W., 1995. Modeling of constant Q: Methodology and algorithm for an efficient and optimally inexpensive viscoelastic technique. *Geophysics*, **60**, 176–184.

Cerjan, C., Kosloff, D., Kosloff, R., and Reshef, M., 1985. A nonreflecting boundary condition for discrete acoustic and elastic wave equations. *Geophysics*, **50**, 705–708.

Clayton, R. W., and Engquist, B., 1977. Absorbing boundary conditions for acoustic and elastic wave equations. *Bulletin. Seismological Society of America*, **6**, 1529–1540.

Collino, F., and Tsogka, C., 2001. Application of the PML Absorbing Layer Model to the Linear Elastodynamic Problem in Anisotropic Heterogeneous Media. *Geophysics*, **66**, 294–1307.

Dablain, M. A., 1986. The application of high-order differencing to the scalar wave equation. *Geophysics*, **51**, 54–66.

Fornberg, B., 1988. Generation of finite difference formulas on arbitrary spaced grids. *Mathematics of Computation*, **51**, 699–706.

Gedney, S. D., 1996. An Anisotropic Perfectly Matched Layer-Absorbing Medium for the Truncation of FDTD Lattices. *IEEE Transactions on Antenna and Propagation*, **44**, 1630–1639.

Geller, R. J., and Takeuchi, N., 1995. A new method for computing highly accurate DSM synthetic seismograms. *Geophysical Journal International*, **123**, 449–470.

Haney, M., 2007. Generalization of von Neumann analysis for a model of two discrete half-spaces: the acoustic case. *Geophysics*, **72**, SM35–SM46.

Hestholm, S., and Ruud, B. O., 1994. 2-D finite difference elastic wave modeling including surface topography. *Geophysical Prospecting*, **42**, 371–390.

Higdon, R. L., 1986. Absorbing boundary conditions for difference approximations to the multi-dimensional wave equation. *Mathematics of Computation*, **47**, 437–459.

Higdon, R. L., 1990. Radiating boundary conditions for elastic wave propagation. *SIAM Journal on Numerical Analysis*, **27**, 831–870.

Holberg, O., 1987. Computational aspects of the choice of operator and sampling interval for numerical differentiation in large-scale simulation of wave phenomena. *Geophysical Prospecting*, **35**, 629–655.

Igel, H., Mora, P., and Riollet, B., 1995. Anisotropic wave propagation through finite-difference grids. *Geophysics*, **60**, 1203–1216.

Kelly, K. R., Ward, R. W., Treitel, S., and Alford, R. M., 1976. Synthetic seismograms: A finite-difference approach. *Geophysics*, **41**, 2–27.

Lax, P. D., and Wendroff, B., 1964. Difference schemes for hyperbolic equations with high order of accuracy. *Communications on Pure and Applied Mathematics*, **17**, 381–398.

Lebedev, V. I., 1964. Difference analysis of orthogonal decompositions of basic differential operators and some boundary value problems I. *U.S.S.R. Comptational Mathmatics and Mathematical Physics*, **4**, 449–465 (in Russian).

Levander, A. R., 1988. Fourth-order finite-difference P-SV seismograms. *Geophysics*, **53**, 1425–1435.

Levander, A. R., 1989. Finite-difference forward modeling in seismology. In James, D. E. (ed.), *The Encyclopedia of Solid Earth Geophysics*. Van Nostrand Reinhold, New York, pp. 410–431.

Lisitsa, V., and Vishnevskiy, D., 2010. Lebedev scheme for the numerical simulation of wave propagation in 3D anisotropic elasticity. *Geophysical Prospecting*, **58**, 619–645.

Moczo, P., Robertsson, J. O. A., and Eisner, L., 2007. The finite-difference time-domain method for modelling of seismic wave

propagation. In Wu, R. S., and Maupin, V. (eds.), *Advances in wave propagation in heterogeneous Earth*, Vol. 48, *Advances in Geophysics* (ed. R. Dmowska). Oxford: Elsevier-Pergamon, pp. 421–516.
Robertsson, J. O. A., 1996. A numerical free-surface condition for elastic/viscoelastic finite-difference modeling in the presence of topography. *Geophysics*, **61**, 1921–1934.
Robertsson, J. O. A., and Chapman, C. H., 2000. An efficient method for calculating finite-difference seismograms after model alterations. *Geophysics*, **65**, 907–918.
Robertsson, J. O. A., Blanch, J. O., and Symes, W. W., 1994a. Viscoelastic finite-difference modeling. *Geophysics*, **59**, 1444–1456.
Robertsson, J. O. A., Blanch, J. O., Symes, W. W., and Burrus, C. S., 1994b. Galerkin-wavelet modeling of wave propagation: Optimal finite-difference stencil design. *Mathematical and Computer Modelling*, **19**, 31–38.
Robertsson, J. O. A., Levander, A., and Holliger, K., 1996. Modeling of the Acoustic Reverberation Special Research Program deep ocean seafloor scattering experiments using a hybrid wave propagation simulation technique. *Journal of Geophysical Research*, **101**, 3085–3101.
Saenger, E. H., 2000. Wave propagation in fracture media: theory and applications of the rotated staggered finite-difference grid. PhD thesis, Karlsruhe University.
Teixeira, F. L., Moss, C. D., Chew, W. C., and Kong, J. A., 2002. Split-field and anisotropic medium PML-FDTD implementations for inhomogeneous media. *IEEE Transactions on Microwave Theory*, **50**, 30–35.
van Manen, D. J., Robertsson, J. O. A., and Curtis, A., 2007. Exact wave field simulation for finite-volume scattering problems. *The Journal of the Acoustical Society of America*, **122**, EL115–EL121.
Virieux, J., 1984. SH-wave propagation in heterogeneous media: Velocity-stress finite-difference method. *Geophysics*, **49**, 1933–1957.
Virieux, J., 1986. P-SV wave propagation in heterogeneous media: Velocity-stress finite-difference method. *Geophysics*, **51**, 889–901.
Zahradník, J., and Moczo, P., 1996. Hybrid seismic modeling based on discrete-wavenumber and finite-difference methods. *Pure and Applied Geophysics*, **148**, 21–38.
Zingg, D. W., 2000. Comparison of high-accuracy finite-difference methods for linear wave propagation. *SIAM Journal of Scientific Computing*, **22**, 476–502.

Cross-references

Earthquakes, Source Theory
Propagation of Elastic Waves: Fundamentals
Seismic Anisotropy
Seismic Imaging, Overview
Seismic, Migration
Seismic, Viscoelastic Attenuation
Seismic, Waveform Modeling and Tomography

NUMERICAL METHODS, FINITE ELEMENT

J. N. Reddy
Department of Mechanical Engineering, Texas A&M University, College Station, TX, USA

Definition

The finite element method is a numerical method, like the finite difference method, for solving differential equations arising in the study of physical phenomena. In the finite element method, a given domain is viewed as a set of nonintersecting subdomains, called *finite elements*, and over each element the governing equation is approximated by any of the traditional variational methods (e.g., Ritz, Galerkin, and least-squares methods). The main reason behind seeking approximate solution on a set of elements is the fact that it is easier to represent a complicated function as a sum of simple polynomials. Of course, each individual segment of the solution should fit with its neighbors in the sense that the function and possibly its derivatives up to a chosen order are continuous along the interface between elements. This entry is a brief introduction to the finite element modeling of diffusion processes and the Navier–Stokes equations governing viscous incompressible fluids. Both phenomena play crucial role in modeling a variety of geological or geomechanics processes.

General introduction

Scientists and engineers model (i.e., develop mathematical models and numerically simulate) natural phenomena with the objective of understanding it. Virtually every phenomenon in nature, whether aerospace, biological, chemical, geological, or mechanical, can be described in terms of algebraic, differential, and/or integral equations relating various quantities of interest. Determining the stress distribution in a mechanical structure with oddly shaped members and numerous stiffeners and subjected to mechanical, thermal, and/or aerodynamic loads, finding the concentration of pollutants in lakes and estuaries or in the atmosphere, predicting geophysical and geological events, and simulating weather in an attempt to predict the formation of tornadoes and thunderstorms are a few examples of many important practical problems that scientists and engineers are occupied with.

In the last four decades, the finite element method has emerged as a powerful computational tool for solving nonlinear partial differential equations over irregular domains with complex domain properties (Reddy, 2003, 2006; Reddy and Gartling, 2010; Bathe, 1996; Belytschko et al., 2000). Applications of the method to geology, geophysics, and geomechanics problems are numerous (see, e.g., Reddy et al., 1982; Wickham et al., 1982; Bird, 1989; Parsons, 2002; Dyksterhuis et al., 2005 and references therein). This brief introduction is meant to provide some insight into the workings of the finite element method as applied to the Poisson equation as well as the Navier–Stokes equations that are used to model certain geomechanics processes in two dimensions.

Finite element model of 2-D Poisson equation

Consider the problem of finding $u(x, y)$ such that the following partial differential equation (a generalized Poisson's equation) is satisfied

$$-\left[\frac{\partial}{\partial x}\left(a_{xx}\frac{\partial u}{\partial x}\right)+\frac{\partial}{\partial y}\left(a_{yy}\frac{\partial u}{\partial y}\right)\right]=f(x,y)\text{ in }\Omega \tag{1}$$

where Ω is a two-dimensional domain with boundary Γ. Here a_{xx} and a_{yy} are material coefficients in the x and y directions, respectively, and $f(x,y)$ is the known source. For example, in a ground water flow problem u denotes the water head (i.e., velocity potential), a_{xx} and a_{yy} are the permeabilities in the x and y directions, respectively, and $f(x,y)$ is distributed water source. Equation 1 also arises in many areas of science and engineering.

In the finite element method, the domain $\bar{\Omega}=\Omega\cup\Gamma$ is divided into a set of subdomains $\bar{\Omega}^e=\Omega^e\cup\Gamma^e$, called finite elements. Any geometric shape qualifies as an element, provided that the approximation functions ψ_i^e can be derived uniquely for the shape.

Suppose that the dependent unknown u is approximated over a typical finite element $\bar{\Omega}^e$ by the expression

$$u(\mathbf{x})\approx u_h^e(\mathbf{x})=\sum_{j=1}^{n}u_j^e\psi_j^e(\mathbf{x}),\ \mathbf{x}=(x,y)\in\bar{\Omega}^e \tag{2}$$

where $u_h^e(\mathbf{x})$ represents an approximation of $u(\mathbf{x})$ over the element $\bar{\Omega}^e$, parameters u_j^e denote the values of the function $u_h^e(\mathbf{x})$ at a selected number of points (i.e., element nodes) in the element $\bar{\Omega}^e$, and ψ_j^e are the Lagrange interpolation functions associated with the element.

We seek to satisfy the governing differential equation in a weak-form sense, with the weight functions being the same as the approximation functions. The resulting finite element model, i.e., set of algebraic equations) is called the *weak-form Galerkin finite element model* (Reddy, 2006). The weak form is

$$\begin{aligned}0=&\int_{\Omega^e}\left(a_{xx}\frac{\partial w_i^e}{\partial x}\frac{\partial u_h^e}{\partial x}+a_{yy}\frac{\partial w_i^e}{\partial y}\frac{\partial u_h^e}{\partial y}-w_i^e f\right)\mathrm{d}x\mathrm{d}y\\&-\oint_{\Gamma^e}w_i^e q_n\mathrm{d}s\end{aligned} \tag{3}$$

Suppose that u_h^e is represented over a typical finite element Ω^e by expression of the form 2. Substituting the finite element approximation 2 into the weak form 3, we obtain

$$\mathbf{K}^e\mathbf{u}^e=\mathbf{f}^e+\mathbf{q}^e\equiv\mathbf{F}^e \tag{4}$$

where the coefficients K_{ij}^e, f_i^e, and q_i^e are defined by

$$\begin{aligned}K_{ij}^e=&\int_{\Omega^e}\left(a_{xx}\frac{\partial\psi_i^e}{\partial x}\frac{\partial\psi_j^e}{\partial x}+a_{yy}\frac{\partial\psi_i^e}{\partial y}\frac{\partial\psi_j^e}{\partial y}\right)\mathrm{d}A\\&+\oint_{\Gamma^e}h_c\psi_i^e\psi_j^e\mathrm{d}s\\f_i^e=&\int_{\Omega^e}f\psi_i^e\,dA,\ q_i^e=\oint_{\Gamma^e}(\hat{q}_n+h_c u_c)\psi_i^e\mathrm{d}s\end{aligned} \tag{5}$$

We note that $K_{ij}^e=K_{ji}^e$ (i.e., $\mathbf{K}^e$ is symmetric).

Finite element models of the 2-D Navier–Stokes equations

In this section, we develop the finite element models of steady flows of viscous, incompressible fluids in two-dimensional domains. The governing equations are the conservation of linear momentum and conservation of mass, expressed in terms of the Cartesian components

$$\rho\left(u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}\right)-\frac{\partial\sigma_{xx}}{\partial x}-\frac{\partial\sigma_{xy}}{\partial y}-f_x=0 \tag{6}$$

$$\rho\left(u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}\right)-\frac{\partial\sigma_{xy}}{\partial x}-\frac{\partial\sigma_{yy}}{\partial y}-f_y=0 \tag{7}$$

$$\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}=0 \tag{8}$$

$$\begin{aligned}&\sigma_{xx}=2\mu\frac{\partial u}{\partial x}-P,\ \ \sigma_{yy}=2\mu\frac{\partial v}{\partial y}-P,\\&\sigma_{xy}=\mu\left(\frac{\partial u}{\partial y}+\frac{\partial v}{\partial x}\right)\end{aligned} \tag{9}$$

Here we present two different finite element models associated with Equations 6–9. The first one is a direct formulation in which the three equations in (u,v,P) are used in their original form. This formulation is known as the *velocity–pressure formulation*. The other formulation is based on the interpretation that the continuity equation (Equation 8) is a constraint on u and v, and the constraint is satisfied in a least-squares (i.e., approximate) sense. This particular method of including the constraint in the formulation is known as the penalty function method, and the model is termed as the *penalty-finite element model*. It is informative to note that the velocity–pressure formulation is the same as the Lagrange multiplier formulation, wherein the constraint is included by means of the Lagrange multiplier. The Lagrange multiplier turns out to be the negative of the pressure.

Velocity–pressure (mixed) model

The weak forms of Equations 6–8 over an element Ω^e can be constructed following the standard procedure (Reddy, 2003). The weight functions have the following physical interpretation:

$$w_1\sim u\ ,\ w_2\sim v\quad\text{and}\quad w_3\sim -P \tag{10}$$

Assuming approximations of the form

$$u=\sum_{j=1}^{m}v_j\psi_j^e,\quad v=\sum_{j=1}^{m}v_j\psi_j^e,\quad P=\sum_{j=1}^{n}P_j\phi_j^e \tag{11}$$

where ψ_j^e and ϕ_j^e denote the Lagrange type interpolation functions $(n<m;\quad n+1=m)$ and substituting into the weak forms of Equations 6–8, we obtain the following finite element model:

$$\begin{bmatrix} \mathbf{K}^{11} & \mathbf{K}^{12} & \mathbf{K}^{13} \\ \mathbf{K}^{21} & \mathbf{K}^{22} & \mathbf{K}^{23} \\ \mathbf{K}^{31} & \mathbf{K}^{32} & \mathbf{K}^{33} \end{bmatrix} \begin{Bmatrix} \mathbf{u} \\ \mathbf{v} \\ \mathbf{P} \end{Bmatrix} = \begin{Bmatrix} \mathbf{F}^1 \\ \mathbf{F}^2 \\ \mathbf{F}^3 \end{Bmatrix} \tag{12}$$

where $F_i^3 = 0$

$$\begin{aligned}
K_{ij}^{11} &= \int_{\Omega^e} \left[\mu \left(2 \frac{\partial \psi_i}{\partial x} \frac{\partial \psi_j}{\partial x} + \frac{\partial \psi_i}{\partial y} \frac{\partial \psi_j}{\partial y} \right) \right. \\
&\quad \left. + \rho \psi_i \left(u \frac{\partial \psi_j}{\partial x} + v \frac{\partial \psi_j}{\partial y} \right) \right] \mathrm{d}x\mathrm{d}y \\
K_{ij}^{12} &= \int_{\Omega^e} \mu \frac{\partial \psi_i}{\partial y} \frac{\partial \psi_j}{\partial x} \mathrm{d}x\mathrm{d}y, \\
K_{ij}^{13} &= -\int_{\Omega^e} \frac{\partial \psi_i}{\partial x} \psi_j \, \mathrm{d}x\mathrm{d}y, \quad K_{ij}^{33} = 0 \\
K_{ij}^{22} &= \int_{\Omega^e} \left[\mu \left(\frac{\partial \psi_i}{\partial x} \frac{\partial \psi_j}{\partial x} + 2 \frac{\partial \psi_i}{\partial y} \frac{\partial \psi_j}{\partial y} \right) \right. \\
&\quad \left. + \rho \psi_i \left(u \frac{\partial \psi_j}{\partial x} + v \frac{\partial \psi_j}{\partial y} \right) \right] \mathrm{d}x\mathrm{d}y \\
K_{ij}^{23} &= -\int_{\Omega^e} \frac{\partial \psi_i}{\partial y} \psi_j \mathrm{d}x\mathrm{d}y, \quad K_{ij}^{21} = K_{ji}^{21}, \\
K_{ij}^{31} &= K_{ji}^{13}, \quad K_{ij}^{32} = K_{ji}^{23} \\
F_i^1 &= \int_{\Omega^e} f_x \psi_i \mathrm{d}x\mathrm{d}y + \oint_{\Gamma^e} t_x \psi_i \mathrm{d}s, \\
F_i^2 &= \int_{\Omega^e} f_y \psi_i \mathrm{d}x\mathrm{d}y + \oint_{\Gamma^e} t_y \psi_i \mathrm{d}s
\end{aligned} \tag{13}$$

Here (t_x, t_y) denote the components of the stress vector $\mathbf{t}$ on the boundary

$$t_x = \sigma_{xx} n_x + \sigma_{xy} n_y, \quad t_y = \sigma_{xy} n_x + \sigma_{yy} n_y \tag{14}$$

and (n_x, n_y) are the components of the unit normal vector $\hat{\mathbf{n}}$.

Penalty-finite element model

Use of the penalty function method amounts to replacing the pressure with

$$P = -\gamma \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right) \tag{15}$$

where γ is known as the penalty parameter. For complete details, the reader may consult (Reddy, 2003, 2006; Reddy and Gartling, 2010). Note that Equation 15 is used to post-compute P once the velocity field is available.

Assuming interpolation of the form

$$u = \sum_{j=1}^{n} u_j^e \psi_j^e, \quad v = \sum_{j=1}^{n} v_j^e \psi_j^e \tag{16}$$

where ψ_i^e are Lagrange interpolation functions. Substituting Equation 16 into the weak forms of Equations 6 and 7, we obtain the finite element model

$$\begin{bmatrix} \mathbf{K}^{11} & \mathbf{K}^{12} \\ \mathbf{K}^{21} & \mathbf{K}^{22} \end{bmatrix} \begin{Bmatrix} \mathbf{u} \\ \mathbf{v} \end{Bmatrix} = \begin{Bmatrix} \mathbf{F}^1 \\ \mathbf{F}^2 \end{Bmatrix} \tag{17}$$

where

$$\begin{aligned}
K_{ij}^{11} &= \int_{\Omega^e} \left[\mu \left(2 \frac{\partial \psi_i^e}{\partial x} \frac{\partial \psi_j^e}{\partial x} + \frac{\partial \psi_i^e}{\partial y} \frac{\partial \psi_j^e}{\partial y} \right) \right. \\
&\quad \left. + \rho \psi_i^e \left(u \frac{\partial \psi_j^e}{\partial x} + v \frac{\partial \psi_j^e}{\partial y} \right) \right] \mathrm{d}x\mathrm{d}y + \int_{\Omega^e} \gamma \frac{\partial \psi_i^e}{\partial x} \frac{\partial \psi_j^e}{\partial x} \, \mathrm{d}x\mathrm{d}y \\
K_{ij}^{12} &= \int_{\Omega^e} \mu \frac{\partial \psi_i}{\partial y} \frac{\partial \psi_j}{\partial x} \mathrm{d}x\mathrm{d}y + \int_{\Omega^e} \gamma \frac{\partial \psi_i^e}{\partial x} \frac{\partial \psi_j^e}{\partial y} \, \mathrm{d}x\mathrm{d}y \\
K_{ij}^{22} &= \int_{\Omega^e} \left[\mu \left(\frac{\partial \psi_i}{\partial x} \frac{\partial \psi_j}{\partial x} + 2 \frac{\partial \psi_i}{\partial y} \frac{\partial \psi_j}{\partial y} \right) \right. \\
&\quad \left. + \rho \psi_i \left(u \frac{\partial \psi_j}{\partial x} + v \frac{\partial \psi_j}{\partial y} \right) \right] \mathrm{d}x\mathrm{d}y + \int_{\Omega^e} \gamma \frac{\partial \psi_i^e}{\partial y} \frac{\partial \psi_j^e}{\partial y} \, \mathrm{d}x\mathrm{d}y \\
F_i^1 &= \int_{\Omega^e} f_x \psi_i^e \mathrm{d}x\mathrm{d}y + \oint_{\Gamma^e} t_x \psi_i^e \mathrm{d}s, \\
F_i^2 &= \int_{\Omega^e} f_y \psi_i^e \mathrm{d}x\mathrm{d}y + \oint_{\Gamma^e} t_y \psi_i^e \mathrm{d}s
\end{aligned} \tag{18}$$

The numerical evaluation of the coefficient matrices appearing in Equation 17 requires special consideration. Equation 17 is of the general form

$$(\mathbf{K}^{\mu} + \mathbf{K}^{\rho} + \mathbf{K}^{\gamma}) \Delta = \mathbf{F} \tag{19}$$

where $\mathbf{K}^{\mu}$, $\mathbf{K}^{\rho}$, and $\mathbf{K}^{\gamma}$ denote the contributions from the viscous, inertia, and penalty terms, respectively. In theory, as we increase the value of the penalty parameter γ, the conservation of mass is satisfied more exactly. However, in practice, for some large value of γ, the contribution from the viscous and inertia terms would be negligibly small compared to the penalty terms in the computer. Thus, if $\mathbf{K}^3$ is a nonsingular (i.e., invertible) matrix, the solution of the final equations associated with Equation 19 for a large value of γ is trivial

$$\lim \gamma \to 0 (\mathbf{K}^{\mu} + \mathbf{K}^{\rho} + \mathbf{K}^{\gamma}) \Delta = \mathbf{F} \to \mathbf{K}^{\gamma} \Delta = \frac{1}{\gamma} \mathbf{F} \tag{20}$$

which will yield $\Delta = 0$. While the trivial solution satisfies the continuity equation, it does not satisfy the momentum equations for nontrivial boundary data or body forces. In this case, the discrete problem (Equation 17) is said to be over-constrained or "locked." If $\mathbf{K}^{\gamma}$ is singular, then the sum $\mathbf{K}^{\mu} + \mathbf{K}^{\rho} + \mathbf{K}^{\gamma}$ is nonsingular (because $\mathbf{K}^{\mu}$ is nonsingular after the imposition of proper boundary conditions), and a nontrivial solution to the problem may be obtained.

The numerical problem described above is eliminated by proper evaluation of the integrals in $\mathbf{K}^{\gamma}$. It is found that if the coefficients of $\mathbf{K}^{\gamma}$ (i.e., penalty terms) are evaluated using a numerical integration rule of one order less than

that required to integrate them exactly, the finite element equations (Equation 17) will give acceptable solutions for the velocity field. This technique of under-integrating the penalty terms is known in the literature as *reduced (order) integration*. For example, if a linear rectangular element is used to approximate the velocity field in a two-dimensional problem, the matrix coefficients $\mathbf{K}^{\mu}$ as well as $\mathbf{K}^{\rho}$ are evaluated using the 2×2 Gauss quadrature, and $\mathbf{K}^{\gamma}$ are evaluated using the one-point 1×1) Gauss quadrature. The one-point quadrature yields a singular $\mathbf{K}^{\gamma}$. Therefore, Equation 17 cannot be inverted, whereas $\mathbf{K}^{\mu} + \mathbf{K}^{\rho} + \mathbf{K}^{\gamma}$ is nonsingular and can be inverted (after assembly and imposition of boundary conditions) to obtain a good finite element solution of the original problem. When a quadratic rectangular element is used, the 3×3 Gauss quadrature is used to evaluate $\mathbf{K}^{\mu}$ and $\mathbf{K}^{\rho}$, and the 2×2 Gauss quadrature is used to evaluate $\mathbf{K}^{\gamma}$.

The choice of the penalty parameter is largely dictated by the ratio of the magnitude of penalty terms to the viscous and convective terms (or compared to the Reynolds number, *Re*), the mesh, and the word length in the computer. The following range of γ is suggested in computations

$$\gamma = 10^4 R_e \text{ to } \gamma = 10^{12} R_e.$$

Summary

Numerical simulation of geomechanical processes requires a good understanding of computational fluid mechanics, heat transfer, and solid mechanics and their couplings. The increase in computing power in both single processor and parallel environments has allowed realistic geomechanics problems of significant complexity and fidelity to be routinely solved and utilized in technological advances. Commercial software has made rapid progress in providing a broad spectrum of analysis capabilities to a variety of industries. Though software is increasingly robust, accurate simulations still require a knowledgeable user, with a background in both mechanics and numerical methods. This entry only provides an introduction to an individual who is interested in the use of the finite element method as a numerical simulation tool for the study and understanding of geomechanical phenomena. The Poisson equation and the Navier–Stokes equations visited here provide the necessary background for the study of diffusion processes and viscous flow problems. Interested readers may consult the references listed.

Acknowledgment

The author gratefully acknowledges the support provided by the Oscar S. Wyatt Endowed Chair.

Bibliography

Bathe, K. J., 1996. *Finite Element Procedures*. Englewood Cliffs: Prentice Hall.

Belytschko, T., Liu, W. K., and Moran, B., 2000. *Nonlinear Finite Elements for Continua and Structures*. Chichester: Wiley.

Bird, P., 1989. New finite element techniques for modeling deformation histories of continents with stratified temperature-dependent rheology. *Journal of Geophysical Research*, **94**(B4), 3967–3990.

Dyksterhuis, S., et al., 2005. Finite-element modelling of contemporary and palaeo-intraplate stress using ABAQUS. *Computers and Geosciences*, **31**(3), 297–307.

Parsons T., 2002. Post-1906 stress recovery of the San Andreas fault system calculated from three-dimensional finite element analysis. *Journal of Geophysical Research* (B8), doi:10.1029/2001JB001051.

Reddy, J. N., 2003. *An Introduction to Nonlinear Finite Element Analysis*. Oxford: Oxford University Press.

Reddy, J. N., 2006. *An Introduction to the Finite Element Method*, 3rd edn. New York: McGraw-Hill.

Reddy, J. N., and Gartling, D. K., 2010. *The Finite Element Method in Heat Transfer and Fluid Dynamics*, 3rd edn. Boca Raton: CRC Press.

Reddy, J. N., Stein, R. J., and Wickham, J. S., 1982. Finite-element modeling of folding and faulting. *International Journal for Numerical and Analytical Methods in Geomechanics*, **6**, 425–440.

Wickham, J. S., Tapp, G. S., and Reddy, J. N., 1982. Finite-element modeling of fracture density in single layer folds. *International Journal for Numerical and Analytical Methods in Geomechanics*, **6**, 441–459.

Cross-references

Numerical Methods, Boundary Element
Numerical Methods, Finite Difference
Numerical Methods, Multigrid

NUMERICAL METHODS, MULTIGRID

Wim A. Mulder
Department of Geotechnology, Delft University of Technology, Delft, The Netherlands

Definition

A multigrid method is an algorithm for the iterative solution of partial differential equations using a sequence of discretizations on multiple scales.

Introduction

The numerical solution of a partial differential equation (PDE) requires its discretization and a method to solve the resulting large system of algebraic equations. For linear equations, the resulting system is often a sparse matrix and a direct solution method suffices if the size of the problem is modest. For problems in three space dimensions, the computational cost of a direct solver may be too large. An iterative method that improves the accuracy of an approximate solution step by step can be a good alternative.

The multigrid method has optimal complexity: the amount of work required to solve a problem with N unknowns is $O(N)$, meaning that it scales with N. It achieves its efficiency by employing several discretization grids for the same problem.

History

Fedorenko (1964) introduced the multigrid method as an iterative scheme for solving Poisson's equation on a square and showed that the number of computations required to determine a solution with a prescribed accuracy is proportional to the number of unknowns, N. Therefore, the method has an optimal computational complexity. Brandt (1973) found that the actual computational cost for a sufficiently accurate result was about 10 work units, where a work unit is the cost of evaluating the discretized equations. He connected the method to local adaptive grid refinement and introduced nonlinear multigrid. Hackbush (1976, 1985) discovered the method independently and provided a mathematical foundation. Since then, the method was developed further to handle PDEs other than Poisson's, which is elliptic, and to go beyond PDEs.

Textbooks include those by Hackbush (1985), who includes convergence proofs; by Wesseling (1991), with chapters on computational fluid dynamics; by Briggs et al. (2000), an easy to read introduction; and by Trottenberg et al. (2001), among others.

Two-grid scheme

A discrete representation of a PDE provides an algebraic system of equations with relations between solution values at neighboring grid points. It is fairly easy to make local corrections that reduce the solution error on a short range, but much harder to correct the long-range or long-wavelength components of the solution that have a more global character. By projecting the solution onto a coarser grid, the long wavelengths become shorter and can effectively be solved for. Combining the corrections to the solution from coarser and finer grids yields an efficient solver.

A simple 1D example can help to understand the fundamentals of the multigrid method. The PDE is $Lu = f$, with $u(x)$ the unknown solution as a function of position x on a finite interval $[x_{min}, x_{max}]$ on the real axis, $f(x)$ a source term or forcing function, and L a linear differential operator, for instance, minus the laplacian, which in 1D is $-\frac{d^2}{dx^2}$. Dirichlet boundary conditions let the solution be zero at the endpoints of the interval. To obtain a discrete representation of the problem, an equidistant 1D grid is defined with grid points $x_k = x_{min} + kh$, index $k = 0, \ldots, N+1$, where $N+1 = 2^M$ and M is a positive integer. The grid spacing is $h = (x_{max} - x_{min})/(N+1)$. A standard second-order finite-difference scheme leads to

$$\frac{-u_{k-1} + 2u_k - u_{k+1}}{h^2} = f_k, \quad k = 1, \ldots, N,$$

where u_k approximates $u(x_k)$ and $f_k = f(x_k)$. At the boundaries, $u_0 = u_{N+1} = 0$. The discrete equations represent the problem $L^h u^h = f^h$, where L^h is a $N \times N$ sparse matrix and u^h and f^h are vectors of length N. The residual is defined as $r^h = f^h - L^h u^h$ and should vanish once the numerical solution has been found.

A simple iterative method is Jacobi relaxation, in which the matrix L^h is replaced by its diagonal $D^h = 2/h^2$. One step of Jacobi relaxation amounts to

$$u^h := u^h + \omega (D^h)^{-1} r^h.$$

The symbol " := " indicates that the solution values are replaced by the expression on the right-hand side. The factor ω controls the amount of damping. Convergence requires $0 < \omega \leq 1$. Fourier analysis (Hackbush, 1985) shows that the convergence rate, the factor by which the norm of the difference between the current and the exact numerical solution is reduced per iteration, is $1 - O(h^2)$ for Jacobi relaxation. Convergence slows down for increasingly finer grids.

The Fourier analysis reveals that the slow convergence is caused by the long-wave components of the solution. Because long waves can be represented on coarser grids, it makes sense to have a look at a two-grid scheme, outlined in Figure 1 and its caption. The coarse grid consists of every other grid point of the fine grid: $x_K = x_0 + KH$, with $H = 2h$ and $K = 0, \ldots, \frac{1}{2}(N+1)$. A restriction operator $\tilde{I}_h^H$ maps the current fine-grid residual r^h to the coarser grid: $r^H = \tilde{I}_h^H r^h$. The simplest restriction operator is injection: $r_K^H = r_{2K}^h$, $K = 1, \ldots, \frac{1}{2}(N-1)$. Full weighting, which lets $r_K^H = \frac{1}{4} r_{2K-1}^h + \frac{1}{2} r_{2K}^h + \frac{1}{4} r_{2K+1}^h$, applies some smoothing to the residual and is more common. The exact solution of the coarse-grid problem yields the coarse-grid correction $v^H = (L^H)^{-1} r^H$, which should be interpolated back to the fine grid and added as a correction to the latest fine-grid solution. The interpolation or prolongation operator is denoted by I_H^h and lets $u^h := u^h + I_H^h v^H$.

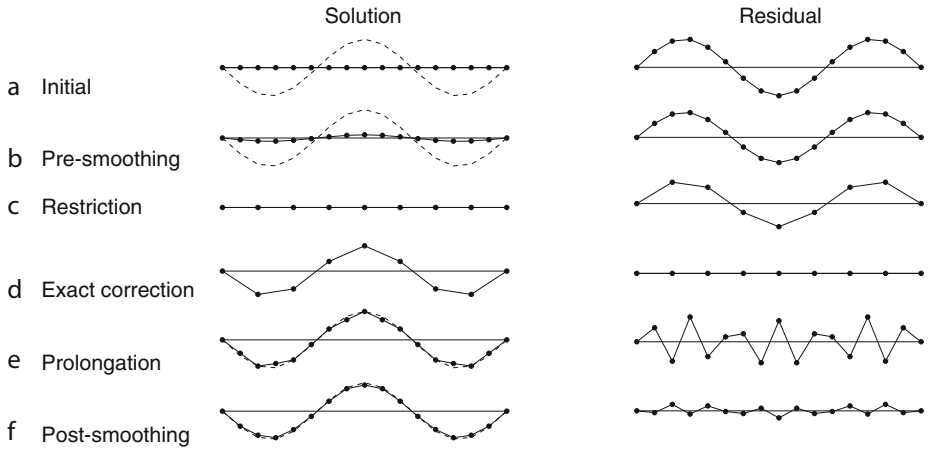
If linear interpolation is used for prolongation, this becomes

$$u_{2K}^h := u_{2K}^h + v_K^H, \quad K = 1, \ldots, \frac{1}{2}(N-1),$$

$$u_{2K-1}^h := u_{2K-1}^h + \frac{1}{2}(v_{K-1}^H + v_K^H),$$

$$K = 1, \ldots, \frac{1}{2}(N+1).$$

Here, it is assumed that $v_0^H = v_{(N+1)/2}^H = 0$. After prolongation, an additional relaxation step with damped Jacobi further removes oscillatory error components of the solution. Jacobi relaxation inside this two-grid scheme has a different purpose than when used as an iterative solver by itself. Instead of removing both the short- and long-wave components of the numerical solution error, it only has to deal with those components that cannot be represented on the coarser grid without aliasing. Therefore, a damped version with $\omega < 1$ can be more effective as it can be optimized to remove the short-wave or



Numerical Methods, Multigrid, Figure 1 Steps in a two-grid cycle. (**a**) The initial solution on the fine grid is set to zero and the residual equals the forcing function. The dashed line represents the exact numerical solution. (**b**) After one step of pre-smoothing with damped Jacobi relaxation, the solution error is still large and the residual has hardly changed in this example. Its restriction to the coarser grid (**c**) is solved exactly, using a coarse-grid version of the discrete differential operator. The resulting correction to the fine-grid solution (**d**) is interpolated or prolongated back to the fine grid and added to the fine-grid solution (**e**). The error in the solution is now dominated by the short wavelengths and appears as an oscillatory function, which is reflected in the corresponding residual. A post-smoothing step (**f**) removes most of the solution error. Repeating the whole cycle will further reduce the error.

oscillatory components. For Figure 1, $\omega = 2/3$ was used. The optimal choice requires a more detailed analysis (Hackbush, 1985).

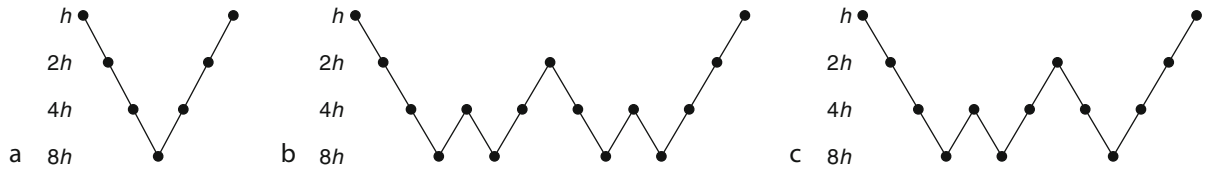
A relaxation scheme geared toward removing oscillatory components is called a smoother. Another popular choice is Gauss–Seidel relaxation. The operator is then approximation by a lower or upper triangular matrix, which is easy to invert. The implementation of the scheme is similar to Jacobi, but the residual is evaluated with the most recent solution available. The result will depend on the order in which the grid is traversed. With lexicographic Gauss–Seidel, one follows the natural index k in increasing order, or in the opposite direction. Symmetric Gauss–Seidel performs both these smoothing steps in sequence. An alternative is red-black Gauss–Seidel, where first the points with an odd and then those with an even index are updated, always using the latest solution values for the residual. In 2D, this would follow a checkerboard pattern.

The grid transfer operator, restriction and prolongation, should obey $m_p + m_r > 2m$ (Hackbush, 1985), where $2m$ is the order of the differential equation and $m_p - 1$ is the highest degree of the polynomial that is interpolated exactly. The scaled adjoint of the restriction operator is an interpolation operator for which m_r can be defined in the same way as m_p. In the example above, $2m = 2$, full weighting has $m_r = 2$, and prolongation based on linear interpolation has $m_p = 2$. More advanced grid transfer operators are based on the differential operator L^h, leading to operator-weighted restriction and prolongation.

The coarse-grid operator L^H can be based on the same discretization as the fine-grid operator L^h. An alternative is the Galerkin coarse-grid approximation $\tilde{I}_h^H L^h I_H^h$. In the current example with full weighting, these happen to be the same. Operator-weighted grid transfer operators can accelerate the convergence by using the coefficients of the differential operator in the construction of restriction and prolongation operators, which in turn will affect the Galerkin coarse-grid approximation of the differential operator.

Multigrid

Instead of using the exact numerical solution on the coarser of the two grids, a two-grid scheme can be applied to the coarse-grid problem. Extending this recursively to a coarsest grid with three points and one unknown, a multigrid solver is obtained. All these grids can be visited in different orders. The standard approach is to start on the finest grid, perform a number of pre-smoothing steps, then move to the coarser grid, again perform pre-smoothing, and continue this until the coarsest grid is reached. There, the exact numerical solution is computed and the resulting coarse-grid correction is prolongated back to the finer level, followed by a number of post-smoothing steps. This is repeated up to the finest grid. Such a sequence is called a V-cycle. A single V-cycle may be insufficient to obtain a convergence rate on the coarser grid that is similar to the smoothing rate on the finer grid. Convergence on coarser grids can be improved by performing more than one cycle. In the W-cycle, the number of cycles doubles on each coarser grid. Figure 2 illustrates the order in which grids are visited for a V-, W-, and F-cycle. The last one is less expensive than a W-cycle, as the number of cycles increases by only one for increasingly coarser grids. Another option, adaptive



Numerical Methods, Multigrid, Figure 2 (**a**) V-cycle, (**b**) W-cycle, and (**c**) F-cycle.

cycling, decides on the sequence of grids by monitoring the decrease of the residuals.

With a proper choice of relaxation scheme or smoother and grid transfer operators, the multigrid method can reach a grid-independent or h-independent convergence rate. This does not, however, mean that the amount of work required to reach a sufficiently accurate solution is proportional to the number of grid points N. The exact numerical solution differs from the true solution of the PDE because of the numerical error due to the discretization, typically with some power p of h. In the earlier example, the discretization is second-order accurate, so $p = 2$. Given this numerical error, convergence to machine precision is more than needed. An iteration error somewhat smaller than the discretization error should suffice. Therefore, more iterations should be carried out if a more accurate solution on a finer grid has to be computed. This leads to an overall complexity of $O(N\log N)$ rather than $O(N)$. Successive grid refinement, in which one first computes a solution on a coarse grid and uses that as an initial guess for the next finer grid and so on, enables removal of the factor $\log N$, assuming that a fixed number of multigrid cycles are used at each level. The combination of successive grid refinement with a multigrid solver is called Full Multigrid (FMG).

Nonlinear multigrid

One way to apply the multigrid method to nonlinear PDEs is the use of Newton's method. A multigrid solver is then applied to a linearization of the discretized nonlinear problem. An example is the computation of galactic gas flow (Mulder, 1986). Another approach is the use of the discretized nonlinear PDE inside the multigrid algorithm. This requires the full solution to be available on the coarser grids. In the example above, only corrections to the solution were represented on coarser grids. The Full Approximation Scheme (FAS) is a reformulation of the multigrid method that includes the full solution (Brandt, 1982). It requires a restriction of the full solution, $u^H = I_h^H u^h$, and of the coarse-grid forcing function, $f^H = L^H(I_h^H u^h) + \tilde{I}_h^H r^h$. The prolongation needs to be changed to $u^h := u^h + I_H^h(u^H - I_h^H u^h)$, where u^H is the sufficiently converged solution on the coarser grid and $I_h^H u^h$ the restriction of the latest fine-grid solution. The restriction operators for the residuals, $\tilde{I}_h^H$, and for the solution, I_h^H, do not have to be the same. Note that the coarse-grid problem, $L^H u^H = f^H$, can be interpreted in a different way: $L^H u^H = f^H = \tilde{I}_h^H f^h + \tau_h^H$. The fine-to-coarse defect correction, $\tau_h^H = L^H(I_h^H u^h) - \tilde{I}_h^H (L^h u^h)$, ensures that the coarse-grid equations maintain the accuracy of the fine-grid equations.

Nonlinear equations of the form $L^h(u^h) = f^h$ fit into this scheme. For $L^H(u^H)$, the same discretization scheme as on the fine grid can be adopted. Smoothing operators can be based on a local linearization of the residuals.

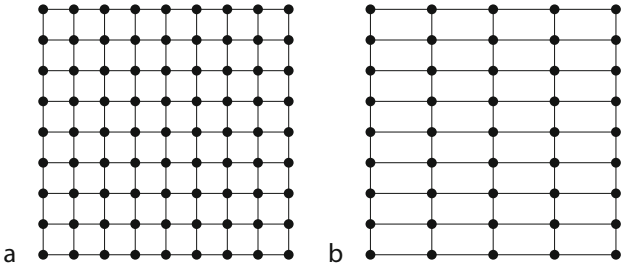
Generalizations

Multigrid is highly efficient for the solution of Poisson's equation on the square, discretized by the finite-difference method on a cartesian grid. For other PDEs, describing convection, diffusion, waves, flow, the method may be less easy to apply. The same is true for unstructured grids, for instance based on triangles or tetrahedra. With suitable modifications and at the expense of additional code complexity, the multigrid method can still be the optimal approach. With unstructured grids, the coarser and finer grid are generally not nested, leading to grid transfer operations that are less easy to code than on a cartesian grid.

Anisotropy in the PDE can degrade performance. Consider the equation

$$-\frac{\partial^2 u}{\partial x^2} - a\frac{\partial^2 u}{\partial y^2} = f(x,y).$$

For $0 < a \ll 1$, smoothing operators may have difficulty with the y-direction where there is only weak coupling between neighboring unknowns. Effective anisotropy in the discretized equation can also happen with grids that have widely different spacings in the various coordinate directions. A more powerful relaxation scheme may repair the poor performance of a simple smoother. For 2D problems, line relaxation will smooth in one direction and solve in the other. An example is line Jacobi relaxation, which approximates the matrix L^h by dropping the off-diagonals in one coordinate direction, while keeping them in the other. The dropped off-diagonals can be subtracted from the main diagonal with a suitable weight factor to obtain better smoothing properties. For 3D problems, plane relaxation is required, which may be too expensive. An alternative is semi-coarsening, where the coarse grid has the same number of points in the direction of weak coupling and is coarsened only in the direction in which the smoothing effectively removes the shorter wavelengths. Figure 3 shows



Numerical Methods, Multigrid, Figure 3 Semi-coarsening of grid (**a**) in the horizontal direction provides grid (**b**).

an example of a grid before and after semi-coarsening in the horizontal direction. In the more general case where semi-coarsening is required in all coordinates, it can be applied in alternating directions. A more costly but powerful approach is simultaneous semi-coarsening in all coordinates (Mulder, 1989), which still retains optimal complexity.

Beyond partial differential equations

The application of the multigrid method is not restricted to the solution of PDEs. There are generalizations to eigenvalue problems (Brandt et al., 1983), integral equations, optimization and optimal control (Borzì and Schulz, 2009), statistical physics (Kandel et al., 1988), image processing (Terzopoulos, 1986), image segmentation, and edge detection. In all cases, the simultaneous use of different scales overcomes the problem of slowly converging solution components.

Algebraic multigrid (AMG), or, better, algebraic multilevel method, refers to a class of solvers that construct sets of coarse-level equations without referring to grids or geometrical properties (Trottenberg et al., 2001; Shapira, 2008). Coarsening is applied to subsets of unknowns for which smoothing is effective. If the problem is specified by a large matrix, these subsets can be determined by examining the coefficients of the matrix. Strongly coupled unknowns, say u_i and u_j, are related by a matrix coefficient a_{ij} that is relatively large. In that case, smoothing is effective and one of the unknowns can be removed from the equations on the coarser level. The coarser level will then be mainly populated by weakly coupled variables. Operator-weighted grid transfer operators are natural in this context, as well as the Galerkin approach for the construction of the coarse-level operator. Algebraic multilevel methods are a popular choice for finite-element discretizations on unstructured grids, as they can be used as a black-box solver and do not require the tedious coding of grid transfer operations between non-nested elements.

Deflation (Vuik et al., 1999) is a technique that accelerates convergence of an iterative procedure if that is slowed down by only a few solution components. Their projection to a subspace that singles them out can be readily solved if the subspace is small.

The multigrid method bears some resemblance to other techniques such as wavelets, hierarchical-basis finite elements, multi-scale techniques, cyclic reduction, fast multipole methods, and other divide-and-conquer methods.

Geophysical applications

Multigrid has been applied successfully to a wide range of partial differential equations, describing static and dynamics problems, diffusion, convection, and flow problems. Examples in geophysics include elliptic problems in potential theory, such as gravity (Boulanger and Chouteau, 2001; Kusche, 2002), magnetostatics (De Gersem and Hameyer, 2001), electrostatics (Bailey and Cheesman, 1996), and Darcy's law in porous media flow (Schmidt, 1995). Controlled-source electromagnetics (Aruliah and Ascher, 2002; Mulder, 2006) and mantle convection (Trompert and Hansen, 1996) are examples of parabolic problems. The wave equation for seismic applications is a hyperbolic problem. Iterative solution of its frequency-domain formulation, the Helmholtz equation, was a notoriously difficult numerical problem but can nowadays be accomplished by using a damped version as preconditioner in a conjugate-gradient-type iterative scheme. Multigrid efficiently deals with the approximate inversion of the preconditioner (Erlangga et al., 2006; Riyanti et al., 2006).

Summary

The multigrid method provides an optimal iterative solution method for a wide class of problems governed by partial differential equations. Its application to a finite-difference discretization of Poisson's equation on a square is the easiest. For other problems, a bit more effort may be required.

Bibliography

Aruliah, D. A., and Ascher, U. M., 2002. Multigrid preconditioning for Krylov methods for time-harmonic Maxwell's equations in three dimensions. *SIAM Journal on Scientific Computing*, **24**(2), 702–718, doi:10.1137/S1064827501387358.

Bailey, R. C., and Cheesman, S., 1996. A multigrid solver for inductive limit EM responses in orebody delineation. *SEG, Expanded Abstracts*, **15**, 265–268, doi:10.1190/1.1826615.

Boulanger, O., and Chouteau, M., 2001. Constraints in 3D gravity inversion. *Geophysical Prospecting*, **49**(2), 265–280, doi:10.1046/j.1365-2478.2001.00254.x

Borzi, A., and Schulz, V., 2009. Multigrid methods for PDE optimization. *SIAM Review*, **51**(2), 361–395, doi:10.1137/060671590.

Brandt, A., 1973. Multi-level adaptive technique (MLAT) for fast numerical solution to boundary value problems. In *Proceedings of the Third International Conference on Numerical Methods in Fluid Mechanics*, Lecture Notes in Physics, Vol. 18, pp 82–89. Springer-Verlag. ISBN 978-3-540-06170-0, doi:10.1007/BFb0118663.

Brandt, A., 1982. Guide to multigrid development. Multigrid Methods. In *Proceedings of the Conference Held at Köln-Porz*, November 23–27, 1981. Hackbush, W., and Trottenberg, U. (eds.), Lecture Notes in Mathematics, Vol. 960: 220–312. Springer-Verlag, ISBN 978-3-540-11955-5, doi:10.1007/BFb0069930.

Brandt, A., McCormick, S., and Ruge, J., 1983. Multigrid methods for differential eigenproblems. *SIAM Journal on Scientific and Statistical Computing*, **4**(2), 244–260, doi:10.1137/0904019.

Briggs, W. L., Henson, V. E., and McCormick, S. F., 2000. *A Multigrid Tutorial*, 2nd edn. Philadelphia: Society for Industrial and Applied Mathematics.

Erlangga, Y. A., Oosterlee, C. W., and Vuik, C., 2006. A novel multigrid based preconditioner for heterogeneous Helmholtz problems. *SIAM Journal on Scientific Computing*, **27**(4), 1471–1492, doi:10.1137/040615195.

Fedorenko, R. P., 1964. The speed of convergence of one iterative process. *USSR Computational Mathematics and Mathematical Physics*, **4**(3), 227–235, doi:10.1016/0041-5553(64)90253-8.

De Gersem, H., and Hameyer, K., 2001. Full multigrid for magnetostatics using unstructured and non-nested meshes. *IEEE Transactions on Magnetics*, **37**(5), 3460–3464, doi:10.1109/20.952637.

Hackbush, W., 1976. Ein iteratives Verfahren zur schnellen Auflösung elliptischer Randwertprobleme, Mathematisches Institut, Universität zu Köln, Report 76–12.

Hackbush, W., 1985. Multi-Grid Methods and Applications, Springer Verlag, Berlin (2nd printing, 2003). http://www.springer.com/math/cse/book/978-3-540-12761-1.

Kandel, D., Domany, E., Ron, D., Brandt, A., and Loh, E., 1988. Simulations without critical slowing down. *Physical Review Letters*, **60**(16), 1591–1594, doi:10.1103/PhysRevLett.60.1591.

Kusche, J., 2002. On fast multigrid iteration techniques for the solution of normal equations in satellite gravity recovery. *Journal of Geodynamics*, **33**(1–2), 173–186, doi:10.1016/S0264-3707(01)00062-X.

Mulder, W. A., 1986. Computation of the quasi-steady gas flow in a spiral galaxy by means of a multigrid method. *Astronomy and Astrophysics*, **156**, 354–380.

Mulder, W. A., 1989. A new multigrid approach to convection problems. *Journal of Computational Physics*, **83**(2), 303–323, doi:10.1016/0021-9991(89)90121-6.

Mulder, W. A., 2006. A multigrid solver for 3D electromagnetic diffusion. *Geophysical Prospecting*, **54**(5), 663–649, doi:10.1111/j.1365-2478.2006.00558.x.

Riyanti, C. D., Erlangga, Y. A., Plessix, R.-E., Mulder, W. A., Vuik, C., and Oosterlee, C. W., 2006. A new iterative solver for the time-harmonic wave equation. *Geophysics*, **71**(5), E57–E63, doi:10.1190/1.2231109.

Schmidt, G., 1995. Operator weighing in a multigrid method for locally refined grids. *Journal of Engineering Mathematics*, **29**(5), 473–496, doi:10.1007/BF00043979.

Shapira, Y., 2008. *Matrix-Based Multigrid: Theory and Applications Numerical Methods and Algorithms* Vol. 2, p. 322, 2nd edn. Springer Verlag, Berlin. ISBN: 978-0-387-49764-8.

Terzopoulos, D., 1986. Image analysis using multigrid relaxation methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **8**(2), 129–139, doi:10.1109/TPAMI.1986.4767767.

Trompert, R. A., and Hansen, U., 1996. The application of a finite volume multigrid method to threedimensional flow problems in a highly viscous fluid with a variable viscosity. *Geophysical and Astrophysical Fluid Dynamics*, **83**(3–4), 261–291, doi:10.1080/03091929608208968.

Trottenberg, U., Oosterlee, C. W., and Schuller, A., 2001. *Multigrid*. London: Academic. ISBN 012701070X.

Vuik, C., Segal, A., and Meijerink, J. A., 1999. An efficient preconditioned CG method for the solution of a class of layered problems with extreme contrasts in the coefficients. *Journal of Computational Physics*, **152**(1), 385–403, doi:10.1006/jcph.1999.6255.

Wesseling, P., 1991. *An Introduction to Multigrid Methods*. Chichester: Wiley. ISBN 0471930830.

Cross-references

Numerical Methods, Finite Difference
Numerical Methods, Finite Element

O

OCEAN BOTTOM SEISMICS

Ingo A. Pecher[1,2], Jörg Bialas[3], Ernst R. Flueh[3]
[1]GNS Science, Lower Hutt, New Zealand
[2]Institute of Petroleum Engineering, Heriot-Watt University, Edinburgh, UK
[3]Leibniz-Institut für Meereswissenschaften, University of Kiel, Kiel, Germany

Synonyms

Ocean bottom seismology

Introduction

Ocean bottom seismics, recording of seismic waves on the ocean floor by seismometers placed on the ocean floor, is becoming increasingly common both in academic research surveys and for hydrocarbon exploration. The main advantages of ocean bottom seismics over reflection seismics with hydrophone streamers (see *Seismic Data Acquisition and Processing*) are that the use of receivers on the seafloor yields sufficient source-receiver offsets to analyze the deeper crustal velocity structure, deploying multicomponent receivers allows recording of shear waves, and that instruments can be deployed for an extended time to record natural seismicity. In most cases, the seismic sources are either airguns towed on the sea surface or natural seismicity. Seafloor sources are rare.

Instrumentation

Seismic seafloor receivers fall broadly into three categories, (1) ocean bottom seismometers (OBSs), individual receivers that can be deployed and recovered without the need of any specialized vessels, (2) ocean bottom cables (OBCs), streamer-like arrays of receivers that are deployed from a specialized vessel, and (3) ocean bottom seismic nodes, individual receivers that require underwater vehicles for deployment and retrieval.

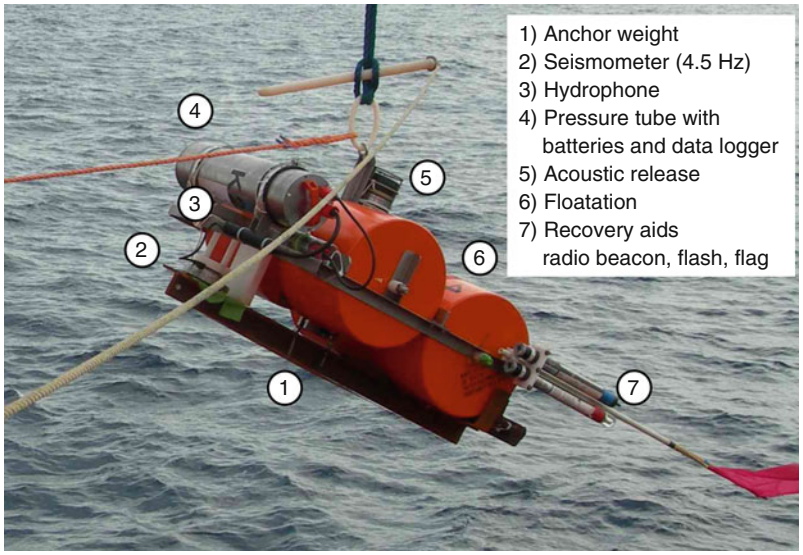
Ocean bottom seismometers

OBSs are individual stations that contain sensors and electronics for data storage (Figure 1). These instruments are deployed from a vessel, and descend to the seafloor weighted by an anchor. A release mechanism is activated toward the end of the survey, detaching the instrument from the anchor. The instrument then rises to the surface buoyed by flotation devices. The instrument is recovered onto the vessel where the recorded data are downloaded. The term ocean bottom seismometer is sometimes used for any instrument that records seismic waves at the seafloor, including OBCs and ocean bottom seismic nodes.

Communication with OBSs while on the seafloor is usually performed with acoustic transmission. This includes releasing the instrument from the anchor; although sometimes, a time-based release is used as backup in case communication fails. Acoustic communication is also used for determining the position of the instrument at the seafloor by triangulation and for checking the status of the instrument.

OBSs that contain only a hydrophone as receiver are often called ocean bottom hydrophones. Hydrophones record pressure waves in the water. Some instruments contain vertically oriented geophones, mostly in combination with hydrophones. But, more commonly used OBSs have 3-component geophones (one vertical, two horizontal components) or 4-component (4-C) OBSs that contain 3-component geophones and a hydrophone. The type of geophone varies, depending on application – usually 4.5 Hz geophones (geophones with a resonance frequency at 4.5 Hz) for active-source experiments. Broadband seismometers with varying frequency bands (with periods as low as several tens of seconds) are used for recording of natural seismicity. In a few cases, accelerometers

Harsh K. Gupta (ed.), *Encyclopedia of Solid Earth Geophysics*, DOI 10.1007/978-90-481-8702-7,



Ocean Bottom Seismics, Figure 1 Ocean bottom seismometer (instruments used by IfM-Geomar, Kiel, Germany).

(measuring particle acceleration rather than velocity as in geophones) are used to record seismic waves.

Seismic data are recorded continuously. Storage media have evolved from analogue tapes to hard drives and flash drives. Exact time stamps from an internal clock are critical for OBS records. GPS time synchronization is completed before and after deployment. Modern high precision clocks keep the deviation as low as 0.05 ppm (4.3 ms/day). For active-source experiments, the records are split into receiver gathers using shot times, usually in SEG-Y format. Data processing varies widely depending on the application and resembles that of land data. Some peculiarities of OBS records include unknown receiver locations, unknown orientations of the horizontal components, and a lack of continuous GPS time stamps. Typical specifications of a modern OBS are listed in Table 1.

Until the 1990s active-source OBS surveys typically involved 10–50 OBSs deployed along a seismic transect. Spacing between instruments for a survey targeting the deeper crustal structure typically was between 5 and 20 km, allowing traveltime analysis of wide-angle reflected and refracted arrivals. In the past decade, significant improvements in hardware and electronics have made instruments less expensive and shortened turnaround times. This has led to a number of experiments with considerably more OBSs allowing imaging of deep crustal structures and 3D traveltime analysis (e.g., Kodaira et al., 2000; Morgan et al., 2002).

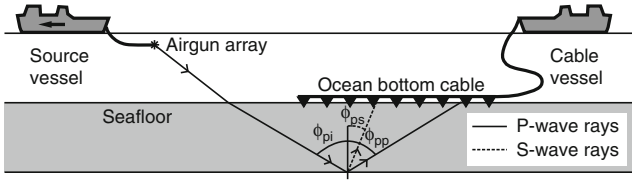
Advances in batteries and storage devices also make it possible to continuously record seismicity with OBSs for up to ~1 year. The use of additional batteries may extend this time window but even for frames and housing made of titanium, corrosion is a problem. Therefore a service recovery after ~1 year is good practice. Such passive-source experiments are becoming more common to study the earth's deep structure, subduction zones on regional scales, as well as seismically active local features such as mud volcanoes (Brueckmann et al., 2009). A combination of onshore and offshore seismological networks has shown to significantly improve the determination of near-shore hypocenters (Husen et al., 1999; Arroyo et al., 2009).

Ocean Bottom Seismics, Table 1 Specification of modern OBS

Recording channels	Min. 4 (hydrophone, 3-component seismometer) Optional additional channels for supplementary sensors (e.g., pressure, CH_4, temperature)
Sampling rate	20 Hz to 1 kHz (passive or active seismology)
Data capacity	Tens of GB with flash memory (lowest power) >50 GB with low power disk drives
Time stability	GPS synchronized prior and after mission Timing accuracy 0.05 ppm/0.03 ppm (2.6/4.3 ms/day; depending on power consumption)
Resolution	20 + bit, 32 bit ideally wanted (not yet available)
Power consumption	<250 mW when active (depending on sensors)
Time of operation	Depending on battery supply up to 1 year
Release system	Acoustic release (optional with additional time release); burn wires, gradually replaced by motor-driven hooks
Recovery aids	Radio beacon, flash light Optional: flag, shroud line

Ocean bottom cables

OBCs are densely spaced receivers attached to a cable. The cable is deployed from a vessel, usually a specialized ship (Figure 2). Seismic data are transmitted along the cable to be recorded on the vessel. Typical receiver spacing is 25 m. Most modern cables are equipped with 4-C receivers. Cables are often dragged



Ocean Bottom Seismics, Figure 2 OBC acquisition. Part of the downgoing P-wave energy is converted upon reflection to upgoing S-wave energy. ϕ_{pi}: angle of incidence, downgoing P-wave. ϕ_{pp}: exit angle, reflected P-wave. ϕ_{ps}: exit angle, P-to-S converted wave. Ray angles according to Snell's Law, that is, $\phi_{pi} = \phi_{pp}$ and $\sin(\phi_{ps})/V_s = \sin(\phi_{pi})/V_p$.

along the seafloor to achieve deployment in a straight line. Some OBCs use geophones that are being pushed into the seafloor by remotely operated vehicles (ROVs) to improve coupling to the seafloor. OBCs may also be deployed in trenches to improve coupling and reduce noise, such as for the Life of Field experiment for long-term monitoring of the Valhall oil field, North Sea (Barkved and Kristiansen, 2005).

OBCs allow imaging of the subsurface with similar coverage as surface-towed streamers. Furthermore, coupling between receivers and seafloor is generally far better than for OBSs partly because the mass of the recording equipment is much smaller, and partly because the deployment techniques for OBCs facilitate slight "burrowing" of cables into the seafloor. OBCs are widely used in the oil and gas exploration industry. The need for a specialized cable deployment vessel, however, normally makes OBCs prohibitively expensive for academic research to date.

Ocean bottom seismic nodes

Ocean bottom seismic nodes are individual receivers with independent recording units – essentially, OBSs without anchors and flotation devices. They are typically deployed with ROVs making their deployment relatively expensive compared to OBSs. Nodes are often deployed to fill coverage gaps in OBC surveys, for example, in the immediate vicinity of platforms where it is unsafe to deploy OBCs (or OBSs). Nodes are often pushed into the seafloor, which potentially leads to improved receiver coupling (Caldwell, 1999).

Ocean bottom seismic sources

By far the majority of active-source OBS surveys use pressure waves generated by airguns towed behind a vessel, although some ocean bottom sources have been pioneered, mainly to generate shear waves.

Explosives have been used as sources at the seafloor for an experiment to record shear waves in boreholes during Ocean Drilling Program Leg 156 but because of the strong increase of shear (S-) wave velocity with depth beneath the seafloor, rays were bent away from the borehole and did not reach borehole receivers (Peacock et al., 1997). Furthermore, environmental and health and safety concerns preclude large-scale use of seafloor explosives.

Imploding glass spheres are the source of the SEEBOSEIS system, for which glass spheres containing air at atmospheric pressure are punctured on the seafloor (Nguyen et al., 2009).

Shear-wave "sleds" are occasionally used to study the high-resolution shear-wave structure of the first few meters beneath the seafloor. One such system uses an electromagnetic coil to generate seismic shear energy at the seafloor. Receivers are dragged along the seafloor behind the source. This system has been used successfully, for example, to measure near-seafloor S-wave velocity offshore California (Huws et al., 2000).

Applications

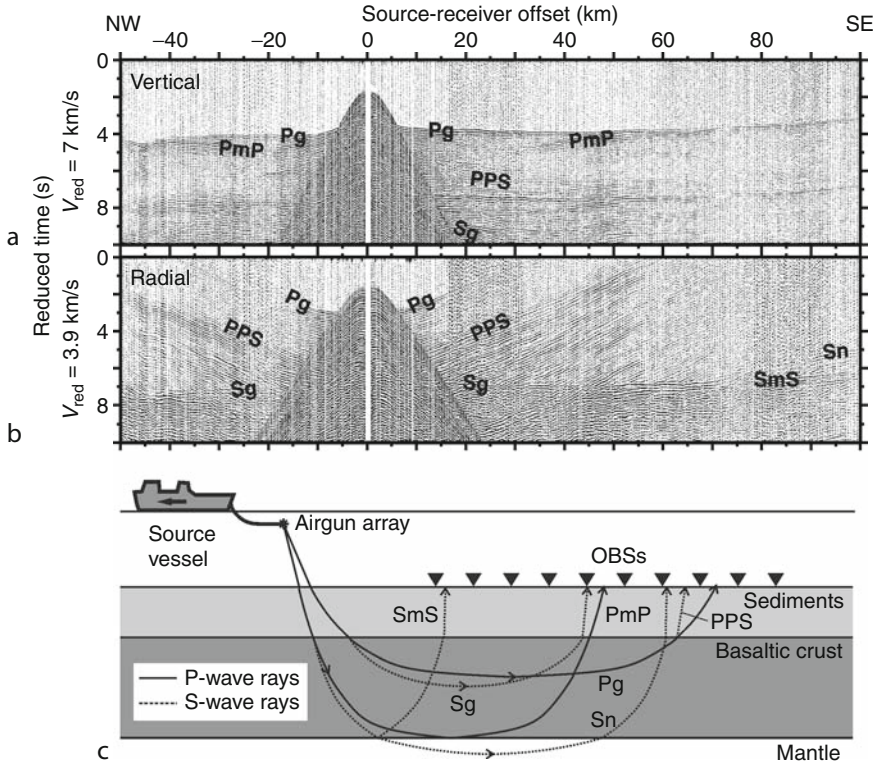
P-waves

OBSs have been used extensively in the last 4 decades to study the crustal structure beneath the seafour (see *Deep Seismic Reflection and Refraction Profiling*) from P-wave traveltime analysis. A typical survey, for example, for a transect across a subduction zone consists of 20+ instruments spaced between 5 and 20 km. Seismic shot lines with maximum offsets of often >100 km are acquired with a large (~4,000–8,000 cu-in = 65–130 l) airgun array. Often, a coincident seismic reflection profile is acquired with the shot vessel although, in order to avoid noise from previous shots to interfere with arrivals from deep targets, shot intervals for crustal OBS surveys should be much larger (~60 s, equivalent to 100–150 m shot spacing) than would be desired for reflection surveying (~15 s) (Christeson et al., 1996) (see *Single and Multichannel Seismics*).

Such a configuration typically allows recording of reflected and refracted arrivals down to the base of the crust as well as refracted waves through the upper mantle (Figure 3). Records are similar to those from land surveys, albeit often at considerably higher signal-to-noise ratio, and receivers and shots are reversed; land crustal wide-angle surveys typically have sparse shot and dense receiver spacing, whereas marine OBS surveys have dense shot but sparse receiver spacing.

Data processing often only involves bandpass filtering and amplitude scaling. Typical sources of noise or data contamination include reverberations particularly on horizontal components, previous-shot noise, water-column multiples, and noise from underwater currents that may shake the instruments.

As for land surveys, data are often displayed using a reduced traveltime ($t_{red} = t - x/v_{red}$, where t_{red}: reduced time, x: source-receiver offset, v_{red}: reduction velocity). The origin of arrivals is identified and their traveltimes are picked, followed by traveltime analysis to determine seismic velocities (see *Traveltime Tomography Using Controlled-Source Seismic Data*). The latter is performed based on ray-trace modeling (e.g., Zelt and Smith, 1992). In the past decade, traveltime inversion has become



Ocean Bottom Seismics, Figure 3 Example of an OBS record from offshore the Faeroe Islands with interpreted arrivals (after Eccles et al., 2009). (**a**) vertical component, (**b**) rotated radial component, (**c**) schematic ray paths. See text in chapter on converted waves for details on radial component. Reduction velocities 7 km/s for vertical component, 3.9 km/s for radial component. Arrivals P-waves: Pg: diving wave in the crust, PmP: reflection from the top of the mantle (Mohorovičić discontinuity). Converted S-waves: PPS: S-wave, converted upon transmission from upgoing P-wave at basalt-sediment interface. Sg: S-wave, converted upon transmission from downgoing P-wave at basalt-sediment interface and diving as S-wave in the basaltic crust. SmS: S-wave, converted upon transmission from downgoing P-wave at basalt-sediment interface and reflected as S-wave at the top of the mantle. Sn: S-wave, converted upon transmission from downgoing P-wave at basalt-sediment interface and diving as S-wave in the mantle. Courtesy Jennifer Eccles (University of Auckland, New Zealand).

common allowing better determination of the uncertainties and uniqueness of resulting velocity models (e.g., Korenaga et al., 2000). However, the uncertainty from potentially misinterpreting the origin of arrivals is difficult to quantify.

Coincident OBS and seismic reflection surveys are being used for high-resolution studies of the first few hundreds of meters beneath the seafloor, in particular to obtain velocity information in the gas-hydrate zone. Data analysis usually comprises joint analysis of reflected and refracted P-waves often with additional structural information from seismic reflection data (Hobro et al., 1998). Many of these experiments are designed to achieve 3D ray coverage, sometimes coincident with 3D seismic reflection data (Plaza-Faverola et al., 2010). OBS experiments designed for 3D traveltime analysis have also become more common for surveys targeting the deeper crust (e.g., Morgan, et al., 2002).

The increased availability of OBSs has also led to a number of surveys using OBSs with sufficiently dense spacing to allow imaging of deep crustal structures similar to seismic reflection data (e.g., Kodaira, et al., 2000). Advancements in wide-angle migration techniques are key toward successful imaging with OBSs (e.g., Talwani and Zelt, 1998). Imaging using OBSs is also improved by increased reflection coverage by utilizing rays from seafloor multiples, that is, rays that are reflected at the seafloor and sea surface, often referred to as mirror imaging (e.g., Grion et al., 2007).

High-resolution (sub-seismic-wavelength) P-wave velocities have been extracted from OBS data using 1D full-waveform inversion techniques similar to those applied to multichannel reflection data (Korenaga et al., 1997). Increases in computing power now also allow 2D visco-acoustic full-waveform tomography (Sourbier et al., 2009a, b) (see *Seismic, Waveform Modeling and Tomography*).

P-to-S converted waves

Identification of P-to-S converted (PS-) waves in horizontal-component records has led to a paradigm shift in ocean bottom seismics in the late 1990s and early 2000s by the

use of OBCs for multicomponent seismic exploration (Figure 2) (see *Energy Partitioning of Seismic Waves*; *Seismic Data Acquisition and Processing*). P-waves are generated on a shooting vessel. They are partly converted at layer interfaces to S-waves. Upgoing S-waves are then recorded with OBCs or sometimes, densely spaced OBSs. The most prominent PS-arrivals are usually arrivals that are converted from downgoing P-waves upon reflection to upgoing S-waves.

Processing sequences for PS-waves require additional steps compared to conventional P-wave seismic reflection data acquired with surface-towed streamers (Stewart et al., 2002). In particular, processing of PS-waves requires separation of P- and S-waves, a rotation of the horizontal components such that a maximum amount of energy is projected into a single component, the radial component, and various modifications to adjust for the asymmetry of the PS-raypaths. A key step of data analysis is the correlation of P- and PS-events. Once P- and PS-events from the same horizons have been identified, determination of the ratio of P-wave velocity (*Vp*) over S-wave velocity (*Vs*) is straight-forward (e.g., Granli et al., 1999).

Multicomponent PS-wave surveys are being conducted routinely by the exploration industry, chiefly with the following objectives:

Imaging through gas clouds. P-waves are attenuated significantly in gas-bearing layers, whereas S-waves are almost unaffected by gas. Converted-wave surveys often yield superior images of structures in and beneath gas-bearing sediments even though in many cases the downgoing P-wave still has to penetrate gas (Granli et al., 1999).

Fluid discrimination. Reflection strength of PS-waves is only marginally affected by the pore fill (through its effect on density). Therefore, flat spots, oil-water or gas-water contacts at the base of hydrocarbon reservoirs, which generate P-wave reflections, usually do not cause any noticeable PS-reflections (e.g., MacLeod et al., 1999). For the same reason, PS-waves may help establish whether high negative-polarity P-wave reflections (bright spots) are likely to be caused by gas in sediment pores.

Mapping of sand–shale interfaces. Sand and shale layers often have similar *Vp* and density, which leads to low P-wave reflectivity. However, because *Vs* in sands is usually considerably higher than in shales, sand–shale interfaces often cause strong P-to-S conversion and have been imaged successfully with PS-waves (e.g., MacLeod et al., 1999).

Lithology from Vp/Vs. *Vp* – *Vp*/*Vs* crossplots may be employed to delineate between sands and shales and to constrain the shale content of reservoir sands (Margrave et al., 1998). *Vp*/*Vs* may also allow to distinguish between different types of carbonates (Rafavich et al., 1984).

Fracture detection from shear-wave splitting. Near-vertical fractures common, for example, in carbonate reservoirs, causes an azimuthal dependence of velocities, azimuthal anisotropy (see *Seismic Anisotropy*). The propagation of S-waves that are polarized perpendicularly to fractures is slower than that of S-waves polarized parallelly to fractures. This leads to PS-waves being split into fast and slow waves, depending on the azimuth of the incoming P-wave, and allows determination of azimuth and magnitude of anisotropy, from which fracture orientation can be inferred (Gaiser et al., 2002).

In recent years, P-to-S converted waves have been utilized increasingly to study *Vs* in the Earth's crust (e.g., Eccles et al., 2009). Unlike for hydrocarbon exploration, with the aim of imaging the sedimentary section with S-waves converted from P-waves upon reflection, studies of the deeper crust often analyze traveltimes of S-waves converted from downgoing P-waves upon transmission. The sediment-basalt interface on the North Atlantic margin between the UK and Iceland, for example, provided a conversion horizon to allow determination of *Vs* in the deeper crust (Figure 3). *Vp* – *Vp*/*Vs* crossplots allowed lithological conclusions indicating mixing between continental crust and mafic intrusions across the ocean-continent transition and the presence of sub-basalt sediments on the Faeroe Margin (Eccles et al., 2009).

Wide-aperture (wide-azimuth) studies

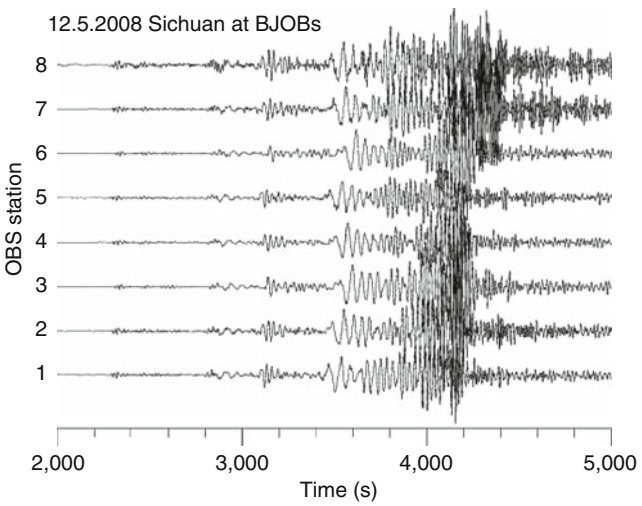
Wide-aperture studies are increasingly conducted in the hydrocarbon exploration for "undershooting" in order to illuminate the subsurface beneath features that are difficult to penetrate with seismic energy and/or exhibit prominent localized velocity anomalies, such as gas clouds or salt domes: Rays in wide-aperture surveys may bypass these features and reach deeper targets of interest. The length of seismic streamers towed behind the same vessel that tows the airgun sources may not be sufficient for wide-aperture studies. Seismic arrivals at wide apertures may be recorded with surface-towed streamers by using separate vessels for sources and receivers. Alternatively, receivers deployed at the seafloor may be used. An acquisition campaign partly aimed at obtaining wide-aperture data across the Atlantis Field in the deep-water Gulf of Mexico involved 900 ocean bottom seismic nodes (Barley and Summers, 2007). 3D surveys with ocean bottom receivers also provide illumination of seismic targets from different azimuths (multi-azimuth surveys), which allows analysis for azimuthal anisotropy and may improve image quality. Wide-aperture seismic in the hydrocarbon exploration industry is often acquired as part of "seismic-on-demand" 4D surveys over permanently installed seismic surveillance systems such as at the Valhall Field (Barley and Summers, 2007 and references therein).

Passive-source seismology

The use of modern storage media has led to a decrease in power consumption and an increase of storage capacity allowing deployments of OBSs for several months to over

a year to record natural seismicity (Figure 4). Because of the lower frequencies of natural seismic sources usually, broadband geophones are used with a lower resonance frequency than the 4.5 Hz geophones typically used for active-source surveys. Both local seismicity and teleseismic events are being evaluated similarly to data from land-seismic records. Unlike for land data, however, receiver location and geophone orientation need to be established. Local noise conditions in the deep oceans vary significantly throughout the seasons, but in general, noise is higher than on well-placed land stations. In addition, the seismic waveforms are contaminated by water-column multiples.

A very different application of passive-source ocean bottom seismology is to measure seafloor compliance. Pressure variations from ocean waves (gravity or infragravity waves) are being transmitted into the seafloor. Seafloor compliance is the transfer function between pressure loading and seafloor deformation. This transfer function is determined as a function of frequency and depends largely on the subsurface shear modulus (Crawford et al., 1991). Pressure variations are typically recorded with differential pressure gauges, seafloor deformation derived from acceleration measured with self-leveling gravimeters. The frequency range of the analyzed pressure variations depends largely on water depth – because of the dispersion of ocean-surface waves, short-period waves are not transmitted into deeper waters. The resolution of subsurface shear-modulus analyses from compliance measurements decreases with decreasing frequencies. As an example, the frequency range for studying gas-hydrate-bearing sediments at 1.3 km water depth and 2 km beneath the seafloor was 0.003–0.049 Hz (Willoughby et al., 2008).



Ocean Bottom Seismics, Figure 4 Teleseismic event on array of broadband OBSs at ~74°N 17°E (between Norway and Spitsbergen), the Mw 7.9 Sichuan Earthquake (China), May 12, 2008. Courtesy Frank Krüger (University of Potsdam, Germany).

Future developments

One of the most significant challenges for autonomous OBSs is still the power supply. Both data loggers and storage media are becoming increasingly energy efficient. Together with new battery technology and further improvement of fuel cells, this is leading to reduction of weight and size of instrumentation. Some study areas may provide additional power sources, such as thermal energy. Weight and size reduction in modern instruments is achieved in most cases by installing all the components in a single glass sphere. On the other hand, glass spheres require increased service efforts and precautions compared to more robust, but larger and heavier titanium pressure casing.

The use of ocean bottom sensors in permanent ocean-floor observatories is likely to become more common. OBSs have already been installed for long-term monitoring at seafloor observatories (e.g., GEOSTAR, Beranzoli et al., 2003) and have been integrated in deployments for tsunami warning systems. These observatories generally have the capability of data transmission to shore. Regional studies usually require a deployment of an array of sensors distributed over an area of several square kilometers. Several seismic seafloor observatories have been established to study the seismogenic zone offshore of Japan and a network of 20 instruments is being developed as part of the Japanese Dense Oceanfloor Network System for Earthquakes and Tsunamis (Kaneda et al., 2009).

Combined seismic and electromagnetic field experiments are becoming more common as the combination of both methods is proving powerful for studying subsurface lithology (Constable and Srnka, 2007 and references therein). Despite significant technical challenges, especially with contamination of electromagnetic signals from the seismic instrumentation, it is conceivable that future ocean bottom instruments will include sensors for both seismic and electromagnetic waves.

The use of OBCs and nodes for hydrocarbon exploration and field appraisal appears to have become a “mainstream” technique. While costs for deployments of either type of instruments are still high compared to conventional 3D streamer seismic, they need to be weighed against the risk of mispositioning considerably more expensive appraisal wells (Barley and Summers, 2007). Furthermore, several hydrocarbon fields are now being monitored with permanent installation of ocean bottom sensors for repeat-seismic surveys in order to guide field development (Barley and Summers, 2007 and references therein).

Conclusions

The last decade has seen a substantial increase of ocean bottom seismic studies. While seafloor seismic sources are rarely deployed, the use of seafloor seismic sensors has established itself primarily for recording offshore wide-angle data, P-to-S converted waves, and natural seismicity. Academic research surveys, for studying the

deeper crustal structure, are usually conducted with autonomous OBSs, which are relatively inexpensive and versatile. Surveys by the exploration industry for evaluating and developing hydrocarbon reservoirs mostly involve deployment of ocean bottom cables or ocean bottom seismic nodes, which give superior data quality compared to OBSs but at much higher cost. In recent years, permanent deployment of instruments has become more common in ocean bottom observatories and for monitoring the development of hydrocarbon fields. The rapid development of ocean bottom seismic instruments is leading to improvements in reliability, data quality, and efficiency, which makes it likely that ocean bottom seismics will continue to expand significantly in the future.

Acknowledgments

We would like to thank Stuart Henrys (GNS Science) and an anonymous external reviewer for constructive comments. We also thank Jennifer Eccles (University of Auckland, New Zealand) and Frank Krüger (University of Potsdam, Germany) for providing the images in Figures 3 and 4, respectively.

Bibliography

Arroyo, I. G., Husen, S., Flueh, E. R., Gossler, J., Kissling, E., and Alvarado, G. E., 2009. Three-dimensional P-wave velocity structure on the shallow part of the Central Costa Rican Pacific margin from local earthquake tomography using off- and onshore networks. *Geophysical Journal International*, **179**, 827–849.

Barkved, O. I., and Kristiansen, T., 2005. Seismic time-lapse effects and stress changes: examples from a compacting reservoir. *The Leading Edge*, **24**, 1244–1248.

Barley, B., and Summers, T., 2007. Multi-azimuth and wide-azimuth seismic: shallow to deep water, exploration to production. *The Leading Edge*, **26**, 450–457.

Beranzoli, L., Braun, T., Calcara, M., Casale, P., DeSantis, A., D'Anna, G., DiMauro, D. G. E., Favali, P., Fuda, J.-L., Frugoni, F., Gamberi, F., Marani, M., Millot, C., Montuori, C., and Smriglio, G., 2003. Mission results from the first GEOSTAR observatory (AdriaticSea, 1998). *Earth Planets Space*, **55**, 361–373.

Brueckmann, W., Tyron, M. D., Bialas, J., Feseker, T., and Lefeldt, M. R., 2009. Monitoring the dynamic properties of an active Mud Volcano in the West Nile Delta. *Eos Transactions AGU*, 90, Fall meeting supplement, Abstract OS21A-1156.

Caldwell, J., 1999. Marine multicomponent seismology. *The Leading Edge*, **18**, 1274–1282.

Christeson, G. L., Nakamura, Y., McIntosh, K. D., and Stoffa, P. L., 1996. Effect of shot interval on ocean bottom seismograph and hydrophone data. *Geophysical Research Letters*, **23**, 3783–3786.

Constable, S., and Srnka, L. J., 2007. Special section – marine controlled-source electromagnetic methods: an introduction to marine controlled-source electromagnetic methods for hydrocarbon exploration. *Geophysics*, **72**, WA3–WA12.

Crawford, W. C., Webb, S. C., and Hildebrand, J. A., 1991. Seafloor compliance observed by long-period pressure and displacement measurements. *Journal of Geophysical Research*, **96**(B), 16151–16160.

Eccles, J. D., White, R. S., and Christie, P. A. F., 2009. Identification and inversion of converted shear waves: case studies from the European North Atlantic continental margins. *Geophysical Journal International*, **179**, 381–400.

Gaiser, J., Loinger, E., Lynn, H., and Vetri, L., 2002. Birefringence analysis at Emilio Field for fracture characterization. *First Break*, **20**, 505–514.

Granli, J. R., Arntsen, B., Sollid, A., and Hilde, E., 1999. Imaging through gas-filled sediments using marine shear-wave data. *Geophysics*, **64**, 668–677.

Grion, S., Exley, R., Manin, M., Miao, X.-G., Pica, A., Wang, Y., Granger, P.-Y., and Ronen, S., 2007. Mirror imaging of OBS data. *First Break*, **25**, 37–42.

Hobro, J. W. D., Minshull, T. A., and Singh, S. C., 1998. Tomographic seismic studies of the methane hydrate stability zone in the Cascadia margin. In Henriet, J. P., and Mienert, J. (eds.), *Gas Hydrates; Relevance to World Margin Stability and Climatic Change*. Geological Society of London Special Publication, 137. London: Geological Society, pp. 133–140.

Husen, S., Kissling, E., Flueh, E., and Asch, G., 1999. Accurate hypocentre determination in the seismogenic zone of the subducting Nazca Plate in northern Chile using a combined on-/offshore network. *Geophysical Journal International*, **138**, 687–701.

Huws, D. G., Davis, A. M., and Pyrah, J. R., 2000. A Nondestructive Technique for Predicting the In Situ Void Ratio for Marine Sediments. *Marine Georesources and Geotechnology*, **18**, 333–346.

Kaneda, Y., Kawaguchi, K., Araki, E., Sakuma, A., Matsumoto, H., Nakamura, T., Kamiya, S., Ariyoshi, K., Baba, T., Ohori, M., and Hori, T., 2009. Dense Ocean floor Network for Earthquakes and Tsunamis (DONET) – development and data application for the mega thrust earthquakes around the Nankai trough. *Eos Transactions AGU*, 90, Fall meeting supplement, Abstract S53A-1453.

Kodaira, S., Takahashi, N., Kato, A., Park, J.-O., Iwasaki, T., and Kaneda, Y., 2000. High pore fluid pressure may cause silent slip in the NankaiTrough. *Science*, **308**, 1295–1298.

Korenaga, J., Holbrook, W. S., Singh, S. C., and Minshull, T. A., 1997. Natural gas hydrates on the southeast US margin: constraints from full waveform inversion and traveltime inversion of wide-angle seismic data. *Journal of Geophysical Research*, **102**(B), 15345–15365.

Korenaga, J., Holbrook, W. S., Kent, G. M., Kelemen, P. B., Detrick, R. S., Larsen, H. C., Hopper, J. R., and Dahl-Jensen, T., 2000. Crustal structure of the southeast Greenland margin from joint refraction and reflection seismic tomography. *Journal of Geophysical Research*, **105**(B), 21,591–21,614.

MacLeod, M. K., Hanson, R. A., and Bell, C. R., 1999. The Alba field ocean bottom cable seismic survey: impact on development. *The Leading Edge*, **18**, 1306–1312.

Margrave, G. F., Lawton, D. C., and Stewart, R. R., 1998. Interpreting channel sands with 3C-3D seismic data. *The Leading Edge*, **17**, 509–513.

Morgan, J. V., Christeson, G. L., and Zelt, C. A., 2002. Testing the resolution of a 3D velocity tomogram across the Chicxulub crater. *Tectonophysics*, **355**, 215–226.

Nguyen, X. N., Dahm, T., and Grevemeyer, I., 2009. Inversion of Scholte wave dispersion and waveform modeling for shallow structure of the Ninetyeast Ridge. *Journal of Seismology*, **13**, 543–559.

Peacock, S., Westbrook, G. K., and Graham, D. P., 1997. Seismic velocities in the Northern Barbados Ridge accretionary comples, Site 949. In Shipley, T. H., Ogawa, Y., Blum, P., and Bahr, J. M. (eds.), *Proceedings ODP, Scientific Results: Ocean Drilling Program*, Vol. 156, pp. 263–275.

Plaza-Faverola, A., Bünz, S., and Mienert, J., 2010. Fluid distributions inferred from P-wave velocity and reflection seismic amplitude anomalies beneath the Nyegga pockmark field of the mid-Norwegian margin. *Marine and Petroleum Geology*, **27**, 46–60.

Rafavich, F., Kendall, C. H. S. C., and Todd, T. P., 1984. The relationship between acoustic propertoes and the petrographic character of carbonate rocks. *Geophysics*, **49**, 1622–1636.
Sourbier, F., Operto, S., Virieux, J., Amestoy, P., and L'Excellent, J.-Y., 2009a. FWT2D: a massively parallel program for frequency-domain full-waveform tomography of wide-aperture seismic data – part 1 algorithm. *Computers and Geosciences*, **35**, 487–495.
Sourbier, F., Operto, S., Virieux, J., Amestoy, P., and L'Excellent, J.-Y., 2009b. FWT2D: a massively parallel program for frequency-domain full-waveform tomography of wide-aperture seismic data – part 2 numerical examples and scalability analysis. *Computers and Geosciences*, **35**, 496–514.
Stewart, R. S., Gaiser, J. E., Brown, R. J., and Lawton, D. C., 2002. Converted-wave seismic exploration: methods. *Geophysics*, **67**, 1348–1363.
Talwani, M., and Zelt, B., 1998. Some recent developments in the acquisition and processing of seismic data. *Tectonophysics*, **286**, 123–142.
Willoughby, E. C., Latychev, K., Edwards, R. N., Schwalenberg, K., and Hyndman, R. D., 2008. Seafloor compliance imaging of marine gas hydrate deposits and cold vent structures. *Journal of Geophysical Research*, **113**, B07107.
Zelt, C. A., and Smith, R. B., 1992. Seismic traveltime inversion for 2-D crustal velocity structure. *Geophysical Journal International*, **108**, 16–34.

Cross-references

Deep Seismic Reflection and Refraction Profiling
Energy Partitioning of Seismic Waves
Seismic Anisotropy
Seismic Data Acquisition and Processing
Seismic, Waveform Modeling and Tomography
Single and Multichannel Seismics
Traveltime Tomography Using Controlled-Source Seismic Data

OCEAN, SPREADING CENTRE

K. S. Krishna
Geological Oceanography, National Institute of Oceanography (Council of Scientific and Industrial Research), Dona Paula, Goa, India

Definition

The *oceanic spreading center*, also known as the midocean ridge, is an underwater mountainous feature lying between the two lithospheric plates, through which new magma material being continuously spread out and lead to growth of the lithospheric plates.

Ridge structure, partial melted mantle upwelled through the midocean ridge, forms a 3–6 km thick basaltic crust, but at a later stage the same crust moves away from the ridge crest and allows growth of mantle component.

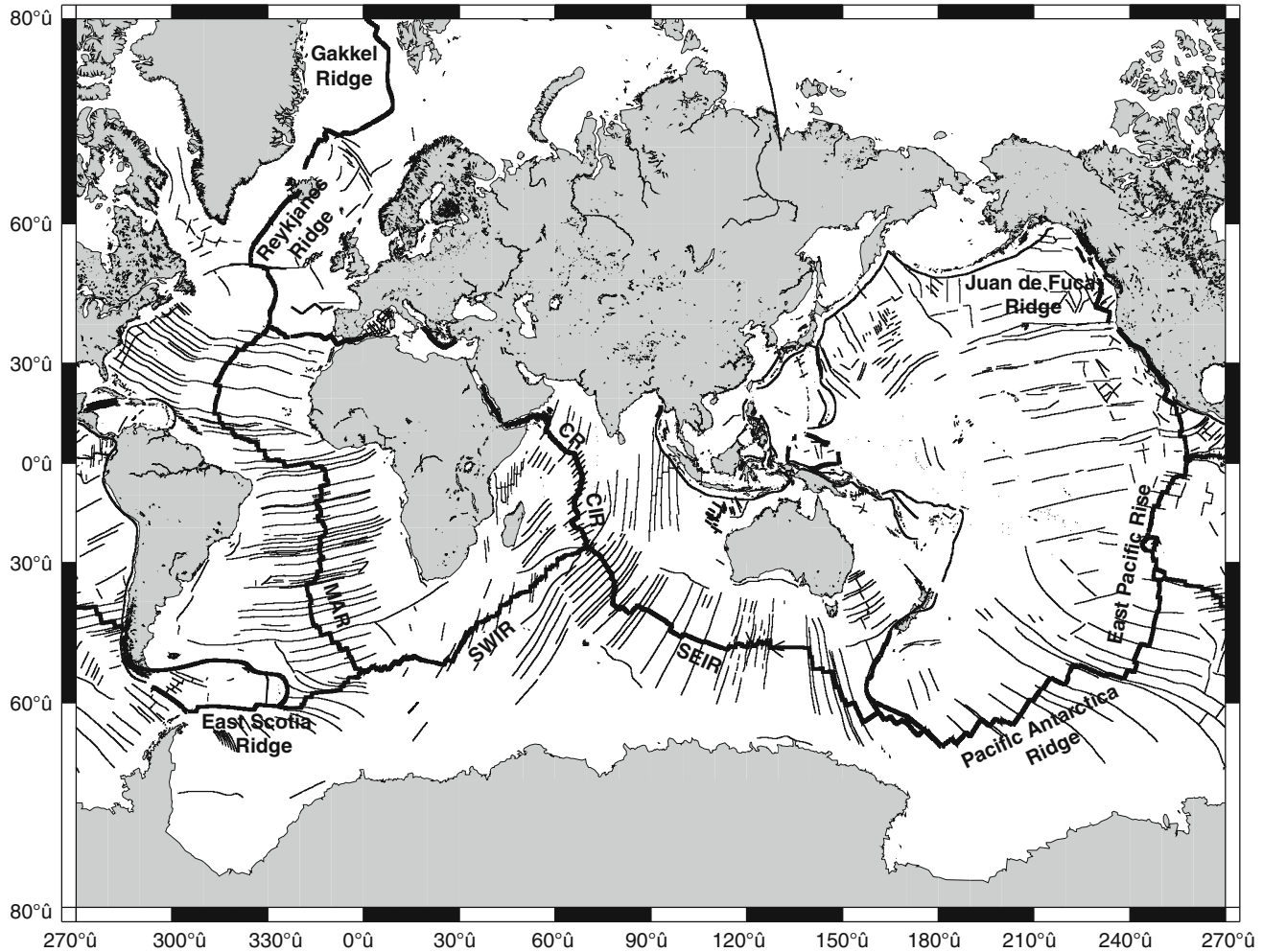
Introduction

The rigid surface layer (generally about 70–100 km thick) of the Earth comprising the crust and uppermost mantle, *termed as lithosphere*, is divided into a number of tectonic plates. Presently, there are about eight primary plates, namely, North American, South American, African, Indian, Australian, Antarctica, Eurasian, and Pacific; seven secondary plates; and several more tertiary plates covering the Earth's surface (for details, see wikipedia web site: http://en.wikipedia.org/wiki/List_of_tectonic_plates). The lithospheric plates are bounded by one of the three main types of geological features: (1) midoceanic ridges, (2) subduction zones, and (3) transform faults. The boundaries are narrow deforming zones and are associated with intense seismicity, whereas the plate's interiors are relatively stable. The plates upon which continents and ocean floor lie are in continuous motion at a speed of few centimeters per year. Each plate is in relative motion with respect to the other on the surface of the Earth. The relative motion produces new crust at midoceanic ridges and consumes old lithosphere at subduction zones. Apart from these tectonic processes, plates do undergo breakups and unifications through geologic time.

Morphology and spreading rates of midoceanic ridge

A midocean ridge is an underwater linear physiographic feature and serves as a divergent tectonic boundary between two lithospheric plates. The midocean ridges, those encircle the entire globe, are physically connected and extend for a total length of about 65,000 km as a single global entity in the world oceans (Figure 1). Many of the major oceanic ridges in the world oceans have been reconfigured in the geological past that led to unification and split of the lithospheric plates. One of the well-mapped ancient oceanic ridges in the Indian Ocean is the Wharton Ridge, once separated the Indian and Australian plates, and the ridge cessation at about 42 million years ago led to unification of both the aforesaid plates into a single Indo-Australian plate (Liu et al., 1983; Krishna et al., 1995).

On the ocean floor, long, linear, and elevated features are observed with variable seafloor undulations (Figure 2). They are the places where magma is continuously upwelling and being added to the previously formed oceanic crust on either side. The upwelling magma rates along the global midoceanic ridges are not fixed; they greatly vary with a range from 6 to 200 mm/year. Following the seafloor expansion rates, the midoceanic ridges are classified into categories of slow, intermediate, and fast spreading ridges. The Gakkel midoceanic ridge located in the Arctic Ocean between Greenland and Siberia (Figure 1) is the slowest spreading ridge with an upwelling magma rate of about 6 mm/year. On the other hand, the East Pacific Rise particularly in equatorial zone generates magma at the fastest rates up to 200 mm/year. The seafloor morphology on the flanks of the midoceanic ridge is largely controlled by the rate of seafloor expansions. Fast expansion spreading rates contribute to generally smooth and flat seafloor surfaces in the vicinity of the ridge, while the slow expansion spreading rates add exceptionally



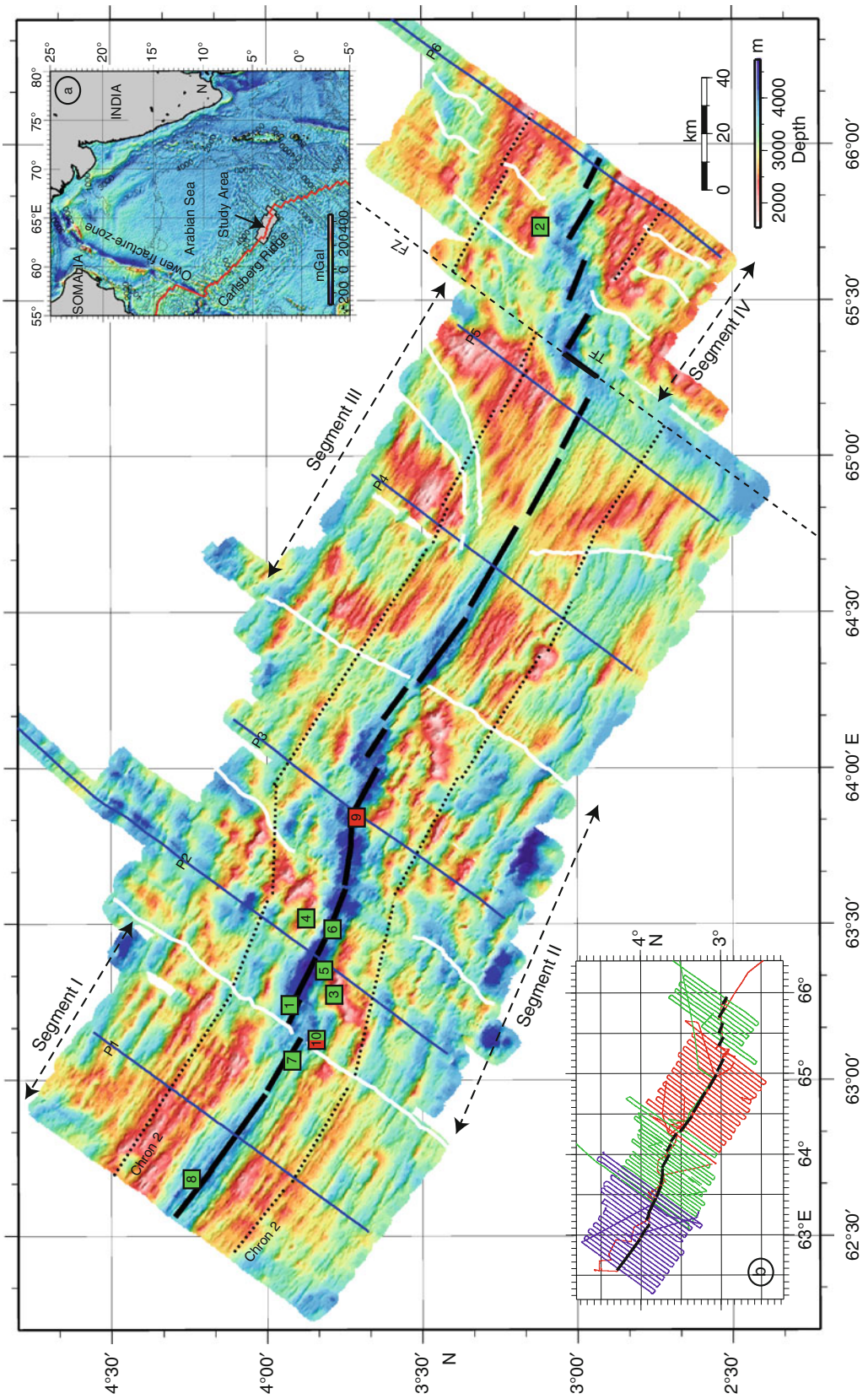
Ocean, Spreading Centre, Figure 1 General map of the global oceans showing midoceanic ridge system, subduction zones, and major fracture zones (after Royer et al., 1989; Müller et al., 1997). MAR indicates Mid-Atlantic Ridge in the Atlantic Ocean. CR, CIR, SWIR, and SEIR represent Carlsberg Ridge, Central Indian Ridge, Southwest Indian Ridge, and Southeast Indian Ridge, respectively, in the Indian Ocean.

rough surfaces. The slow spreading ridges are generally marked in their crest parts by an axial valley, called rift, which is clearly distinguishable along the Carlsberg Ridge in northern Indian Ocean (Figure 2). The ridge crest is strongly characterized by rugged topography, steep valley walls, and wide rift valley floor (Kamesh Raju et al., 2008). Further, the long ridge segment is bisected by less-known non-transform discontinuities.

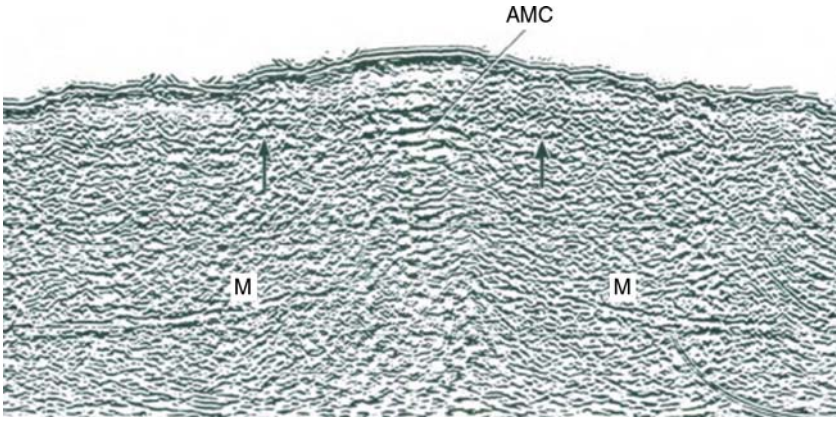
Internal structure of the midoceanic ridge

Along the midocean ridge, the lithospheric plates grow, which eventually lead to partial melting of the upwelling mantle and to form a layer of basaltic crust 3–6 km thick. Close to the ridge crest, the lithosphere may consist of oceanic crust only, but at a later stage the same crust moves away from the ridge crest and allows growth of mantle component due to thermal cooling. Both lithosphere and elastic plate thicken continuously with the age of the ocean floor until about 60 Ma and stabilize thereafter.

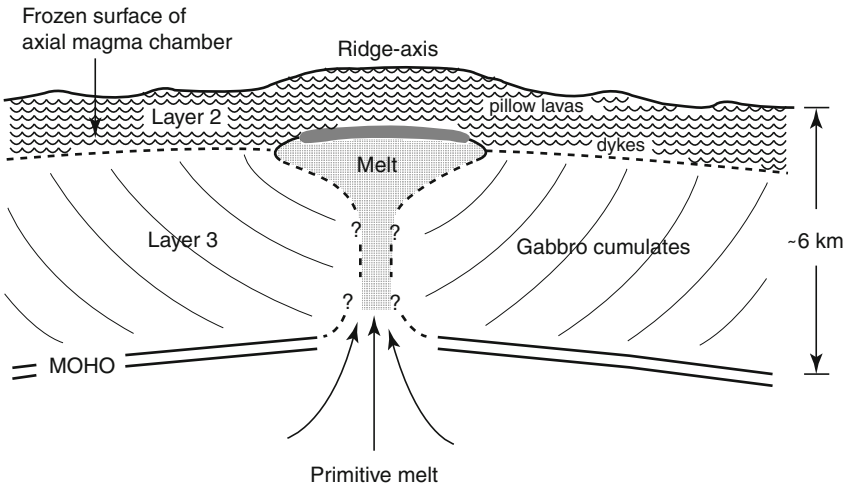
A seismic image across the East Pacific Rise is shown in Figure 3 (Detrick et al., 1987). It shows the reflections received from an interface present at about 6 s two-way travel time possibly marking a Moho boundary and reflections from a shallow event indicating the top of a crustal low-seismic-velocity zone (LVZ). Detrick et al. (1987) interpreted the top of the LVZ as the roof of an axial magma chamber (AMC) within the crust beneath the ridge axis. On either side of the chamber, the reflections are weaker and perhaps represent the frozen top of the magma chamber body (shown with arrows in Figure 3). Also, the reflections from Moho boundary are absent below the magma chamber. A sketch of the interpreted seismic structure of the East Pacific Rise is presented in Figure 4.



Ocean, Spreading Centre, Figure 2 Seafloor image of the Carlsberg Ridge segments, northern Indian Ocean. Inset maps a and b show the study area (*shaded box*) and cruise tracks over the segment of the Carlsberg Ridge. The image is reproduced from Kamesh Raju et al. (2008).



Ocean, Spreading Centre, Figure 3 Seismic reflection cross section across the East Pacific rise shows the internal structure of the oceanic crust beneath the ridge axis. AMC represents significant seismic signals reflected from the roof of an axial magma chamber. Solid arrows on either side mark relatively weaker reflections, perhaps representing frozen top of magma chamber, away from the ridge axis. M indicates the Moho boundary. The image is reproduced from Detrick et al. (1987).



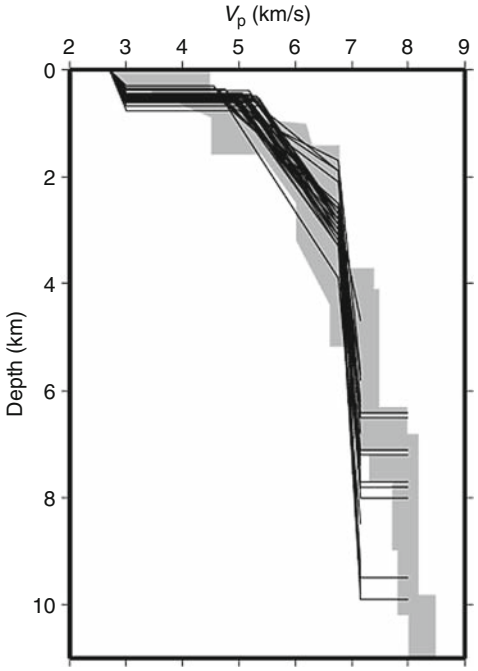
Ocean, Spreading Centre, Figure 4 A sketch showing the crustal structure beneath the East Pacific rise (Reproduced from Detrick et al., 1987).

Under some midoceanic ridge segments, it is found that the axial magma chamber is either small or absent altogether. For example, multichannel reflection experiment carried out on the Mid-Atlantic Ridge (Detrick et al., 1990) did not find evidences for the presence of intracrustal reflections associated with an axial magma chamber. Seismic experiments, in general, demonstrated that along slow-spreading ridges, magma chambers are either absent or extremely short-lived, if at all they exist.

Seismic experiment carried out on a ridge segment between 8° and 9°S of the Mid-Atlantic Ridge revealed largely a different seismic structure from that of the East Pacific Rise. In general, an axial magma chamber is not traceable within the Mid-Atlantic Ridge crust. Velocity models of the Mid-Atlantic Ridge segment (Minshull et al., 2003) suggest a total thickness of crust ranging from 6.5 to 10 km with individual layers 2 and 3 having thickness of approximately 3 and 5 km, respectively (Figure 5). The velocity heterogeneities are much greater in layers 2A and 2B, ranging from 2.65 to 3.0 km/s and 4.6–6.8 km/s, respectively, while the velocities in layer 3 are remarkably uniform.

Relationships between spreading rate, seismic structure, and ridge-axis morphology

Geophysical experiments over the global midoceanic ridges have led to the discovery of some explicit relationships between spreading rate, seismic structure, and ridge-axis morphology. Purdy et al. (1992) found that as full spreading rate decreases from 150 to 20 mm/year, the depth to the top of the axial low-seismic-velocity zone



Ocean, Spreading Centre, Figure 5 Velocity-depth profiles (*solid lines*) beneath the Mid-Atlantic Ridge segment between 8° and 9°S. Shaded region marks the range of velocities for 0–7 Ma Atlantic oceanic crust (White et al., 1992). The figure is reproduced from Minshull et al. (2003).

(LVZ) increases from ~1 to ~4 km. In another study, Minshull et al. (2003) established a relationship between crustal thickness and ridge-axis morphology at constant spreading rate. The ridge segments having large crustal thicknesses are associated with rifted axial highs, whereas the segments with reduced crustal thickness have deep axial valleys. Investigations over the midoceanic ridges have also proven that they are important locations for metallic mineral deposition (Rona, 1980).

Summary

A midocean ridge, an underwater mountainous feature, extends for a length of about 65,000 km as a single continuous body in the global oceans. It is a window to the Earth's interior, through which magma is continuously upwelling and drives the lithospheric plates on either side in order to accommodate newly accreted crust. Many of the oceanic ridges in the world oceans have been abandoned in the geologic past and led to resume the activity elsewhere either in the intra-oceanic or intracontinental plate regions. Partial melted mantle upwelled through the midocean ridge forms a layer of basaltic crust 3–6 km thick, but at a later stage the same crust moves away from the ridge crest and allows growth of mantle component. The upwelling magma rates along the mid-oceanic ridges, in general, control the internal structure. Fast spreading ridge such as East Pacific Rise has an axial magma chamber (AMC) at a shallow depth within the crust, whereas beneath an intermediate spreading ridge, the Mid-Atlantic Ridge, an axial magma chamber, is not traceable. Geophysical experiments over the global midoceanic ridges have found some explicit relationships between spreading rate, seismic structure, and ridge-axis morphology.

Bibliography

Detrick, R. S., Buhl, P., Vera, E., Mutter, J., Orcutt, J., Madsen, J., and Brocher, T., 1987. Multi-channel seismic imaging of a crustal magma chamber along the East Pacific Rise. *Nature*, **326**, 35–41.

Detrick, R. S., Mutter, J. C., Buhl, P., and Kim, I. I., 1990. No evidence from multichannel reflection data for a crustal magma chamber in the MARK area on the Mid-Atlantic Ridge. *Nature*, **347**, 61–64.

http://en.wikipedia.org/wiki/List_of_tectonic_plates

Kamesh Raju, K. A., Chaubey, A. K., Amarnath, D., and Mudholkar, A. V., 2008. Morphotectonics of the Carlsberg Ridge between 62° 20' and 66° 20'E, northwest Indian Ocean. *Marine Geology*, **252**, 120–128.

Krishna, K. S., Gopala Rao, D., Ramana, M. V., Subrahmanyam, V., Sarma, K. V. L. N. S., Pilipenko, A. I., Shcherbakov, V. S., and Radhakrishna Murthy, I. V., 1995. Tectonic model for the evolution of oceanic crust in the northeastern Indian Ocean from the late cretaceous to the early tertiary. *Journal of Geophysical Research*, **100**, 20011–20024.

Liu, C. S., Curray, J. R., and McDonald, J. M., 1983. New constraints on the tectonic evolution of the Eastern Indian Ocean. *Earth and Planetary Science Letters*, **65**, 331–342.

Minshull, T. A., Bruguier, N. J., and Brozena, J. M., 2003. Seismic structure of the Mid-Atlantic Ridge, 8–9°S. *Journal of Geophysical Research*, **108**, 2513, doi:10.1029/2002JB002360.

Müller, R. D., Roest, W. R., Royer, J.-Y., Gahagan, L. M., and Sclater, J. G., 1997. Digital isochrons of the worlds ocean floor. *Journal of Geophysical Research*, **102**, 3211–3214.

Purdy, G. M., Kong, L. S., Christenson, G. L., and Solomon, S. C., 1992. Relationship between spreading rate and the seismic structure of mid-ocean ridges. *Nature*, **355**, 815–817.

Rona, P. A., 1980. TAG hydrothermal field: Mid-Atlantic Ridge crest at latitude 26°N. *Journal of the Geological Society of London*, **137**, 385–402.

Royer, J.-Y., Sclater, J. G., and Sandwell, D. T., 1989. A preliminary tectonic fabric chart of the Indian Ocean. *Proceedings of the Indian Academy of Sciences-Earth and Planetary Science*, **98**, 7–24.

White, R. S., McKenzie, D., and O'Nions, R. K., 1992. Oceanic crustal thickness from seismic measurements and rare earth element inversions. *Journal of Geophysical Research*, **97**, 19683–19715.

Cross-References

Earth's Structure, Global
Earth's Structure, Upper Mantle
Earthquakes and Crustal Deformation
Lithosphere, Oceanic
Lithosphere, Oceanic: Thermal Structure
Plate Driving Forces
Plate Motions in Time: Inferences on Driving and Resisting Forces
Seafloor Spreading
Seismic Structure at Mid-Ocean Ridges

OCEANIC INTRAPLATE DEFORMATION: THE CENTRAL INDIAN OCEAN BASIN

D. Gopala Rao[1], D. A. Bhaskara Rao[2]
[1]Geology Department, Osmania University, Hyderabad, India
[2]National Geophysical Research Institute, Hyderabad, India

Definition

Intense deformation of the oceanic crust in the middle of the Indo-Australian plate is not anticipated as per the theory of plate tectonics, because the plate's interiors are considered rigid, and deformation area is limited to narrow zones occurring at the plate boundaries only. Compression is an accepted mechanism for the deformation and related to plate end dynamics, spreading ridge push and slab drag at trench, its temporal and spatial evolution is still being debated. The recent scientific studies results should provide more evidences in support of early Neogene origin of the Himalaya and its backward push resisting northward motion of the Indian Plate resulting in compression and the deformation of large areal extant in the Indian Ocean.

Introduction

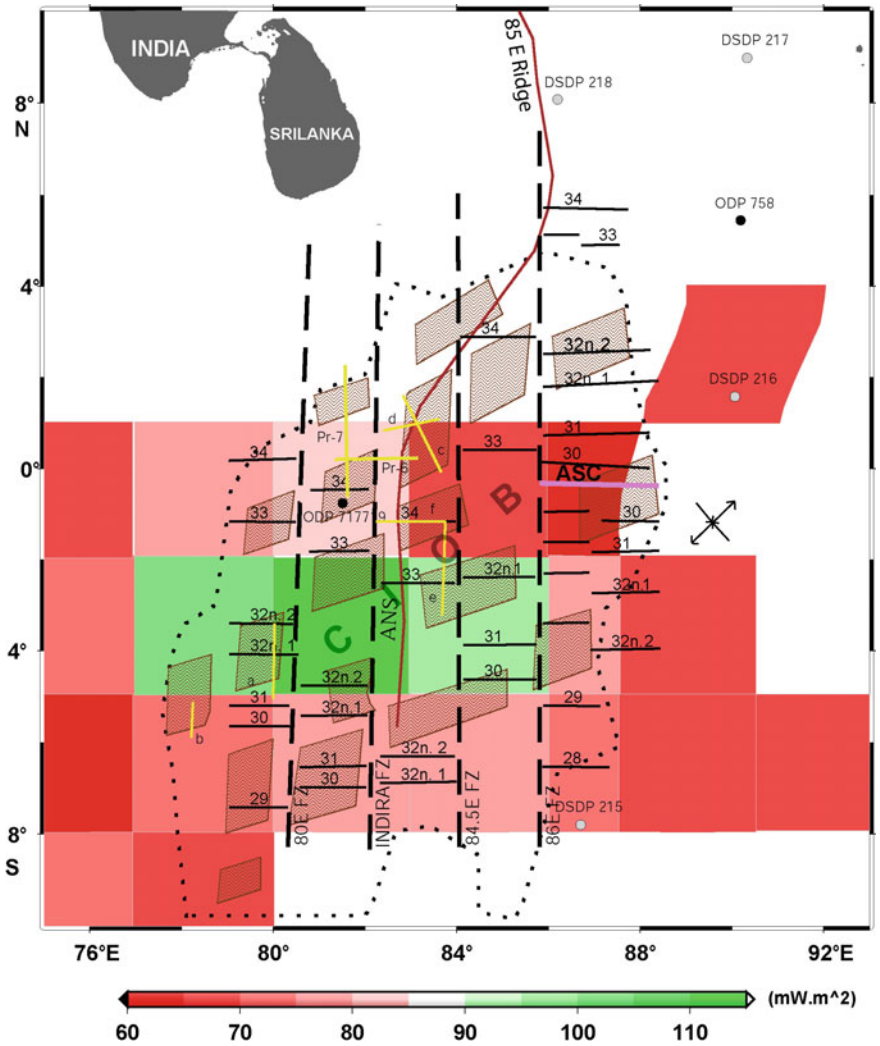
Deformation of the late Cretaceous crust of the Central Indian Ocean Basin (CIOB) in the middle of the Indo-Australian plate between 5° N and 10° S (Figure 1) is puzzling and is contrary to the paradigm that plate's interiors are rigid and deformation occurs at plate ends only. Seismic imaging of the upper crust and overlying sediments of the CIOB area have revealed intensely deformed crust consisting of anticline structures. Satellite Free-air gravity anomalies show E-W structural trends. The mid-plate deformation area is ~1,600 km wide and ~3,300 km long between the Central Indian Ridge in the west (CIR) and Sunda Trench (ST) in the east. Gordon et al. (1998) named it as "Diffuse Plate Boundary" between India and Australia. The putative boundary has deformed along its entire length as the plates jostle against each other (Kerr, 1988). But earthquake focal mechanism studies suggest north-south crustal deformation due to compression in CIOB and right-lateral strike-slip faults along the Ninetyeast Ridge (NyR), bounding the deformation area in the east (Stein and Okla, 1978). Microseismicity studies indicated episodic tectonic activity in the deformation zone (Neprochnov et al., 1988). Recent seismic reflection investigations have noted periodic deformation of Miocene (7.5–8.0 Ma), Pliocene (5.0– 4.0 Ma), and Pleistocene (0.8 Ma) ages due to in-plane (spreading ridge push from south) compressive forces (Krishna et al., 2001b). Closely spaced high-angle faults (Figures 2b and c) and tight folds are the characteristic features marking intense deformation. Subsidence of the crest of anticline structures against vertical faults (Figure 2d) indicates extension in compressional regime. Strike-slip faults with some dip-slip component near the CIR, right-lateral strike-slip faults and north-south compression of the CIOB, and left-lateral strike-slip faults and north-south compression of the Wharton Basin (WB) adjoining CIOB in the east are the striking dissimilarities in strain pattern noted from earthquake focal mechanism solutions (Delescluse and Chamot-Rooke, 2007). But the time of initiation of deformation in this region is not known for certain.

The characteristic anticline structures, 100–300 km long, 1–3 km high, and sediment-filled troughs, are bounded by longitudinal fractures zones (Fzs.) and east-west faults extending to basement (Figure 1). They are superposed by short wavelength E-W folds within the sediments at the crest (Figures 2a, c, and d) and high-angle reverse faults (Figure 2b). These are explicit expressions of ongoing tectonics of the deformed crust and overlying sediments. The style of deformation and strain pattern in the CIOB and at the three plate (India, Australia, and Capricorn) boundaries are site-specific and mark varied mechanisms of deformation and localized tectonics of the present and past times as well.

Several geophysical investigations including Deep Sea Drilling studies have been carried out in the area to understand physical processes and onset time of deformation. According to these studies, the deformation occurred by one of the following processes: (1) in-plane compression created by the plate end geodynamics and resistive forces, (2) thermo-mechanical subsidence under excess loading of igneous material initiating lithosphere flexure and eventually leading to crustal deformation under compression, or (3) periodic deformation due to resistive forces (backward push) due to the stationary Himalayas and plate end geodynamics – ridge push from the south and slab drag at the trench in the north. Even though, it is generally accepted that the mountain-building activity at the plate end in north and deformation of the CIOB are tectonically related, the Himalaya's temporal and spatial evolution has always been debated and it still remains a big scientific question. Furthermore, it is not clear whether the lower crust and entire lithosphere are involved in the deformation. The lower crustal (oceanic crust 3A) P-wave velocity (7.9 km/s) beneath the anticlines was found to be anomalously low (Louden, 1995; Neprochnov et al., 1998) and could not be unambiguously explained by any single physical process in the absence of further evidences. An overview of recent investigations on temporal evolution of the Himalayas and Tibet plateaus, structure of the northern Indian Ocean crust and sediments and sedimentary basins tectonics of the northwest continental margin of India (WCMI) is presented in the following, which shall provide some crucial evidences to address some of these points raised above.

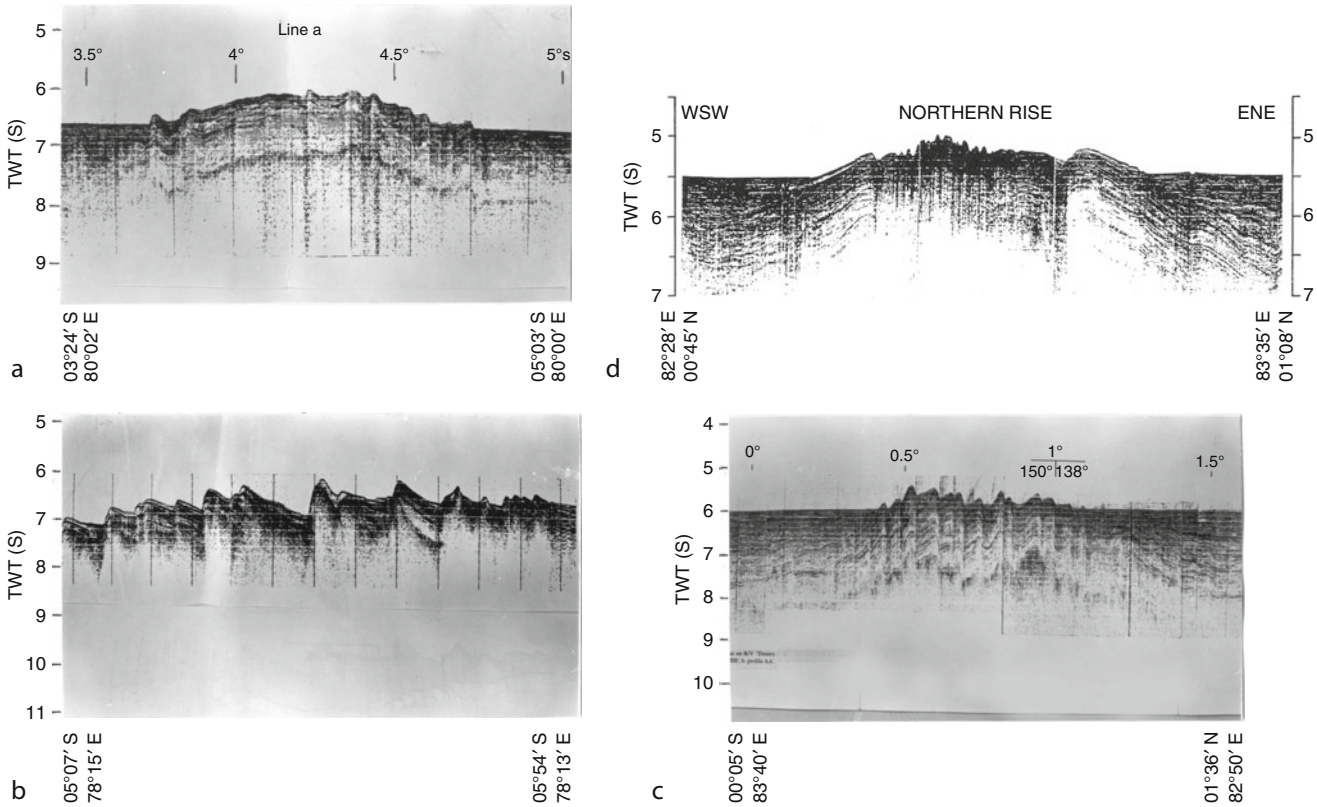
Temporal evolution of the Himalaya and Tibetan Plateau

Geologic structure and tectonics of the Tibetan Plateau have been investigated extensively to constrain temporal



Oceanic Intraplate Deformation: The Central Indian Ocean Basin, Figure 1 Mosaic of deformed crustal blocks. Heat flow budget shown in color coded rectangular grid. Squares with patterned lines inside are deformed blocks of the Central Indian Ocean Basin. Open and closed circles are Deep Sea Drilling Project (*DSDP*) and Ocean Drilling Project (*ODP*) drill sites, respectively. Black continuous lines with numbers are seafloor spreading magnetic isochrons. Black dotted (closed line) is the limit of the deformed blocks imaged. *FZ* fracture zone, *ASC* Abandoned Spreading Center, *ANS* Afanasy Nikitin Seamount. Arrows are compression/extension directions. (Modified after Gopala Rao et al. (2004), heat flow budget from Delescluse and Chamot-Rooke (2007)).

evolution of the Himalayas. The most significant observations are: (1) fossil leaf assemblages (Spicer et al., 2003) and stable isotope Paleo-altimetry studies (Rowley and Currie, 2006) of the southern and central Tibetan plateaus, respectively, which indicate that the plateau attained its present elevations, (~5,000 m) long before 8 Ma and sometimes between 15 and 35 Ma ago instead of rapid increase in height. (2) Rapid denudation of the Tibetan Plateau occurred during 21 to 15 Ma ago (Harrison et al., 1992), leading to flat surfaces of the plateau of the western Himalayas and the southwest Tibetan Plateau areas requiring the same geological processes and the time of formation (Unsworth et al., 2005), (3) the existence of high Himalayas prior to 15 Ma leading to push southwards causing the crust deformation in CIOB (Gordon, 2009). Paleogene granetoids, relicts of the intra-oceanic island-arc assemblages are observed beneath the Himalayas and Tibetan Plateau. Aitchison et al. (2007) have proposed the existence of Cretaceous back-arc basin island-arc and its closure ~20–25 Ma after the initial collision 50–55 Ma (Muller, 2010), (4) diachronous uplift of the Tibet Plateau (Chung et al., 1998) and trench migration (Capitanio et al., 2010) immediately after initial collision of the Indian Plate with Eurasian Plate infer the early mountain ranges building. The contemporaneous tectonics and geologic processes can be considered to support the early temporal evolution of the Himalayas and its impact on the deformation of the CIOB crust.



Oceanic Intraplate Deformation: The Central Indian Ocean Basin, Figure 2 (**a**–**d**): Seismic image of the deformed crustal blocks. (**a**) Long wavelength deformed crust shows primary and secondary deformations. (**b**) High angle thrust faults. (**c**) Long wavelength fold superposed by closely spaced folds bounded by vertical faults extending to basement. (**d**) Long wavelength fold with subsided crest part against normal faults. (Modified after Neprochnov et al. (1998)).

Crustal processes in the northern Indian Ocean and CIOB

Recent geophysical investigations have brought to light several anomalous crustal processes that could be explained by plate end geodynamics, and are discussed in the following sections.

The northern Indian Ocean

Pause in the seafloor spreading during Oligocene, 42 to 24 Ma, in the East Arabian Sea (EAS) (Chaubey et al., 1998), late Oligocene-mid Miocene turbidite sediments in the EAS (Clift et al., 2002; Gopala Rao, 2008) and in the Bay of Bengal (Gopala Rao et al., 1997; Gopala Rao, 2008), rapid subsidence of sedimentary basins of WCMI during late Oligocene and Miocene (Whiting et al., 1994), and the deformation of the CIOB crust/lithosphere prior to 20 Ma are contemporaneous and unique imprints noting active geological processes of the northern Indian Ocean. Lower Miocene age turbidites are possibly eroded from the high Himalayas and Tibetan Plateau and some of them reaching beyond equatorial region. Therefore, the anomalous crustal features shall mark the period of active tectonics in the northern Indian Ocean crustal processes.

The central Indian Ocean Basin (CIOB)

The CIOB crust in Figure 1 is late Cretaceous according to seafloor spreading magnetic isochrons 34 to 28 (Norton and Sclater, 1979). It is diachronous in age in the east as an extra oceanic crust was transferred between 86° E Fz (Fracture Zone) and NyR (parallel to 90° E) from Antarctic plate to Indian Plate during 60 and 42 Ma (Krishna et al., 1998). Furthermore, Abandoned Spreading Centers (ASCs) are also observed in the CIOB (Figure 1) and its adjoining regions. A 1.5-km high and 10-km wide structure associated with 8 mGal gravity high is present in center of a trough between anticlines associated with long wavelength gravity low of 20 mGal. It has been identified as serpentinite diapires structures of Paleogene age (Krishna et al., 2002). The low P-wave velocities of the oceanic layer 3A of the anticline structures infer serpentinites at subcrustal level.

Presence of serpentinites rocks along flanks of the volcanic Afanasy Nikitin Seamount (ANS), an edifice (~3.5 km in length, ~1.6 km wide and ~3.5 km high) of subaerial origin of the late Cretaceous in the middle of the deformation area, has been reported by Murdmaa et al. (1998). The tectono-magmatism, spreading ridge migration onto earlier evolved oceanic crust and presence of heterogeneous

(semi-ductile-serpentinites) crustal rocks point to reheating of the diachronous age crustal rocks in the CIOB area during the geological past (Figure 1). Due to the ongoing tectonism the crustal layers response might vary resulting varied physical properties of the lower crust rocks and contributing to selective deformation in the anomalous area. Kasintsev and Sborshchikov (1998) from analyses of the magmatic rocks collected along flanks of the ANS during MIR submersible cruise have noted evidences for continued tectono-magmatism since Paleocene. The tectonism might have as well prevailed in the CIOB as the ANS is placed in the middle of the deformed zone under compression. The diapires are characteristic features of the deformed crust blocks that evidently confirm compression tectonics as major component (Figures 2a and b) and early tectonics. A mosaic of the deformed crustal blocks/anticlines imaged in the CIOB (Neprochnov et al., 1998) and superposed gridded heat flow (Delescluse and Chamot-Rooke, 2007) (Figure 1) show wide deformation around the chain of the ANS of large dimension and excess heat flow around the emplaced igneous rocks.

Pelagic (pre-fan sediments) and pre-deformation sediments above the basement are ~1.6 s and 1.2–1.4 s thick, respectively. They are 1.5–2.2 s (TWT) thick in the south and 2.3–2.6 s in the north of the deformation area (Levchenko and Milanovsky, 1998) (Figures 2a–d). Numerous elongated seafloor rises of 15–20 m highs and about 45 km wide features are superposed along crests and flanks of the deformed blocks (Figure 2a).

The NE to ENE – SW to WSW trends of the free-air satellite gravity anomalies of the Wharton Basin (WB) to the east of the CIOB (Louden, 1995) differ from E-W trends of the CIOB. Furthermore, the anticline structures consist of low velocity (~7.9 km/s) oceanic crust, layer 3b (Figure 3a) (Neprochnov et al., 1998). High average heat flow of 65 mW/m², which is 12 mW/m² higher than average of 54 mW/m² for the 70 Ma age crust of the CIOB (vary from 44 to 161 mW/m²), was observed by Verzhbitsky and Lobkovsky (1998). Each deformed blocks heat flow and the high anomalous heat flow associated with the ANS reflect excess heating of rocks of the CIOB area.

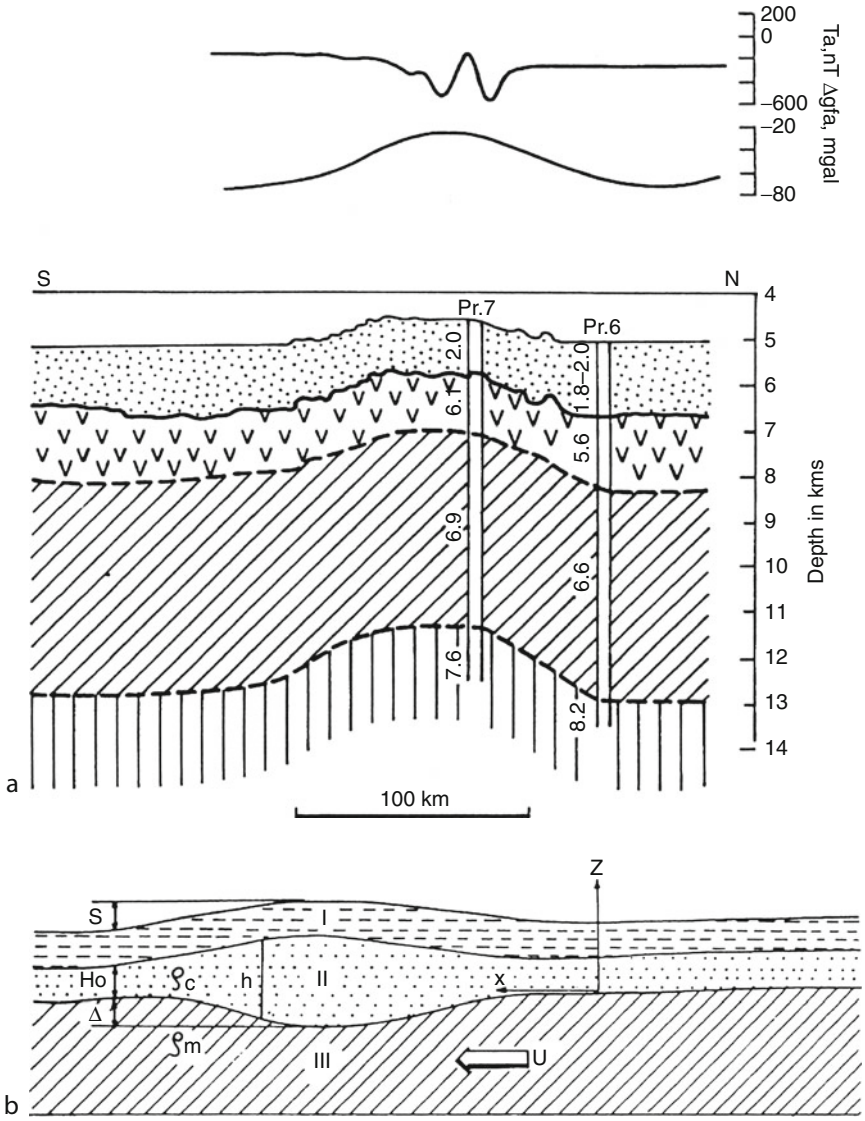
The mass sediment fluid flows, turbidites of the northern Indian Ocean, seafloor spreading processes in the EAS, and rapid subsidence of the WCMI sedimentary basins mark intraplate kinematics. The high mountain ranges' eroded sediments have reached the ocean basins since beginning of Neogene. The Cretaceous Paleo-Tethys Ocean (comprises of continental margin, mid-plate volcanic island-arc and back-arc basin) volcanic island-arc subduction was suggested 20–25 Ma later to initial plate's collision (Muller, 2010). So, it is possible that there was pause in plate motions due to the subduction of island-arc during the initial stage (locking) as it is likely to diminish northward slab pull decelerating India's northward motion and seafloor spreading in the EAS. On completion of subduction of the arc and closure of Cretaceous back-arc basin (unlocking) the subduction slab pull force had resumed at plate boundary in the north, which might have allowed resumption in seafloor spreading in EAS. Therefore, the high raised Himalaya's erosion has contributed to the turbidite sediments that are traced up to 8° S in the northeast Indian Ocean. Periodic crustal deformation was observed in CIOB and was attributed to the Himalaya's Tectonics (Krishna et al., 2009b). Paleogene granetoids of island-arc beneath the Himalayas/Tibetan Plateau mark suturing process at plate end in the north prior to the subduction. The turbidites, seafloor spreading, and the basin's tectonics are result of plate end geodynamics at least since ~30 Ma. The onshore and offshore geologic records can be considered to mark tectonics of the Himalayas; especially the outward push shall indicate crustal deformation (extension and compression) of the CIOB crust in a wide area. It implies that the selective deformation in CIOB is due to anisotropic geologic structure, which favored the deformation besides plate end geodynamics.

The observed intense intraplate oceanic crust deformation and the associated geophysical signatures can be explained assuming three-layer oceanic crust consisting of oceanic layers 2, 3A, 3B (Figure 3a), and two-level plate tectonic model (Figure 3b) similar to the one suggested earlier by Lobkovisky (1998). The low velocity viscoelastic serpentinites creep reported in the lower crust might have evolved either in situ due to dewatering during the deformation process or was part of crust/lithosphere during the seafloor spreading process and moved northward as a part of oceanic plate. The changes in rheology and the presence of the viscoelastic creep in the form of serpentinites possibly favored the selective deformation in the middle of the large Indo-Australian plate, culminating in the spatial and temporal variation of deformation and their geophysical signatures. Beneath the Tibetan plateau also mid-crustal fluid flows are inferred from resistivity and seismic reflection images and was inferred for the tectonics and uprise of the Himalayas/Tibet (Unsworth et al., 2005). The records of folded sediments along flanks and subsidence at the crest of the deformed blocks indicate later stage tectonics, reactivation, and deformation. During the ongoing tectonic process, the penetration of faults to subcrustal lithosphere, thrust faults reactivation, and development of series of syn-genetic faults at upper brittle crustal levels might have occurred.

The fault and structural style, temporal and spatial variations in compression and stress/strain distribution, and the long and short wavelength deformations of the crustal blocks are explicit expressions of the regional and localized tectonic processes. Despite the best effort to unravel the temporal and spatial evolution of deformation, it is not yet possible to explain whether the deformation is confined to crustal (upper?) level only or the entire lithosphere is involved.

Summary

Intense deformation of the late Cretaceous oceanic crust of the Central Indian Ocean Basin (CIOB) in the Indian Ocean equatorial region, between the Central Indian Ridge



Oceanic Intraplate Deformation: The Central Indian Ocean Basin, Figure 3 (**a**–**b**): (**a**) Illustrates modeled three layered crustal structure associated gravity and magnetic anomalies and refraction velocities across deformed crustal block. Dotted-sediments, v-oceanic layer 2, oblique lines-oceanic layer 3 and vertical lines-mantle. Numbers-refraction velocities in km/s obtained along profiles across deformed block. (**b**) Crustal model of two-level plate tectonics. ρ_m and ρ_c are the densities of the mantle and crust, respectively. I, II, and III are upper crust, lower crust, and mantle, respectively. *Open arrow* shows direction of the compressional stress. S-amount of deformation at the top of the upper crustal level. Δ-amount of deformation at mantle. *h* and H_0 are thickness of viscoelastic layer beneath the deformed block and in case of normal crust, respectively. (Modified after Lobkovisky (1998) and Neprochnov et al. (1998)).

(CIR) in the west and the Sunda Trench (ST) in the east, is noted for ongoing, periodic, and plate end geodynamics especially resistive forces of the high Himalayas against the northward moving Indian plate oceanic crust.

The tectono-magmatism, spreading ridge southward migration onto the already evolved late Cretaceous oceanic crust and large extrusion of volcanics in the form of voluminous seamounts of the area during early Tertiary and Eocene are unique geodynamic crustal processes that might as well contribute to altered thermal regime of the past and changes in crustal rocks properties responding differently to external forces.

The numerous onshore and offshore proxy evidences shall support the early Neogene origin of the Himalayas, which contributed to the southward push onto northward moving oceanic crust since their formation.

The question unanswered is whether deformation is limited to crust only or entire lithosphere, which needs to be understood by probing deeper parts of the lithosphere as it is classical example of crustal deformation.

Bibliography

Aitchison, J., Ali, J. R., and Davis, A. M., 2007. When and where did India and Asia collide? *Journal of Geophysical Research*, **112**, 1–19, doi:10.1029/2006 JB 004706.

Capitanio, F. A., Morra, G., Goes, S., Weinberg, R. F., and Moresi, L., 2010. India-Asia convergence driven by subduction of the Greater Indian continent. *Nature Geoscience*, **3**, 136–139, doi:10.1038/NGEO725.

Chaubey, A. K., Bhattacharya, G. C., Murthy, G. P. S., Srinivas, K., Ramprasad, T., and Gopala Rao, D., 1998. Early teritiary seafloor spreading magnetic anomalies and paleo-propogators in the northern Arabian Sea. *Earth Planetary Science Letters*, **154**, 41–52.

Chung, S. L., et al., 1998. Diachronous uplift of the Tibetan plateau starting 40 Myr ago. *Nature*, **394**, 769–773.

Clift, P., Gaedicke, C., Edwards, R., Lee, Jae Il, Hildebrand, P., Amjad, S., White, R. S., and Schluter, H.-U., 2002. The stratigraphic evolution of the Indus Fan and the history of sedimentation in the Arabian Sea. *Marine Geophysical Research*, **23**, 223–245.

Cochran, J. R., 1990. Himalayan uplift, sea level, and the record of Bengal Fan sedimentation at the ODP Leg 116 sites. In Cochran, J. R., et al. (eds.), *Proceedings of Ocean Drilling Program, Scientific results*. Ocean Drilling Program, Texas, Vol. 116, 397–414.

Delescluse, M., and Chamot-Rooke, N., 2007. Instantaneous deformation and kinematics of the India-Australia Plate. *Geophysical Journal International*, **168**, 818–842.

Gopala Rao, D., Krishna, K. S., and Sar, D., 1997. Crustal evolution and sedimentation history of the Bay of Bengal since Cretaceous. *Journal of Geophysical Research (B: Solid Earth)*, **102**, 17747–17768.

Gopala Rao, D., Krishna, K. S., Neprochnov Yu, P., and Grinko, B. N., 2004. Satellite gravity anomalies and crustal features of the Central Indian Ocean Basin. *Current Science*, **86**, 948–957.

Gopala Rao, D., 2008. *Marine Geophysics in India*. Memior 68. Geological Society of India, 1–250.

Gordon, R. G., DeMets, C., and Royer, J.-Y., 1998. Evidence for long-term diffuse deformation of the lithosphere of the equatorial Indian Ocean. *Nature*, **395**, 370–374.

Gordon, R. G., 2009. Lithospheric deformation in the equatorial Indian Ocean. *Timing and Tibet Geology*, **37**(3), 287–288, doi:10.1130/focus032009.1.

Harrison, T. M. P., Copeland, W. S., Kidd, F., and Yin, A., 1992. Raising Tibet. *Science*, **255**, 1663–1670, doi:10.1126/Science. 255.5052.1663.

Kasintsev, G. L., and Sborshchikov, I. M., 1998. Magmatic rocks and intraplate volcanism. In Neprochnov, Yu, P., Gopala Rao, D., Subrahmanyam, C., and Murthy, K. S. R. (eds.), *Intraplate Deformation in the Central Indian Ocean Basin*. Bangalore: Geological Society of India Memoir, 39, pp. 125–131.

Kerr, R. A., 1988. Puzzling out the tectonic plates. *Science*, **247**, 808.

Krishna, K. S., Ramana, M. V., Gopala Rao, D., Murthy, K. S. R., Rao, M. M. M., Subrahmanyam, V., and Sarma, K. V. L. N. S., 1998. Periodic deformation of oceanic crust in the central Indian Ocean. *Journal of Geophysical Research*, **103**, 17859–17875.

Krishna, K. S., Bull, J. M., and Scrutton, R. A., 2001b. Evidence for multiphase folding of the central Indian Ocean lithosphere. *Geology*, **29**, 715–718.

Krishna, K. S., Gopala Rao, D., and Neprochnov, Y. P., 2002. Emplacement of serpentinite rocks near the oceanic basement in the deformation zone, central Indian Ocean: a model from gravity and seismic reflection data. *Proceedings Indian Academy of Sciences*, **111**, 17–28.

Krishna, K. S., Bull, J. M., and Scrutton, R. A., 2009b. Early (pre-8 Ma) fault activity and temporal strain distribution in the central Indian Ocean. *Geology*, **37**, 227–230.

Levchenko, O. V., and Milanovsky, V. E., 1998. Structural style of intraplate deformation area. In Neprochnov, Yu, P., Gopala Rao, D., Subrahmanyam, C., and Murthy, K. S. R. (eds.), *Intraplate deformation in the Central Indian Ocean Basin*. Bangalore: Geological Society of India Memoir, 39, pp. 125–131.

Lobkovisky, L. I., 1998. Concept of two-level plate model. In Neprochnov, Yu, P., Gopala Rao, D., Subrahmanyam, C., and Murthy, K. S. R. (eds.), *Mechanism of heating up of the Indo-Australian Plate*. Bangalore: Geological Society of India Memoir, 39, pp. 210–214.

Louden, K. E., 1995. Variations in crustal structure related to intra plate deformation; evidence from seismic refraction and gravity profiles in the Central Indian Basin. *Geophysical Journal International*, **120**, 375–392.

Murdmaa, I. O., Levitan, M. A., Reddy, N. P. C., Emelyanov, E. M., and Jafri, S. H., 1998. Sedimentary geology of the deformation zone. In Neprochnov, Yu, P., Gopala Rao, D., Subrahmanyam, C., and Murthy, K. S. R. (eds.), *Intraplate deformation in the Central Indian Ocean Basin*. Bangalore: Geological Society of India Memoir, 39, pp. 177–186.

Müller, R. D., 2010. Sinking continents. *Nature Geosciences*, **3**, 79–80.

Neprochnov, Yu, P., Levchenko, O. V., Merklin, L. V., and Sedov, V. V., 1988. The structure and tectonics of the intraplate deformation area in the Indian Ocean. *Tectonophysics*, **156**, 89–106.

Neprochnov, Yu, P., Gopala Rao, D., Subrahmanyam, C., and Murthy, K. S. R., 1998. *Intraplate deformation in the Central Indian Ocean Basin*. Bangalore: Geological Society of India Memoir, 39, pp. 1–250.

Norton, I. O., and Sclater, J. G., 1979. A model for the evolution of the Indian Ocean and the breakup of Gondwanaland. *Journal of Geophysical Research*, **84**, 6803–6830.

Rowley, D. B., and Currie, B. S., 2006. Paleo-altimetry of the late Eocene to Miocene Lunpola basin, central Tibet. *Nature*, **39**, 677–681, doi:10.1038/nature04506.

Spicer, R. A., Harris, N. B. W., Widdowson, M., Herman, A. B., Guo, S., Valdes, P. J., Wolfe, J. A., and Kelley, S. P., 2003. Constant elevation of southern Tibet over the past 15 million years. *Nature*, **421**, 622–624, doi:10.1038/nature01356.

Stein, S., and Okla, E., 1978. Seismicity and tectonics of the Ninetyeast Ridge area evidence for internal deformation of the Indian plate. *Journal of Geophysical Research*, **83**, 2233–2245.

Unsworth, M. J., Jones, A. G., Wei, W., Marquis, G., Golan, S. G., Spratt, S. G., and the INDEPTH-MT team, 2005. Crustal rheology of the Himalayas and southern Tibet inferred from magneto-telluric data. *Nature*, **438**, 78–81, doi:10.1038/nature0415.

Verzhbitsky, E. V., and Lobkovsky, L. I., 1998. Heat flow studies of the northeast Indian Ocean. In Neprochnov, Yu, P., Gopala Rao, D., Subrahmanyam, C., and Murthy, K. S. R. (eds.), *Intraplate deformation in the Central Indian Ocean Basin*. Bangalore: Geological Society of India Memoir, 39, pp. 77–81.

Whiting, B. M., Karner, G. D., and Driscoll, N. W., 1994. Flexural and stratigraphic development of the west Indian continental margin. *Journal of Geophysical Research, (B: Solid Earth)*, **99**(B7), 13791–13811.

Cross-references

Earth's Structure, Global
Lithosphere, Oceanic
Ocean, Spreading Centre
Plate Driving Forces
Seafloor Spreading

P

PALEOMAGNETIC FIELD INTENSITY

Andrew Biggin, Neil Suttie, John Shaw
Geomagnetism Laboratory, University of Liverpool, Liverpool, UK

Synonyms

Archeointensity; Paleointensity

Definition

Absolute paleointensity: a paleomagnetic measurement of geomagnetic field intensity recorded at the time and place that a measured material acquired a thermoremanent magnetization.

Archeointensity: an absolute paleointensity estimate produced from a material (frequently, a heated archeological artifact) which acquired its thermoremanent magnetization in archeological times.

Relative paleointensity: the record of relative geomagnetic intensity variations measured from the natural remanent magnetization of sedimentary rock samples.

Cosmogenic isotope paleointensity: the record of relative geomagnetic intensity variations measured using the concentration of cosmogenic isotopes as a proxy.

Introduction

Direct measurements of the intensity of the geomagnetic field only became possible in 1832 when a method of obtaining them was developed by Carl Frederich Gauss (Stern, 2002). If we wish to know how the field intensity has varied before this time, then we are required to use the geological and archeological records to make indirect measurements. Paleomagnetic approaches are based on the principle that, for weak fields such as the Earth's, the intensity (M_{anc}) of the natural remanent magnetization (NRM) recorded by a material is linearly proportional to the strength (H_{anc}) of the paleofield. For absolute paleointensity measurements, the material in question must have acquired a thermoremanent magnetization (TRM) whereas, for relative paleointensity, it is required to be a depositional (DRM) or postdepositional (pDRM) remanent magnetization. Unfortunately, for all types of remanence, the constant of proportionality linking the magnitude of the NRM to the paleofield intensity is a sensitive and complicated function of the magnetic properties of the recording material and extremely difficult to measure directly. Absolute paleointensity experiments obviate this difficulty by imparting the sample with a new TRM (M_{lab}) in the laboratory by heating the sample up and cooling it in a controlled field of known intensity (H_{lab}). The linear relationships between M and H can then be combined to eliminate the constant and solve for the ancient field intensity:

$$H_{anc} = \frac{M_{anc}}{M_{lab}} H_{lab} \quad (1)$$

Depositional-type remanences cannot be reliably reproduced in the laboratory and therefore paleointensity experiments performed on sediments produce records only of relative *variations* in H_{anc}.

In order for any type of paleointensity measurement to be useful, an independent estimate for the age of the magnetization (often assumed or demonstrated to be the same age as the material) is usually required.

Absolute paleointensity measurements

Absolute paleointensity experiments may only be performed using materials which have cooled from elevated temperatures in nature and which therefore possess a TRM imparted by the ambient (geomagnetic) field. Materials that may be used include any fired archeological materials (ceramics, kiln lining, hearths, burnt bricks,

Harsh K. Gupta (ed.), *Encyclopedia of Solid Earth Geophysics*, DOI 10.1007/978-90-481-8702-7,

etc.), most types of igneous rocks, and metamorphic rocks (e.g., baked sediments), which have been extensively reheated.

Producing unequivocally accurate measurements of the absolute paleointensity is extremely demanding since there are a large number of factors which can potentially bias the estimate. The list below describes the criteria that a hypothetically ideal absolute paleointensity estimate would meet.

1. The remanence carriers must retain a pure TRM and have undergone no chemical or physical changes since cooling beneath their Curie temperature (see e.g., Fabian, 2009).
2. Any secondary components of magnetization (e.g., viscous or isothermal overprints, see *Remanent Magnetism*) must be removed so that the primary TRM is isolated.
3. The remanence carriers must undergo no chemical or physical alteration as a consequence of the laboratory heating that is required to obtain the paleointensity (or this must be corrected for; see, e.g., Valet et al., 1996).
4. The remanence carriers should be noninteracting single domain grains or at the least, it must be shown that multidomain, vortex state, or magnetically interacting grains are not biasing the measurement by violation of Thellier's laws of thermoremanence.
5. Any anisotropy of TRM must be avoided or corrected for (see e.g., Selkin et al., 2000).
6. Any effect that the difference in the cooling rate between the natural and laboratory cooling cycles has on the calculated paleointensity must be corrected for (see e.g., Perrin, 1998).
7. Any deviation from Thellier's law of linearity must be detected and corrected for (Selkin et al., 2007).
8. The number of individual sample estimates from a single rock unit or archeological horizon must be sufficient to obtain a reasonable mean and to demonstrate internal consistency (see e.g., Biggin et al., 2003).

The high failure rates and time-consuming nature of many absolute paleointensity studies has stimulated a wealth of innovation in terms of both the types of materials which are used and the manner in which the experiments are performed. In the former respect, two noteworthy materials are submarine basaltic glass (SBG) generally taken from ocean drilling cores and ophiolites (Tauxe, 2006), and single silicate crystals isolated from both extrusive and intrusive igneous rocks (Tarduno et al., 2006). Both of these have been argued to be higher fidelity recorders (less prone to problems numbered 1, 3, 4, and/or 6 in the list above) than more conventional paleointensity recorders (e.g., whole-rock subaerial lava samples).

In addition to providing an estimate, the choice of absolute paleointensity methodology is also made to enable verification of some or all of the criteria that are listed above. To some extent, this may be achieved directly through the paleointensity experiment itself, but it is also very common for supporting rock magnetic, microscopic, and/or paleodirectional experiments to be undertaken for this purpose. Table 1 summarizes some of the more commonly used absolute paleointensity techniques and provides the relevant references. The Thellier method (Thellier and Thellier, 1959) and its derivatives remains the most widely used and trusted technique. These are based on stepwise removal of the NRM alongside simultaneous stepwise replacement with a laboratory TRM. For a general introduction to this method and some of the other techniques listed in Table 1, the reader is directed to comprehensive reviews by Tauxe and Yamazaki (2007) and Valet (2003).

Relative paleointensity measurements

Paleomagnetic analyses of sedimentary rock and soft sediment samples can produce continuous time series of geomagnetic intensity variations, which may be calibrated using absolute paleointensity estimates or used alone as records of relative geomagnetic behavior. Relative paleointensity studies are generally undertaken on the assumption that the samples retain a depositional remanent magnetization (DRM) acquired as the magnetic grains in the sediment physically align themselves during or shortly after deposition from the water column. The DRM acquisition process is sensitively related to the size, shape, and composition of the remanence carriers and is further complicated by flocculation effects in saline environments. Much later, the intensity of the magnetization can be significantly affected by consolidation and compaction processes (see review by Tauxe and Yamazaki, 2007).

Some normalization of the measured magnetization intensity is required to account for changes (which should ideally be small) in the concentration and nature of the magnetic recorders and produce a record of geomagnetic intensity variations. Candidates for normalization include anhysteretic remanent magnetization (ARM), isothermal remanent magnetization (IRM), and low-field susceptibility. These should ideally display a strong coherence with the NRM signal (after removal of any secondary components of magnetization) but no coherence with the normalized record (Valet, 2003). The reliability of a relative paleointensity record can be further supported by the presence of antipodal directions (suggesting that secondary components of magnetization have been removed) and by the coherence of multiple curves from the same region covering the same time interval (as calibrated using independently obtained timescales).

The sedimentation rate and the interval over which the magnetization is "locked in" sets a limit on the time resolution of any relative paleointensity record. In deep-marine sediments, a standard 2.5 cm paleomagnetic sample can represent a time period of 10 kyr or more,

Paleomagnetic Field Intensity, Table 1 Selection of absolute paleointensity methods currently in use

Family	Variants	Benefits	Criticisms
Thellier (Thellier and Thellier, 1959)	Coe (1967), Aitken et al. (1988), Perpendicular (Kono and Ueno, 1977), IZZI (Tauxe and Staudigel, 2004), QP (Biggin et al. 2007)	Sound theoretical basis for single-domain grains (Néel, 1955); potential to recover estimate from a sample which both contains secondary magnetizations at low temperatures and which alters at high temperatures; additional checks for alteration (Prévot et al., 1981) and multidomain behavior (Riisager et al., 2001; Krasa et al., 2003) can easily be added	Multiple heatings are time consuming and increase likelihood of sample alteration; bias due to multidomain effects may be prevalent in lava samples (Biggin, 2010)
Microwave (Shaw et al., 1996; Hill and Shaw, 2000)	All Thellier-type variants may be used	In principle, all benefits of Thellier but with reduced sample heating in terms of both temperature and time; faster experiments which can be individually tailored to samples	Requires highly specialist equipment (microwave paleointensity system); theoretical basis not yet firmly established; samples may not unblock as cleanly as in thermal experiments
Shaw (1974)	ARM-alteration corrections (Kono, 1978; Rolph and Shaw, 1985); Double heating (Tsunakawa and Shaw, 1994); low-temperature demagnetization (Yamamoto et al., 2003); Microwave (Yamamoto and Shaw, 2008)	Original method is fast; arguably domain-state independent as uses full TRMs	The use of ARMs to check and correct for alteration in terms of the TRM properties is not always valid (Tanaka and Komuro, 2009)
High-temperature measurement	Wilson (1961) Le Goff and Gallet (2004)	Very fast methods; Wilson method is arguably domain-state independent as it uses full TRMs (Muxworthy, 2010); Le Goff–Gallet method designed to incorporate cooling rate and anisotropy corrections	Requires specialist equipment (high-temperature magnetometer); samples must survive heating to Curie temperature without altering in Wilson method; unknown effect of multidomain grains in Le Goff–Gallet method
Multispecimen (Hoffman et al., 1989)	Hoffman and Biggin (2005), Dekkers and Böhnel (2006)	Fast; *inter*-specimen consistency is implicitly considered; domain state effects can be minimized in Dekkers–Böhnel experiments; fewer heating steps implies less alteration	Fewer checks for *intra*-specimen consistency; no internal checks for alteration; domain-state benefits of Dekkers–Böhnel method may be lost if secondary magnetizations require removing (Michalk et al., 2010)
IRM Normalization (Cisowski and Fuller, 1986)		No heating required	Precision limited to orders of magnitude

which will produce significant smoothing of secular variation (see *Geomagnetic Field, Secular Variation*).

Cosmogenic isotope paleointensity measurements

This method of paleointensity measurement, applicable to the last few hundreds of kyr, is based on the fact that production rate of certain radioisotopes (^{14}C, ^{36}Cl, and ^{10}Be) is strongly influenced by the degree of shielding the geomagnetic field provides the Earth from incoming cosmic rays. Records of the concentration of these isotopes in sediment and ice cores (and tree rings and corals in the case of ^{14}C) can therefore provide a proxy for geomagnetic dipole moment once the necessary corrections for latitude and relevant physical, chemical, and biological processes have been made. See review by Valet (2003) for more information.

Geophysical implications

Paleointensity measurements provide information about the geodynamo across a variety of timescales. Direct observations indicate that the strength of the main field has fallen by around 10% since the first measurements were made in the 1830s. A key aim of archeointensity studies has been to establish the longevity of this recent trend. A recent analysis of the global dataset of archeointensity data from 1600 onward showed little evidence of a rapidly decaying field prior to 1840 (Gubbins et al., 2006). Longer time series

have been constructed for specific regions and Genevey et al. (2009) show periods of rapidly changing strength in France over the past 800 years. Archeointensity and paleointensity data are particularly important for constructing global models of field behavior on millennial timescales (e.g., CALS7K.2; Korte and Constable, 2005). Each of the observable paleomagnetic field components (declination, inclination, and intensity) can be mapped, via their respective Green's functions, to the radial field at the core–mantle boundary (Johnson and Constable, 1997). Intensity measurements sample the field at the core–mantle boundary at different locations (tending to higher latitudes) than directional data from the same site (Johnson and Mcfadden, 2007). Recently, updated databases of paleointensity estimates include measurements from the last 10 kyr (Genevey et al., 2008), 50 kyr (Korhonen et al., 2008), and the period spanning 50 kyr to 3.5 Gyr (Biggin et al., 2010).

Relative paleointensity records from varved sediments are capable of providing good temporal resolution over the Holocene and agree well with absolute intensity records as well as records of cosmogenic nuclide production (Snowball and Sandgren, 2002). Correspondingly, absolute archeointensities have been combined with the ^{14}C record from tree rings and the ^{10}Be record from ice cores to infer variation in solar activity, with implications for understanding solar forcing of climate (Solanki et al., 2004). Marine sediments have provided relative paleointensity records over longer timescales. Here, it is usually assumed that a typical sample will represent a sufficiently long period for nondipole features to be averaged out (Valet, 2003) allowing composite stacks of relative paleointensity from different areas of the world to be constructed. A global stack for the past 2 Myr, Sint 2000 (Valet et al., 2005), agrees well with absolute paleointensity from lava flows and confirms a low field strength during excursions. Importantly, this record is long enough to include reversals and finds a correlation between polarity interval length and field strength., This same correlation is seen, albeit weakly, in the 11 Myr Oligocene record from site Deep Sea Drilling Project, site 522 (Constable et al., 1998).

It becomes increasingly difficult to obtain relative paleointensities from sediments more than a few million years old. Consequently, longer term variations in field strength are studied using absolute paleointensity data derived from igneous rocks. By studying paleointensity throughout the Phanerozoic, it may be possible to understand the relationship between reversal frequency and intensity. Theoretical models have been proposed that link a stable nonreversing field with a high dipole moment (Glatzmaier et al., 1999; Driscoll and Olson, 2009). On the other hand, by considering properties of self-exiting dynamos, Hide (2000) finds a high field strength is likely to promote reversal. The study of paleointensity during superchrons is a natural approach to testing these models (Fuller and Weeks, 1992). A fertile field of study has been the Cretaceous normal superchron (CNS) but the suite of results obtained so far has as yet been unable to settle the issue. Pick and Tauxe (1993) found samples of submarine basaltic glass (SBG) from the CNS gave paleointensities less than half the present day value today (although later SBG data gave higher values; Tauxe and Staudigel, 2004). Prévot et al. (1990) found evidence of a longer period of low field intensity which they called the *Mesozoic dipole low*, but only one result in their analysis came from the CNS. Subsequent analyses have argued both for (Perrin and Shcherbakov, 1997; Biggin and Thomas, 2003) and against (Selkin and Tauxe, 2000; Goguitchaichvili et al., 2002) the existence of the Mesozoic Dipole Low. Studies based on the measurement of single silicate crystals have found evidence of high fields during the CNS (Tarduno et al., 2001; Tarduno and Cottrell, 2005) and a clear inverse relationship between reversal frequency and dipole moment. Other more recent studies based on whole rock samples have not confirmed this (Granot et al., 2007; Hill et al., 2008) although it has been claimed that these data are biased (Tarduno and Cottrell, 2005). There remains no consensus regarding the magnitude of the field during a superchron. However, the hypothesis of a positive relationship between reversal frequency and dipole moment (i.e., weaker field strength in superchrons) has now arguably been rejected (Biggin and Thomas, 2003). Tauxe (2006) argued that the average paleointensity is proportional to the length of the *specific* chron from which it is derived as opposed to the *average* length of chron during the time period in question. This is a subtle but important difference from the case generally presented.

The dynamo process producing the geomagnetic field is powered, in large part, by the growth of the inner core. The nucleation and subsequent growth of the inner core may then have had a profound effect on the geomagnetic field strength (Stevenson et al., 1983). Estimates for the age of the inner core range from 3.5 billion years (Gubbins et al., 2004) to 1 billion years (Labrosse et al., 2001) or younger, so paleointensities from Archean to Proterozoic rocks may carry a signal emanating from the onset of its crystallization. Studies on rocks of this age have been hampered by the alteration of the remanence carriers and once again no clear picture has yet emerged. Studies based on single silicate crystals from Paleoarchean (3.2–3.5 Ga) intrusive rocks suggest a dipole moment in the range of 50–70% of today's value (Tarduno et al., 2007, 2010). This could be taken as indicating either an early age for the onset of inner core nucleation or that thermal convection alone was capable of producing a field with intensity comparable to today's. In contrast, the paleointensity record has been interpreted as indicating a significant increase in field strength sometime between 1000 and 300Myr, suggesting a later formation of the inner core (Macouin et al., 2004). Biggin et al. (2009) raised the possibility that both observations were true and that a *Proterozoic dipole low* separates periods of dominantly high field strength in the Archean and Phanerozoic.

Paleointensity experiments have not been confined to terrestrial rocks. Lunar rocks from the Apollo missions

showed what appeared to be a thermoremanence and a variety of paleointensity techniques seemed to hint at a lunar dynamo operating between 3.9 and 3.6 Gyr (Cisowski, 1982). However, recently the reliability of the lunar paleointensity record has been questioned and the evidence for a lunar dynamo reconsidered (Lawrence et al., 2008). Paleointensities from meteorites have been used to estimate the ancient Martian field (Shaw et al., 2001; Weiss et al., 2008b) and may even be evidence of dynamos operating within the metallic cores of planetisimals in the early solar system (Weiss et al., 2008a).

Bibliography

Aitken, M. J., Allsop, A. L., Bussell, G. D., and Winter, M. B., 1988. Determination of the intensity of the Earth's magnetic field during archeological times: reliability of the Thellier technique. *Reviews of Geophysics*, **26**, 3–12.

Biggin, A., McCormack, A., and Roberts, A., 2010. Paleointensity database updated and upgraded. *Eos, Transactions American Geophysical Union*, **91**, 15.

Biggin, A. J., 2010. Are systematic differences between thermal and microwave Thellier-type palaeointensity estimates a consequence of multidomain bias in the thermal results? *Physics of the Earth and Planetary Interiors*, **180**, 16–40.

Biggin, A. J., Böhnel, H. N., and Zuniga, F. R., 2003. How many paleointensity determinations are required from a single lava flow to constitute a reliable average? *Geophysical Research Letters*, **30**(11), 10.1029/2003GL017146.

Biggin, A. J., Perrin, M., and Shaw, J., 2007. A comparison of a quasi-perpendicular method of absolute palaeointensity determination with other thermal and microwave techniques. *Earth and Planetary Science Letters*, **257**, 564–581.

Biggin, A. J., Strik, G., and Langereis, C. G., 2009. The intensity of the geomagnetic field in the late-Archaean: new measurements and an analysis of the updated IAGA palaeointensity database. *Earth Planets and Space*, **61**, 9–22.

Biggin, A. J., and Thomas, D. N., 2003. Analysis of long-term variations in the geomagnetic poloidal field intensity and evaluation of their relationship with global geodynamics. *Geophysical Journal International*, **152**(2), 392–415.

Cisowski, S., and Fuller, M., 1986. Lunar paleointensities via theIRM(s) normalization method and the early magnetic history of the moon. In Hartmann, W., Phillips, R., and Taylor, G. (eds.), *The Origin of the Moon*. Houston: Lunar and Planetary Science Institute, pp. 411–421.

Cisowski, S. M., 1982. A review of lunar paleointensity data. *Advances in Space Research*, **2**(12), 31–34.

Coe, R. S., 1967. Palaeointensities of the Earth's magnetic field determined from Tertiary and Quaternary rocks. *Journal of Geophysical Research-Solid Earth*, **72**, 3247–3262.

Constable, C. G., Tauxe, L., and Parker, R. L., 1998. Analysis of 11 Myr of geomagnetic intensity variation. *Journal of Geophysical Research B: Solid Earth*, **103**(8), 17735–17748.

Dekkers, M. J., and Böhnel, H. N., 2006. Reliable absolute palaeointensities independent of magnetic domain state. *Earth and Planetary Science Letters*, **248**(1–2), 508–517.

Driscoll, P., and Olson, P., 2009. Polarity reversals in geodynamo models with core evolution. *Earth and Planetary Science Letters*, **282**(1–4), 24–33.

Fabian, K., 2009. Thermochemical remanence acquisition in single-domain particle ensembles: A case for possible overestimation of the geomagnetic paleointensity. *Geochemistry Geophysics Geosystems*, **10**(6), doi: 10.1029/2009GC002420.

Fuller, M., and Weeks, R., 1992. Superplumes and superchrons. *Nature*, **356**(6364), 16–17.

Genevey, A., Gallet, Y., Constable, C. G., Korte, M., and Hulot, G., 2008. ArcheoInt: an upgraded compilation of geomagnetic field intensity data for the past ten millennia and its application to the recovery of the past dipole moment. *Geochemistry Geophysics Geosystems*, **9**, Q04038, doi:10.1029/2007GC001881.

Genevey, A., Gallet, Y., Rosen, J., and Le Goff, M., 2009. Evidence for rapid geomagnetic field intensity variations in Western Europe over the past 800Ã' years from new French archeointensity data. *Earth and Planetary Science Letters*, **284** (1–2), 132–143.

Glatzmaier, G. A., Coe, R. S., Hongre, L., and Roberts, P. H., 1999. The role of the Earth's mantle in controlling the frequency of geomagnetic reversals. *Nature*, **401**(6756), 885–890.

Goguitchaichvili, A., Urrutia-Fucugauchi, J., and Alva-Valdivia, L. M., 2002. Mesozoic dipole low: myth or reality? *Eos, Transactions American Geophysical Union*, **83**(41), 457.

Granot, R., Tauxe, L., Gee, J. S., and Ron, H., 2007. A view into the Cretaceous geomagnetic field from analysis of gabbros and submarine glasses. *Earth and Planetary Science Letters*, **256**(1–2), 1–11.

Gubbins, D., Alfè, D., Masters, G., Price, G. D., and Gillan, M., 2004. Gross thermodynamics of two-component core convection. *Geophysical Journal International*, **157**(3), 1407–1414.

Gubbins, D., Jones, A. L., and Finlay, C. C., 2006. Fall in earth's magnetic field is erratic. *Science*, **312**(5775), 900–902.

Hide, R., 2000. Generic nonlinear processes in self-exciting dynamos and the long-term behaviour of the main geomagnetic field, including polarity superchrons. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **358**(1768), 943–955.

Hill, M. J., Pan, Y. X., and Davies, C. J., 2008. An assessment of the reliability of palaeointensity results obtained from the Cretaceous aged Suhongtu section, Inner Mongolia, China. *Physics of the Earth and Planetary Interiors*, **169**(1–4), 76–88.

Hill, M. J., and Shaw, J., 2000. Magnetic field intensity study of the 1960 Kilauea lava flow, Hawaii, using the microwave palaeointensity technique. *Geophysical Journal International*, **142**(2), 487–504.

Hoffman, K. A., and Biggin, A. J., 2005. A rapid multiple-sample approach to the determination of absolute paleointensity. *Journal of Geophysical Research-Solid Earth*, **110**(B12), doi: 10.1029/2005JB003646.

Hoffman, K. A., Constantine, V. L., and Morse, D. L., 1989. Determination of Absolute Palaeointensity Using a Multi-Specimen Procedure. *Nature*, **339**(6222), 295–297.

Johnson, C., and Constable, C. G., 1997. The time-averaged geomagnetic field: global and regional biases for 0–5 Ma. *Geophysical Journal International*, **131**, 643–666.

Johnson, C., and Mcfadden, P. L., 2007. Time-averaged field and paleosecular variation. In Kono, M. (ed.), *Treatise on Geophysics*. New York: Elsevier. Chap. 11.

Kono, M., 1978. Reliability of Paleointensity methods using alternating-field demagnetization and anhysteretic remanence. *Geophysical Journal of the Royal Astronomical Society*, **54**(2), 241–261.

Kono, M., and Ueno, N., 1977. Palaeointensity determination by a modified Thellier method. *Physics of the Earth and Planetary Interiors*, **13**, 305–314.

Korhonen, K., Donadini, F., Riisager, P., and Pesonen, L. J., 2008. GEOMAGIA50: an archeointensity database with PHP and MySQL. *Geochemistry Geophysics Geosystems*, **9**, Q04029, doi:10.1029/2007GC001893.

Korte, M., and Constable, C. G., 2005. Continuous geomagnetic field models for the past 7 millennia: 2. CALS7K.

Geochemistry, Geophysics, Geosystems, **6**, Q02H16, doi:10.1029/2004GC000801.

Krasa, D., Heunemann, C., Leonhardt, R., and Petersen, N., 2003. Experimental procedure to detect multidomain remanence during Thellier-Thellier experiments. *Physics and Chemistry of the Earth*, **28**(16–19), 681–687.

Labrosse, S., Poirier, J. P., and Le Mouel, J. L., 2001. The age of the inner core. *Earth and Planetary Science Letters*, **190**(3–4), 111–123.

Lawrence, K., Johnson, C., Tauxe, L., and Gee, J., 2008. Lunar paleointensity measurements: implications for lunar magnetic evolution. *Physics of the Earth and Planetary Interiors*, **168** (1–2), 71–87.

Le Goff, M., and Gallet, Y., 2004. A new three-axis vibrating sample magnetometer for continuous high-temperature magnetization measurements: applications to paleo- and archeo-intensity determinations. *Earth and Planetary Science Letters*, **229** (1–2), 31–43.

Macouin, M., Valet, J. P., and Besse, J., 2004. Long-term evolution of the geomagnetic dipole moment. *Physics of the Earth and Planetary Interiors*, **147**(2–3), 239–246.

Michalk, D. M., et al., 2010. Application of the multispecimen palaeointensity method to Pleistocene lava flows from the Trans-Mexican Volcanic Belt. *Physics of the Earth and Planetary Interiors*, **179**, 139–156.

Muxworthy, A. R., 2010. Revisiting a domain-state independent method of palaeointensity determination. *Physics of the Earth and Planetary Interiors*, **179**, 21–31.

Néel, L., 1955. Some theoretical aspects of rock magnetism. *Advances in Physics*, **4**, 191–242.

Perrin, M., 1998. Paleointensity determination, magnetic domain structure, and selection criteria. *Journal of Geophysical Research-Solid Earth*, **103**(B12), 30591–30600.

Perrin, M., and Shcherbakov, V., 1997. Paleointensity of the earth's magnetic field for the past 400 Ma: Evidence for a dipole structure during the Mesozoic Low. *Journal of Geomagnetism and Geoelectricity*, **49**(4), 601–614.

Pick, T., and Tauxe, L., 1993. Geomagnetic palaeointensities during the Cretaceous normal Superchron measured using submarine basaltic glass. *Nature*, **366**(6452), 238–242.

Prévot, M., Derder, M. E., Mcwilliams, M., and Thompson, J., 1990. Intensity of the earth's magnetic-field - evidence for a Mesozoic dipole low. *Earth and Planetary Science Letters*, **97**(1–2), 129–139.

Prévot, M., Lecaille, A., and Mankinen, E. A., 1981. Magnetic effects of maghematization of oceanic crust. *Journal of Geophysical Research*, **86**, 4009–4020.

Riisager, J., Perrin, M., Riisager, P., and Vandamme, D., 2001. Palaeomagnetic results and palaeointensity of Late Cretaceous Madagascan basalt. *Journal of African Earth Sciences*, **32**(3), 503–518.

Rolph, T. C., and Shaw, J., 1985. A new method of paleofield magnitude correction for thermally altered samples and its application to Lower Carboniferous lavas. *Geophysical Journal of the Royal Astronomical Society*, **80**, 773–781.

Selkin, P. A., Gee, J. S., and Tauxe, L., 2007. Nonlinear thermoremanence acquisition and implications for paleointensity data. *Earth and Planetary Science Letters*, **256**(1–2), 81–89.

Selkin, P. A., Gee, J. S., Tauxe, L., Meurer, W. P., and Newell, A. J., 2000. The effect of remanence anisotropy on paleointensity estimates: a case study from the Archean Stillwater Complex. *Earth and Planetary Science Letters*, **183**(3–4), 403–416.

Selkin, P. A., and Tauxe, L., 2000. Long-term variations in palaeointensity. *Philosophical Transactions of the Royal Society of London Series a-Mathematical Physical and Engineering Sciences*, **358**(1768), 1065–1088.

Shaw, J., 1974. A new method of determining the magnitude of the palaeomagnetic field, application to five historic lavas and five archaeological samples. *Geophysical Journal of the Royal Astronomical Society*, **39**, 133–141.

Shaw, J., Hill, M. J., and Openshaw, S. J., 2001. Investigating the ancient Martian magnetic field using microwaves. *Earth and Planetary Science Letters*, **190**(3–4), 103–109.

Shaw, J., Walton, D., Yang, S., Rolph, T. C., and Share, J. A., 1996. Microwave archaeointensities from Peruvian ceramics. *Geophysical Journal International*, **124**(1), 241–244.

Snowball, I., and Sandgren, P., 2002. Geomagnetic field variations in northern Sweden during the Holocene quantified from varved lake sediments and their implications for cosmogenic nuclide production rates. *Holocene*, **12**(5), 517–530.

Solanki, S. K., Usoskin, I. G., Kromer, B., Schüssler, M., and Beer, J., 2004. Unusual activity of the Sun during recent decades compared to the previous 11, 000 years. *Nature*, **431**(7012), 1084–1087.

Stern, D. P., 2002. A millennium of geomagnetism. *Reviews of Geophysics*, **40**(3), 1007.

Stevenson, D. J., Spohn, T., and Schubert, G., 1983. Magnetism and thermal evolution of the terrestrial planets. *Icarus*, **54**(3), 466–489.

Tanaka, H., and Komuro, N., 2009. The Shaw paleointensity method: Can the ARM simulate the TRM alteration? *Physics of the Earth and Planetary Interiors*, **173**(3–4), 269–278.

Tarduno, J. A., and Cottrell, R. D., 2005. Dipole strength and variation of the time-averaged reversing and nonreversing geodynamo based on Thellier analyses of single plagioclase crystals. *Journal of Geophysical Research B: Solid Earth*, **110**(11), 1–10.

Tarduno, J. A., Cottrell, R. D., and Smirnov, A. V., 2001. High geomagnetic intensity during the Mid-Cretaceous from Thellier analyses of single plagioclase crystals. *Science*, **293**(5530), 607–607.

Tarduno, J. A., Cottrell, R. D., and Smirnov, A. V., 2006. The paleomagnetism of single silicate crystals: recording geomagnetic field strength during mixed polarity intervals, superchrons, and inner core growth. *Reviews of Geophysics*, **44**(1), RG1002, doi:10.1029/2005RG000189.

Tarduno, J. A., et al., 2010. Geodynamo, solar wind, and magnetopause 3.4 to 3.45 billion years ago. *Science*, **327**, 1238–1240.

Tarduno, J. A., Cottrell, R. D., Watkeys, M. K., and Bauch, D., 2007. Geomagnetic field strength 3.2 billion years ago recorded by single silicate crystals. *Nature*, **446**, 657–660.

Tauxe, L., 2006. Long-term trends in paleointensity: the contribution of DSDP/ODP submarine basaltic glass collections. *Physics of the Earth and Planetary Interiors*, **156**(3–4), 223–241.

Tauxe, L. and Staudigel, H., 2004. Strength of the geomagnetic field in the Cretaceous Normal Superchron: New data from submarine basaltic glass of the Troodos Ophiolite. *Geochemistry Geophysics Geosystems*, **5**, Q02H06.

Tauxe, L., and Yamazaki, T., 2007. Paleointensities. In Kono, M. (ed.), *Geomagnetism. Treatise on Geophysics*. Amsterdam: Elsevier, pp. 510–563.

Thellier, E., and Thellier, O., 1959. Sur l'intensité du champ magnétique terrestre dans la passé historique et géologique. *Ann. Géophys.*, **15**, 285–376.

Tsunakawa, H., and Shaw, J., 1994. The Shaw method of paleointensity determinations and its application to recent volcanic-rocks. *Geophysical Journal International*, **118**(3), 781–787.

Valet, J. P., 2003. Time variations in geomagnetic intensity. *Reviews of Geophysics*, **41**(1), 1004, doi:10.1029/2001RG000104.

Valet, J. P., et al., 1996. Absolute paleointensity and magnetomineralogical changes. *Journal of Geophysical Research-Solid Earth*, **101**(B11), 25029–25044.

Valet, J. P., Meynadier, L., and Guyodo, Y., 2005. Geomagnetic dipole strength and reversal rate over the past two million years. *Nature*, **435**(7043), 802–805.
Weiss, B. P., et al., 2008a. Magnetism on the angrite parent body and the early differentiation of planetesimals. *Science*, **322**(5902), 713–716.
Weiss, B. P., Fong, L. E., Vali, H., Lima, E. A., and Baudenbacher, F. J., 2008b. Paleointensity of the ancient Martian magnetic field. *Geophysical Research Letters*, **35**(23), L23207, doi:10.1029/2008GL035585.
Wilson, R. L., 1961. Palaeomagnetism in Northern Ireland .1. The thermal demagnetization of natural magnetic moments in rocks. *Geophysical Journal of the Royal Astronomical Society*, **5**(1), 45–58.
Yamamoto, Y., Tsunakawa, H., and Shibuya, H., 2003. Palaeointensity study of the Hawaiian 1960 lava: implications for possible causes of erroneously high intensities. *Geophysical Journal International*, **153**(1), 263–276.
Yamamoto, Y., and Shaw, J., 2008. Development of the microwave LTD-DHT Shaw method fpr absolute paleointensity determination. *Physics of the Earth and Planetary Interiors*, **170**, 15–23.

Cross-references

Archaeomagnetism
Curie Temperature
Geomagnetic Excursions
Geomagnetic Field, Polarity Reversals
Geomagnetic Field, Secular Variation
Magnetic Domains
Paleomagnetism, Measurement Techniques and Instrumentation
Remanent Magnetism

PALEOMAGNETISM, MAGNETOSTRATIGRAPHY

Donald R. Prothero
Department of Geology, Occidental College,
Los Angeles, CA, USA

Synonyms

Magnetic polarity stratigraphy; Magnetic stratigraphy

Definition

Magnetic stratigraphy. Application of magnetic methods to obtain the magnetic polarity pattern in stratified rocks, allowing them to be correlated to the global magnetic timescale.
Remanent magnetism. Permanent magnetism in rocks, caused by the orientation of the Earth's magnetic field at the time the rock was formed.

Magnetic stratigraphy

Introduction

Since the 1960s, one of the most important tools for dating and correlating fossiliferous sedimentary rocks has been magnetic stratigraphy. First developed to aid in the correlation of deep-sea sediment cores from the Deep Sea Drilling Project, by the 1970s it had also been applied to dating terrestrial sedimentary rocks as well. The early efforts were hampered by the slowness and insensitivity of old spinner magnetometers, but with the widespread availability and the increased reliability, sensitivity, speed, and precision of cryogenic magnetometers in the 1970s, the field has grown tremendously. It has now been applied to many of the important marine and terrestrial sequences that were the type sections for biostratigraphic zones and timescales, and is routinely used whenever there are suitable outcrops and age control to allow global correlation.

Magnetic stratigraphy works by obtaining a dense sampling of oriented samples from stratigraphic sections of layered sedimentary rocks or volcanic rocks. These samples record the direction of the magnetic field (*normal*, as in today's field, or *reversed*, about 180° from today's field) in a random pattern of polarity zones. If there is a sufficiently long pattern of polarity zones, it can be matched up to the global magnetic polarity timescale by calibrating it with radiometric ages, or by means of biostratigraphy, or both. Once sections have been correlated, the magnetic polarity reversals represent time planes, and their ages can be resolved to less than 10,000 years, much better dating and resolution than obtainable by any other means.

Detrital remanent magnetization

Although it is possible to correlate lava flows by magnetic methods, by far the most common application of magnetic stratigraphy is to layered sedimentary rocks. These rocks become magnetized by *detrital remanent magnetization* (DRM). When the sediment is loose and fluid, it will contain a small number of tiny (less than 60 μm in diameter) grains of magnetic minerals, such as magnetite or hematite. While these sediments settle out and become liquefied, the magnetic mineral grains are influenced by the ambient magnetic field at the time the sediment was deposited, and thus line up to be parallel to it. Thus, the magnetic grains act like tiny bar magnets, or like iron filings near a bar magnet. As the sediment becomes compacted or cemented and turned into sedimentary rock, the orientation of these magnetic grains is preserved. Although there are instances of sediment compaction causing the grains to rotate and thus flattening the inclination angle, most sedimentary rocks seem to record field directions that closely match those produced from the cooling of igneous rocks, which are not subject to compaction (Butler, 1992, pp. 183–187).

Field and laboratory techniques

In order to undertake a magnetostratigraphic study, a large number of oriented samples densely spanning the total stratigraphic interval of exposed rock are required. The density of sampling is dictated by how many exposures are suitable, and the logistics of how many people and man-hours of field work are available to conduct the sampling, since it is a tedious and exhausting task. Typically, each stratigraphic level or

"site" is sampled at least in three places, and more if possible, so that Fisher statistics can be calculated. In deep-sea cores, sampling is conducted on the split core, with small subsamples taken from the core so that an excessive amount of material is not removed.

The lithology of rock is also critical. Most remanence is held in tiny (less than a few tens of microns) *single-domain grains* of magnetic mineral, or in slightly larger *pseudo-single domain* grains. Above a certain size, it is possible to have two or more magnetic domains in a grain, and then they will tend to form antipodally oriented domains within a single grain and cancel each other out, giving no net remanence. Thus, a fine-grained component is essential for a sedimentary rock to retain useful remanence. For this reason, sampling usually concentrates on shales, mudstones, and fine-grained sandstones. Coarser sandstones may work if they have a fine-grained matrix. A well-sorted, very clean coarse sandstone with no matrix should be avoided for this reason.

If the rocks are very hard and well cemented, it is possible to sample using a coring drill, as is commonly practiced in igneous and metamorphic rocks. Most sedimentary rocks, however, are too friable and crumbly to stand up to the stress of drilling and the flow of water through the drill bit (required to cool the bit and flush out the drilled rock chips). In this case, samples of terrestrial rocks are taken by scraping oriented surfaces on hand samples, and then subsampling them into cubes or cores in the laboratory using a coring drill or band saw, or hammer and sandpaper for very soft rocks. If the samples are very friable (such as loose soils or unconsolidated sediments), it is still possible to obtain oriented samples. Typically, this is done by pressing open tubes of a high-temperature glass (such as quartz glass or sometimes Pyrex) into the exposure, measuring the orientation, then removing it and pouring a high-temperature-tolerant hardener (such as sodium silicate) into the sediment. The tube is then taped shut, and in the lab it can be sealed at the open end by a high-temperature ceramic, such as Zircar aluminum cement. The sealed glass tube is then measured as if it were a normal drill core. In addition, samples may crumble during preparation until they are too small for a core or cube. They can be salvaged by placing them oriented face down on a piece of wax paper, surrounding them with a plastic ring the diameter of a core (such as a cut piece of plastic test tube), and pouring in Zircar aluminum ceramic to make them into a molded ring that will survive the high temperatures of the thermal demagnetization.

Once samples are fully prepared, they can be measured in a magnetometer to determine the direction and intensity of their magnetic remanence. Before the late 1970s, all the studies were done with older-generation *spinner magnetometers*, which required over 30 min to measure a single sample, and were not sensitive enough to measure most sedimentary rocks. For this reason, only a few measurements could be made on each sample, and only a limited number of samples could be taken if the project were to remain reasonable in size.

In the 1970s, the invention and widespread adoption of *cryogenic magnetometers* revolutionized the study of paleomagnetism. These devices have a sealed jacket of liquid helium at 4 K surrounding a superconducting region in the bottom of the magnetometer. The electronics in the cryogenic chamber are thus sensitive to extremely small magnetic moments, and can make measurements in a few seconds. They are four to five orders of magnitude more sensitive than spinner magnetometers, so samples can be measured that are extremely weakly magnetized. These include not only every kind of rock but even organic tissues like the magnetite in bacteria or birds or bees or butterflies. Early cryogenic magnetometers were very temperamental and prone to problems. They bled off the evaporating helium at such a fast rate that a single shipment costing hundreds of dollars would only last a few weeks. Only a few measurements could be taken on each sample, and the lab had to work 24/7 to use every precious second of magnetometer time. Now most cryogenic magnetometers have a cryocooler pump that circulates liquid helium or nitrogen around the cryogenic container, and keeps it cold for many months at a time, greatly reducing the cost of operation, wear and tear on the magnetometer, and improving reliability. In many labs (especially those equipped with a Caltech-style automatic sample changer), it is possible to make dozens of measurements in an hour, allowing paleomagnetists to make dozens of different measurements on each sample.

Each sample contains a mixture of remanences acquired during the long history of the rock unit. There will be a primary, or characteristic, component from when the rock originally formed, which may be overprinted by one or more magnetizations that were acquired at later times. For a reliable paleomagnetic study, the researcher must attempt to remove these overprinted components by gradually demagnetizing the sample, and gradually stripping off the younger, less stable magnetizations until only the characteristic remanence is left. Consequently, it is typical for a modern paleomagnetic study to measure the samples first at NRM (*natural remanent magnetization*, or the magnetization of the sample before treatment). The samples can then be demagnetized using *alternating fields* (AF), which helps remove remanence due to low-coercivity minerals such as magnetite, and eliminates magnetization from transient fields (such as lightning strikes) or from VRM (*viscous remanent magnetization*). Examination of the magnetic behavior and the drop in intensity at each AF step allows the paleomagnetist to determine how much of the remanence is held in a low-coercivity mineral, or whether significant high-coercivity minerals (which do not respond to AF demagnetization) are present. AF treatment is usually sufficient for samples from deep-sea cores, which have not undergone any significant diagenesis or alteration. Another method, known as *thermal demagnetization*, heats the samples in field-free space to a preset temperature, then cools them down to room temperature without acquiring any new magnetization from the earth's field. In some cases, the

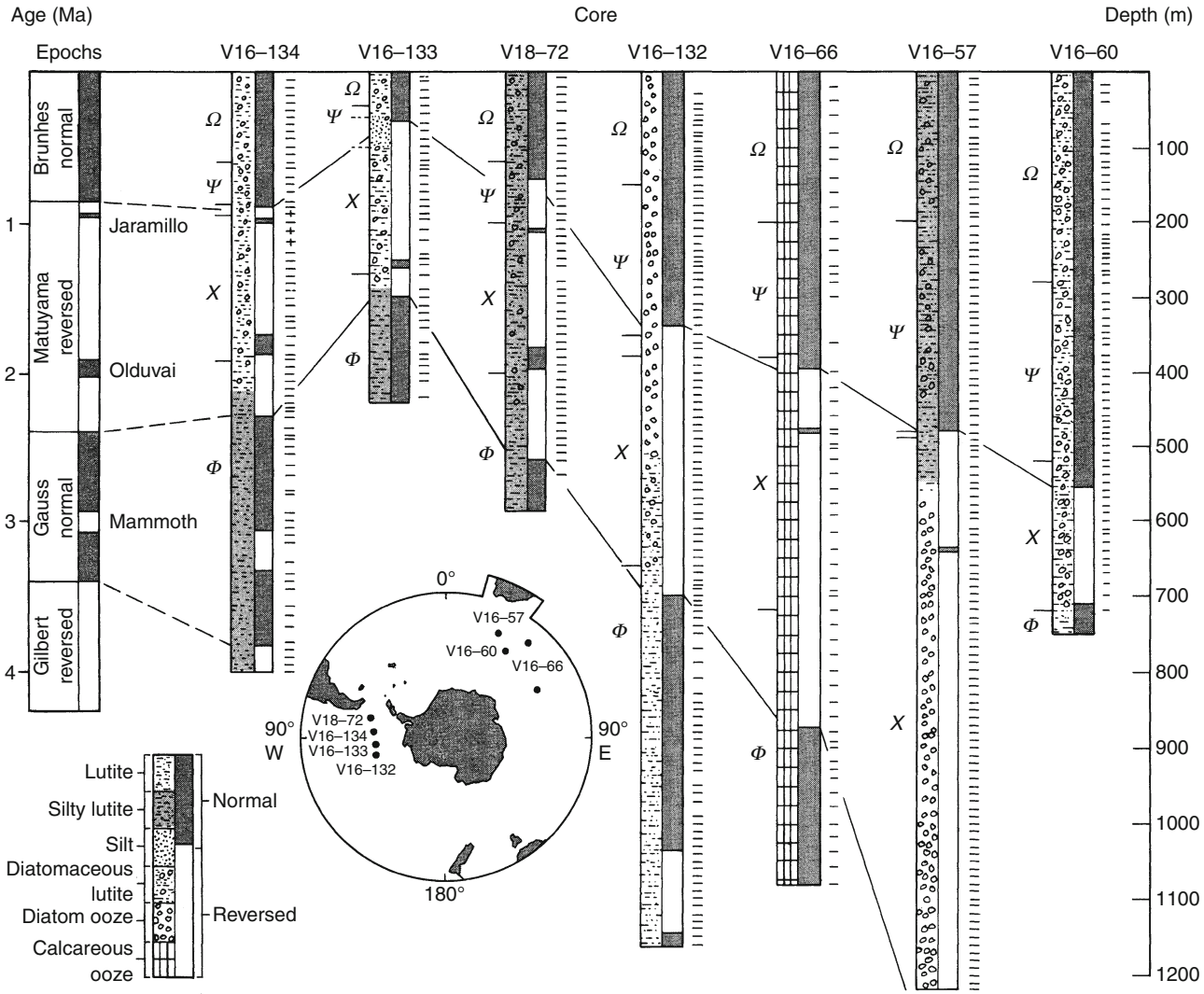
oven is saturated by a jet of nitrogen gas to reduce the possibility of minerals oxidizing as they cool down. Thermal demagnetization in many steps of increasing temperature allows the paleomagnetist to see how the remanence changes as higher and higher temperatures are reached and the *blocking temperature* of the magnetic minerals (578°C for pure magnetite; 630° for hematite) is approached. At high temperatures, the behavior of the sample above the blocking temperature of magnetite tells the researcher about the intensity and direction of the magnetic components that are held in hematite. Thermal demagnetization has an additional advantage. Above 200°C, iron oxides and hydroxides such as goethite are converted to hematite, and their remanence is removed. These "rusty" minerals are often formed by weathering of sedimentary rocks at the surface, and usual form as a *chemical remanence* deposited in the rock long after the primary or characteristic remanence was formed. Thus, they are a form of overprinting that must be removed, but cannot be affected by AF demagnetization. For this reason, nearly all terrestrial sedimentary rocks (even if they are not visibly reddish) should be thermally demagnetized to insure that there are no false overprinted magnetic components. Many studies use both AF and thermal demagnetization on every sample to strip off overprints due to iron oxides, to determine the relative contribution of high-coercivity and low-coercivity minerals, and to obtain a clear sense of the dominant magnetic mineralogy of each sample. The samples can also be studied in polished thin sections, where the light reflects differently off magnetite, goethite, and hematite, and the depositional fabric of the magnetic mineral and size of the grains can be determined.

These magnetic directions can be plotted and analyzed in a number of different ways. For simple visual plotting of the three-dimensional changes in directions, a conventional *stereonet* can be used. But a stereonet only shows direction, where as an *orthogonal demagnetization plot* ("Zijderveld plot") shows both the direction and intensity of each demagnetization step (Zijderveld, 1967). Most paleomagnetics labs have software that rapidly creates these plots, so the behavior of the sample can be studied as the demagnetization proceeds. Once a stable remanence has been obtained and overprints isolated and removed, the inferred characteristic remanence components are usually summarized by the *least squares method* of Kirschvink (1980). These directions for each sample can then be averaged using Fisher (1953) statistics, which give the mean and ellipse of 95% confidence for each sample. Using Watson's (1956) criteria, the statistical significance of each site can be assessed. Frequently they are ranked using the criteria of Opdyke et al. (1977), where Class I sites are statistically significant at the 95% confidence level, that is, there is a 95% chance that the sample distribution differs from a random distribution. These sites were labeled "Class A" by Johnson et al. (1985). Class II sites have fewer than three samples (so statistics cannot be calculated), and Class III sites ("Class B" sites of Johnson et al., 1985) have at least two sample directions that are in agreement, but one which is divergent due to incomplete removal of overprints. These directions can then be plotted against a stratigraphic column to show the overall polarity pattern for the section, and also indicate the degree of reliability of individual samples at each site.

As in any other paleomagnetic study, there are a number of techniques for assessing the stability and reliability of the sample directions (see Butler, 1992, pp. 123–128). The most commonly used in magnetostratigraphy is the *reversal test*, which compares the statistical mean of all normal directions and reversed directions to see if they are antipodal (as they should be if they represent the original magnetic field of the earth). Most long stratigraphic sections yield both normal and reversed polarity samples, providing the data for a stability test like this. In some cases, there are conglomerates in the section, so a *conglomerate test* can be performed. In this test, samples are taken from a large sample of clasts in a conglomerate. If their directions are randomly scattered, then they have not been remagnetized since the conglomerate formed. If, on the other hand, they yield directions that are clustered, then there has been overprinting not only of that conglomerate, but also of the rocks in the rest of the section. Another common test in rocks that have been deformed is the *fold test*. The researcher can compare the directions of the samples with and without dip correction. If these directions are clustered with dip correction, and more scattered without dip correction, then the directions were acquired before folding. If the directions are more clustered in their uncorrected, folded orientation, then the directions were acquired after folding (and thus are not primary or characteristic directions). Not all these stability tests are available in every study, but at least one (usually the reversal test) should be attempted to obtain credible results. If other tests are also available, they should also be analyzed to increase the rigor of the study.

Calibration and correlation

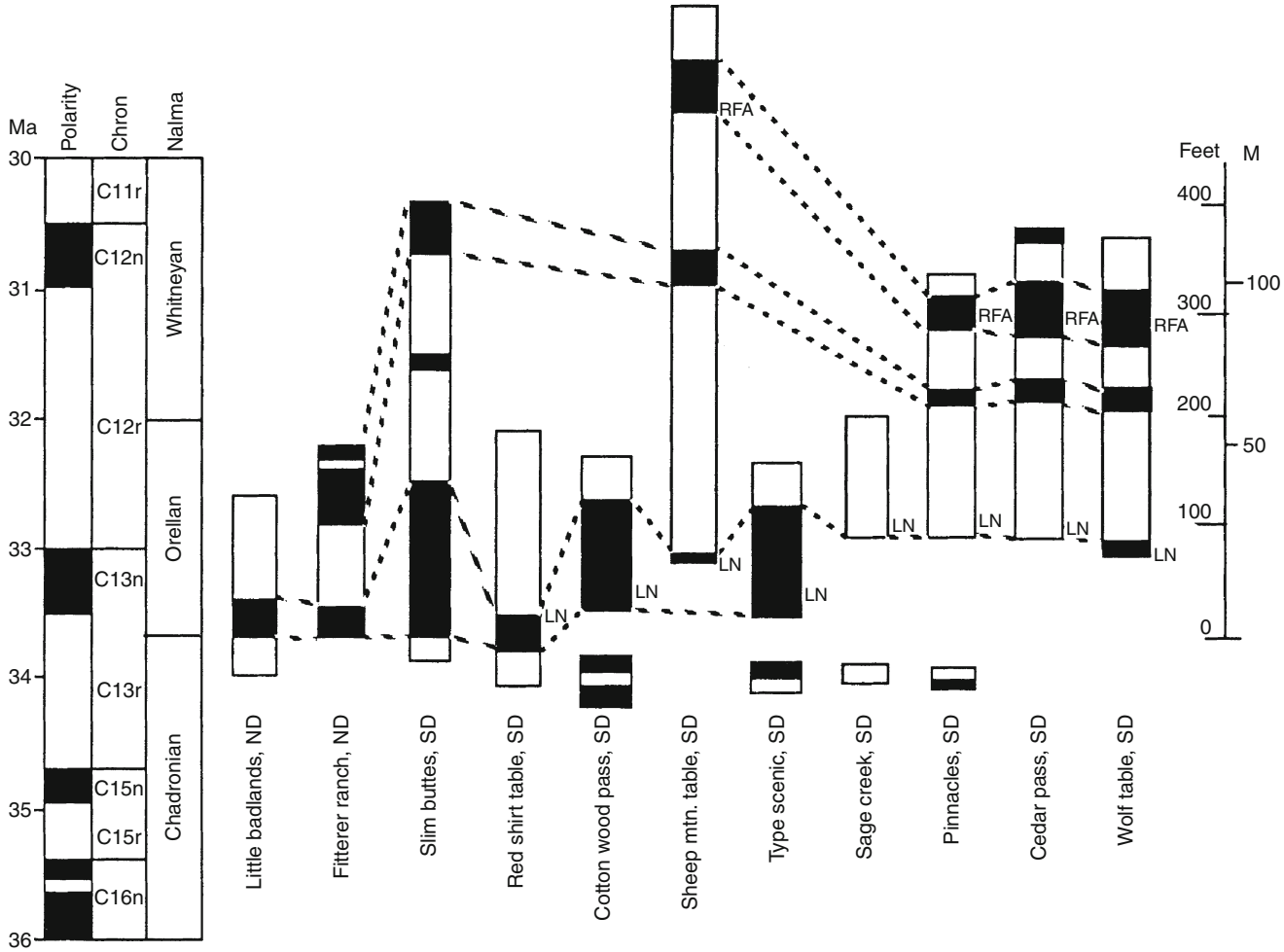
Once characteristic directions that have passed stability tests have been obtained from each site and they are plotted, then the overall polarity pattern for each individual section can be assessed. Polarity zones are not diagnostic or unambiguous by themselves, but must be correlated based on other criteria, such as lithostratigraphy or biostratigraphy. In the early days of magnetic stratigraphy, deep-sea cores (Figure 1) could be sampled down from the modern sediments at the top of the core, and each polarity zone could be calibrated by the Brunhes normal polarity magnetozone at the top. Most deep-sea cores, however, do not have a continuous sedimentary record from the present backward, but represent an older geologic span of time separated by unconformities. In such cases, the biostratigraphy of the planktonic microfossils enables the scientist to calibrate a core directly against the magnetic polarity timescale.



Paleomagnetism, Magnetostratigraphy, Figure 1 Magnetostratigraphic correlation of Antarctic cores, based on the modern Brunhes normal polarity sediments at the top of each section, and the biostratigraphic zones of radiolaria (*shown by the capital Greek letters*) in each core. To the right of each column, minus signs indicate normally magnetized specimens, while plus signs are reversed specimens (After Opdyke et al., 1966, Figure 1; by permission of the American Association of the Advancement of Science).

In another example, Figure 2 shows the correlation of magnetozones of many different stratigraphic sections across the Eocene-Oligocene White River Group from northeastern Colorado to Nebraska and Wyoming to the Big Badlands of South Dakota to North Dakota. The correlations are based on the lithostratigraphic units and the biostratigraphic zonation, with the common datum based on a widespread ash horizon found throughout most of the sections. Note that the polarity magnetozones are roughly consistent in thickness from one section to another, although each of the magnetozones is proportionally thicker where the total sediment accumulation rate is higher (e.g., Slim Buttes, South Dakota).

But these polarity matches do not establish correlation with the global magnetic polarity timescale. Magnetic polarity can only be "normal" or "reversed," and that pattern has occurred thousands of times in the geologic past, so it is not sufficient just to get a raw polarity pattern and attempt to correlate it to the global timescale (unless you are working down from modern sediments at the top of a section, in which case the uppermost normal magnetozone is the Brunhes Chron, or Chron 1n). For reliable correlation, the section needs to be calibrated by some sort of independent means of time control. Ideally, this is provided by at least one and preferably more radiometric dates, which will place the local section on the global



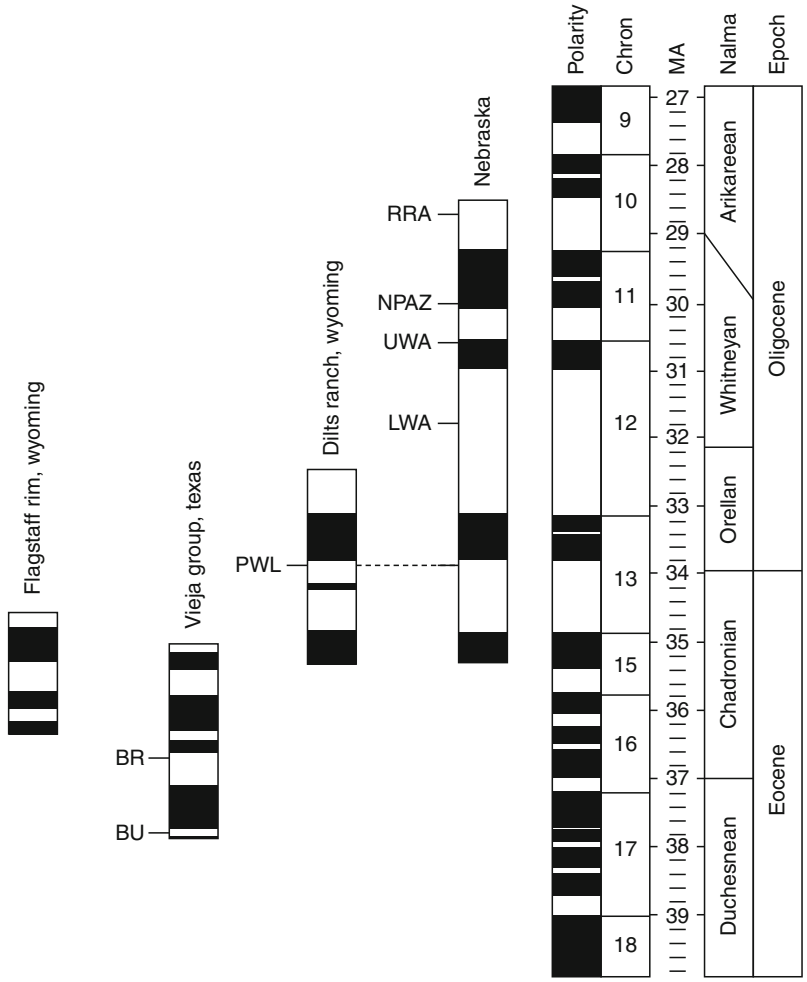
Paleomagnetism, Magnetostratigraphy, Figure 2 Correlation of magnetostratigraphic sections in the White River Group from North and South Dakota. All of the sections are correlated based on the datum of the base of the Orellan. Abbreviations as follows: NALMA = North American land mammal ages; LN = Lower Nodular Zone of the Big Badlands; RFA = Rockyford Ash. From Prothero (1996, Figure 9).

magnetic polarity timescale (Figure 3). Where radiometric dates are not available, it is often possible to use a finely resolved well-calibrated biostratigraphy (whose precise correlation to the global timescale is already known) to correlate another section containing those key fossils.

Strengths and weaknesses of magnetic stratigraphy

Like any other stratigraphic technique (see Prothero and Schwab, 2003), magnetic stratigraphy has great strengths – and also some weaknesses. The strengths include:

1. *Worldwide*: magnetic polarity events happen around the globe simultaneously. By contrast, most other stratigraphic markers (volcanic ashes, biostratigraphic events, lithologic boundaries, even isotopic events) occur only locally.
2. *Independent of facies or lithology*: any fine-grained sedimentary rock or volcanic rock will work, allowing us to correlate between deep-sea muds to shallow-marine silts to terrestrial floodplain mudstones. By contrast, rock units are typically time-transgressive as facies migrate, and are not found on both land and sea. Stable isotope curves are usually restricted to marine rocks.
3. *Geological instantaneous*: polarity reversals take place over 4,000–5,000 years, which is a geological instant for most studies of the deep past. By contrast, most lithostratigraphic units are time-transgressive over millions of years.
4. *Constant resolution regardless of age*: the duration of magnetic polarity transitions is constant whether the rocks are Pleistocene or Permian, so the resolution is always in the order of a few thousand years.



Paleomagnetism, Magnetostratigraphy, Figure 3 Calibration of the White River-Arikaree Group magnetic stratigraphy of Figure 2, based on $^{40}Ar/^{39}Ar$ dating of key volcanic ashes. Abbreviations: BR = Bracks Rhyolite; BU = Buckshot Ignimbrite; LWA = Lower Whitney Ash; NPAZ = Nonpareil ash zone; RRA = Roundhouse Rock Ash; UWA = Upper Whitney Ash (Modified from Prothero and Swisher, 1992, Figure 2.3).

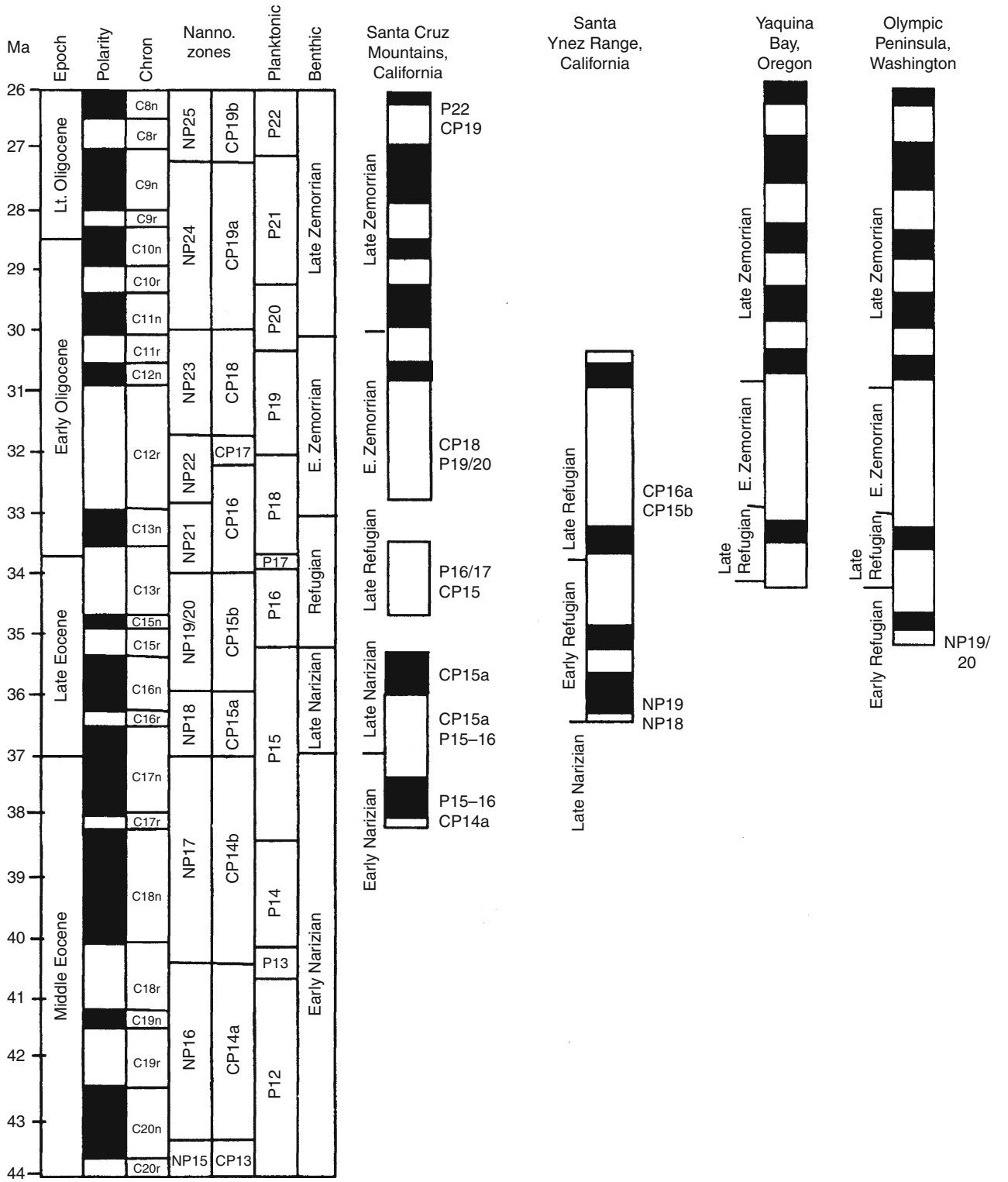
By contrast, radiometric dates have an error bar that is a proportion of the mean age and therefore gets larger as the rock gets older. For example, if the error estimates are ±5% for a rock of 1 Ma, then the true age is 1 million ± 50,000 years. If the age is 100 Ma, however, then the true age is 100 million ± 5 million years. In other words, there is a 95% probability that the true age lies between 95 and105 Ma. These error bars get larger with older and older rocks.

There are some limitations to magnetic stratigraphy as well:

1. Only *finer-grained lithologies*, or sandstones with clay matrix, are suitable (as discussed above).
2. If there are *unconformities* in the section, then magnetozones can be shortened or deleted altogether. This problem can best be addressed by biostratigraphic methods, the usual way in which we test for evidence of slow or no deposition.
3. *Demagnetization problems* can hamper the quality of the data, particularly in yielding spurious normal overprinted zones that have not yet been removed by thermal demagnetization of iron oxides and hydroxides (e.g., Butler and Lindsay, 1985).
4. *Independent form of age control* is required to calibrate the section.

Some applications and implications

Since the late 1960s, magnetic stratigraphy has become one of the most powerful tools available for precise dating of sedimentary sequences and geologic events, and for correlation of events between the terrestrial, shallow-marine, and deep-sea records. In most detailed



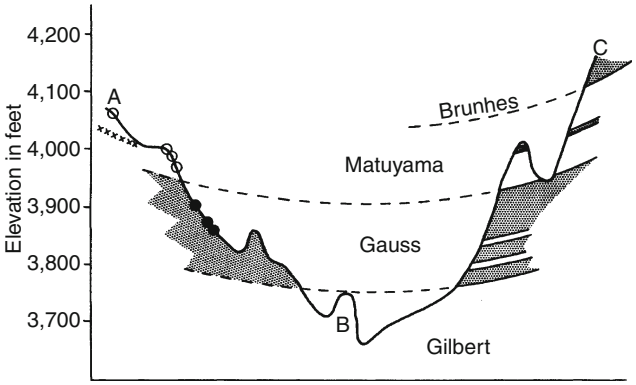
Paleomagnetism, Magnetostratigraphy, Figure 4 Magnetostratigraphic correlation of marine sections from the Pacific Coast, based on the magnetic polarity pattern and the biostratigraphy of microplankton. On the left is the standard geological time scale and polarity chrons with the planktonic zones of nannofossils and foraminifera (after Prothero, 2001).

stratigraphic studies, it is routine to do magnetostratigraphic analysis along with biostratigraphy to form a network of correlation now called *magnetobiostratigraphy* (Figure 4). Combined with stable isotope stratigraphy (especially in deep marine cores), it is possible to obtain very high-resolution, high-precision dating of many stratigraphic sections and correlate their events with high precision on a global basis. This, in turn, has allowed reliable and detailed comparison between climatic and biotic events on the land and the sea, and has revolutionized our understanding of climatic change and global correlations of climate and evolution.

The application of magnetic stratigraphy has also changed our conceptions of many geologic problems. For example, the dating of strata before the advent of magnetic stratigraphy has been radically rethought as these correlations improve. In North America, for example, the Uintan and Duchesnean land mammal ages of Wood et al. (1941) were once thought to be late Eocene, the Chadronian land mammal age early Oligocene, and the Orellan and Whitneyan land mammal ages middle and late Oligocene, respectively. However, with the advent of magnetobiostratigraphy and high-resolution argon/argon dating, all these correlations have changed radically (Prothero and Swisher, 1992). The Uintan and Duchesnean are now middle Eocene, the Chadronian late Eocene, and the Orellan and Whitneyan early Oligocene. The Arikareean, once thought to be early Miocene, is now largely late Oligocene. Every one of the original mid-Cenozoic stratigraphic interpretations of the Wood Committee (1941) was miscorrelated to the standard European timescale by at least one stage of the Eocene or Oligocene (and the same was true of their North American Miocene-Pliocene stages). Prior to magnetic stratigraphy and radiometric dating, the Deseadan land mammal age of South America was thought to be early Oligocene (27–37 Ma). Today it is considered late Oligocene to early Miocene (22–27 Ma) in age (MacFadden et al., 1985), or about 10 million years younger than originally thought.

Magnetic stratigraphy has also been used as a device for testing other geologic ideas. Because it produces geologically instantaneous time planes, it is valuable in testing whether biostratigraphic datums are synchronous or not. By and large, this has proven true, confirming the importance of biostratigraphy as a time indicator. Prothero (1982) showed this in the Eocene-Oligocene strata and fossils of the White River Group of the American High Plains. Lindsay et al. (1984) did the same for the biostratigraphy of the late Cenozoic of southwestern North America and Europe, as did Flynn et al. (1984) for the Eocene of the Rocky Mountains. A few biostratigraphic events, like *Lepus* event in North America (Lindsay et al., 1990), or the *Hipparion* horse datum in Europe, are demonstrably time-transgressive (Woodburne et al., 1981; Woodburne, 1989), but these are the exceptions to the rule.

In some cases, magnetic stratigraphy produces powerful inferences about the rock record. For example,



Paleomagnetism, Magnetostratigraphy,
Figure 5 Topographic profile across the San Pedro Valley of Arizona, showing the interpretation of "invisible magnetic zones" lying within the outcrops based on the sampling of Johnson et al. (1975). (Modified from Johnson et al., 1975, Figure 8).

Johnson et al. (1975) showed that magnetozones could be mapped along the Miocene-Pliocene exposures in the San Pedro Valley of Arizona. These "invisible magnetic time planes" (Figure 5) exist in the outcrops and can be mapped and interpreted using standard magnetostratigraphic methods. Magnetic stratigraphy allows precise determination of sedimentation rates in marine and terrestrial sections, and relative completeness of sections, which are important to many different kinds of stratigraphic studies (Sadler, 1981; May et al., 1985; Johnson et al., 1988).

Summary

Magnetic stratigraphy is now one of the principal tools that allow the precise, high-resolution correlation of geologic events in the stratigraphic record, especially climatic events. It is unique among stratigraphic methods in that global high-resolution correlations are possible between terrestrial and marine records, and does not suffer from issues of local facies changes or limited lateral extent of most other stratigraphic markers. Of course, the researcher must be aware of the limitations: relatively complete well-exposed sections of fine-grained rocks are required, as is an independent form of time control to allow calibration. Every sample must be carefully analyzed by multiple techniques to insure that the primary or characteristic remanence has been recovered and overprinting removed. Nonetheless, magnetic stratigraphy has improved our correlations of global events by several orders of magnitude, and corrected many erroneous correlations that were once widely accepted. It has allowed for many other interesting studies, such as analysis of rates of sedimentation, and tests of the isochroneity of biostratigraphic events. In short, the horizon is limitless for the application of paleomagnetism to stratified rocks.

Bibliography

Aïssaoui, D. M., McNeill, D. F., and Hurley, N. F., 1993. Applications of paleomagnetism to sedimentary geology. *SEPM. Special Publication*, **49**, 1–215.

Butler, R. F., 1992. *Paleomagnetism*. Boston: Blackwell, p. 320.

Butler, R. F., and Lindsay, E. H., 1985. Mineralogy of magnetic minerals and revised magnetic polarity stratigraphy of continental sediments, San Juan Basin, New Mexico. *Journal of Geology*, **93**, 535–554.

Fisher, R. A., 1953. Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A*, **217**, 295–305.

Flynn, J. J., MacFadden, B. J., and McKenna, M. C., 1984. Land mammal ages, faunal heterochrony, and temporal resolution in Cenozoic terrestrial sequences. *Journal of Geology*, **92**, 687–705.

Hailwood, E. A., 1989. *Magnetostratigraphy.* Blackwell: Oxford, pp. 1–84.

Johnson, N. M., Opdyke, N. D., and Lindsay, E. H., 1975. Magnetic polarity stratigraphy of Pliocene-Pleistocene terrestrial deposits and vertebrate faunas, San Pedro Valley, Arizona. *Geological Society of America Bulletin*, **86**, 5–12.

Johnson, N. M., Stix, J., Tauxe, L., Cerveny, P. F., and Tahirkheli, R. A. K., 1985. Paleomagnetic chronology, fluvial processes, and tectonic implications of the Siwalik deposits near Chinji Village, Pakistan. *Journal of Geology*, **93**, 27–40.

Johnson, N. M., Shiekh, K. A., Dawson-Sanders, E., and McRae, L., 1988. The use of magnetic reversal time lines in stratigraphic analysis: a case study in measuring variability in sedimentation rates. In Kleinspehn, K., and Paola, C. (eds.), *New Perspectives in Basin Analysis*. New York: Springer, pp. 307–330.

Kennett, J. P., (ed.), 1980. *Magnetic Stratigraphy of Sediments*. Stroudsburg: Dowden, Hutchinson & Ross, 438 pp.

Kirschvink, J. L., 1980. The least-squares line and plane and the analysis of paleomagnetic data: examples from Siberia and Morocco. *Geophysical Journal of the Royal Astronomical Society*, **62**, 699–718.

Lindsay, E. H., Opdyke, N. M., and Johnson, N. M., 1984. Blancan-Hemphillian land mammal ages and late Cenozoic dispersal events. *Annual Reviews of Earth and Planetary Sciences*, **12**, 445–488.

Lindsay, E. H., Smith, G. A., Haynes, G. V., and Opdyke, N. D., 1990. Sediments, geomorphology, magnetostratigraphy, and vertebrate paleontology in the San Pedro Valley, Arizona. *Journal of Geology*, **98**, 605–619.

MacFadden, B. J., Campbell, K. E., Cifelli, R. L., Siles, O., Johnson, N. M., Naeser, C. W., and Zeitler, P. K., 1985. Magnetic polarity stratigraphy and mammalian fauna of the Deseadan (late Oligocene-early Miocene) Salla beds in northern Bolivia. *Journal of Geology*, **93**, 223–250.

May, S. R., Butler, R. F., and Roth, F. A., 1985. Magnetic polarity stratigraphy and stratigraphic completeness. *Geophysical Research Letters*, **12**, 341–344.

Opdyke, N. D., and Channell, J. E. T., 1996. *Magnetic Stratigraphy.* San Diego: Academic, p. 341.

Opdyke, N. D., Glass, B., Hays, J. D., and Foster, J., 1966. Paleomagnetic study of Antarctic deep-sea cores. *Science*, **154**, 349–357.

Opdyke, N. D., Lindsay, E. H., Johnson, N. M., and Downs, T., 1977. The paleomagnetism and magnetic polarity stratigraphy of the mammal-bearing section of Anza-Borrego State Park, California. *Quaternary Research*, **7**, 316–329.

Prothero, D. R., 1982. How isochronous are mammalian biostratigraphic events? *Proceedings of the 3rd North American Paleontological Convention*, **2**, 405–409.

Prothero, D. R., 1988. Mammals and magnetostratigraphy. *Journal of Geological Education*, **34**(4), 227–236.

Prothero, D. R., 1996. Magnetostratigraphy of the White River Group in the High Plains. In Prothero, D. R., and Emry, R. J. (eds.), *The Terrestrial Eocene-Oligocene Transition in North America*. Cambridge: Cambridge University Press, pp. 247–262.

Prothero, D. R. (ed.), 2001, *Magnetic Stratigraphy of the Pacific Coast Cenozoic*. Pacific Section SEPM Special Publication, **91**, 394 p.

Prothero, D. R., and Schwab, F., 2003. *Sedimentary Geology*, 2nd edn. New York: W.H. Freeman, p. 557.

Prothero, D. R., and Swisher, C. C., III, 1992. Magnetostratigraphy and geochronology of the terrestrial Eocene-Oligocene transition in North America. In Prothero, D. R., and Berggren, W. A. (eds.), *Eocene-Oligocene Climatic and Biotic Evolution*. Princeton: Princeton University Press, pp. 46–74.

Sadler, P. M., 1981. Sediment accumulation rates and completeness of the stratigraphic record. *Journal of Geology*, **89**, 569–584.

Watson, G. S., 1956. A test for randomness of directions. *Monthly Notices of the Royal Astronomical Society Geophysics Supplement*, **7**, 160–161.

Wood, H. E., II, Chaney, R. W., Clark, J., Colbert, E. H., Jepsen, G. L., Reeside, J. B., Jr., and Stock, C., 1941. Nomenclature and correlation of the North American continental Tertiary. *Bulletin of the Geological Society of America*, **52**, 1–48.

Woodburne, M. O., 1989. Hipparion horses: a pattern of worldwide dispersal and endemic evolution. In Prothero, D. R., and Schoch, R. M. (eds.), *The Evolution of Perissodactyls*. Oxford: Oxford University Press, pp. 197–233.

Woodburne, M. O., MacFadden, B. J., and Skinner, M. F., 1981. The North American Hipparion Datum and the implications for the Neogene of the Old World. *Geobios*, **14**(4), 493–524.

Zijderveld, J. D. A., 1967. AC demagnetization of rocks: analysis of results. In Collinson, D. W., Creer, K. M., and Runcorn, S. K. (eds.), *Methods in Paleomagnetism*. Amsterdam: Elsevier, pp. 254–286.

Cross-references

Magnetic Domains
Magnetometers
Paleomagnetism, Principles
Remanent Magnetism

PALEOMAGNETISM, MEASUREMENT TECHNIQUES AND INSTRUMENTATION

Tallavajhala Radhakrishna[1], J. D. A. Piper[2]
[1]Geoscience Division, Centre for Earth Science Studies, Trivandrum, India
[2]Geomagnetism Laboratory, University of Liverpool, Liverpool, UK

Synonyms

Ancient magnetism; Fossil magnetism

Definition

Paleomagnetism is the study of the Earth's magnetic field in past times. A permanent record of the ancient magnetic field is recorded by certain ferromagnetic minerals formed in rocks either as they form, and/or when they are

subjected to later geological events. The past direction, and sometimes the intensity, of the geomagnetic field can be determined from this fossil magnetism. This information is essential for quantifying the motions of the tectonic plates that comprise the outer shell of the Earth (Plate Tectonics) and for evaluating the history of the geodynamo seated within the Earth's core. Although this contribution is concerned primarily with the record of fossil magnetism in the rock, the acquisition of magnetism is an ongoing process within materials forming during the present day; thus it is found, for example, in soft sediment, soils, and in organic material such as magnetotactic bacteria.

Introduction

The permanent magnetism residing in a natural material, lodestone, has been known for many centuries for its ability to align, and it has long been used for this purpose by navigators. It was generally believed that the pole star (Polaris) or a large magnetic island at the North Pole attracted lodestone compasses. However, William Gilbert (1544–1603) in his book *De Magnete, Magneticisque Corporibus, et de Magno Magnete Tellure* published in 1600, concluded that the Earth was a giant magnet. He argued that the centre of the Earth was iron. He documented an important and related property of magnets by showing that they could be cut so that each new magnet possessed north and south poles. In the nineteenth century, it was found that rocks were capable of becoming permanently magnetized in the direction of the ambient geomagnetic field, and Bernard Brunhes (1867–1910) found that young volcanic lavas in central France were magnetized in a direction opposite to the prevailing field. This is the phenomenon of geomagnetic reversal and in 1929 the Japanese scientist Motonori Matuyama (1884–1958) was the first person to show that reversals of the field had occurred at specific times during the geological past.

The geometry of the magnetic field direction at the surface of the Earth is described in terms of declination and inclination. Declination refers to the angle in the horizontal plane between true north and the geomagnetic direction along which the lines of magnetic field are directed. Since the magnetic field is also directed toward, or away from, the center of the earth it is usually inclined to the horizontal. The angle above or below the horizontal is termed the inclination; it is latitude-dependant and varies from vertically up or down at the magnetic poles to zero at the equator. For paleomagnetic study to yield a useful geometrical parameter for resolving ancient continental movements, it is important to demonstrate that the magnetic poles are constrained to the geographic poles of rotation. Records kept in London show that the declination of the magnetic field direction changed from 11°E in 1580 to 24°W in 1820, while the calculated position of the magnetic pole executed a path around, and close to, the pole of rotation. This is secular variation. When averaged over intervals of 1,000 years or more, the mean magnetic pole has been found to lie very close to the pole of rotation and this led P.-L. Mercanton (1876–1963) to suggest that fossil magnetism could be used to monitor the movements of the continents. An outcome of research conducted during World War II was the development by Patrick Blackett (1897–1974) of a sensitive magnetometer able to accurately measure the direction of magnetism in rock samples. It was first employed in the mid-1950s by Blackett and Keith Runcorn (1922–1995) to show that continents had moved over the surface of the Earth. This evidence placed the theory of continental drift, which had languished since being proposed by Alfred Wegener (1880–1930) in 1915, on a firm observational basis. When it was realized that the ocean crust could also preserve paleomagnetism, continental drift was rapidly integrated in the 1960s with the hypothesis of sea floor spreading to develop the theory of Plate Tectonics, a paradigm that has prevailed, with much refinement, to the present day. During this time, paleomagnetism has proved to be one of the most important techniques in the Earth Sciences with applications to many other fields including environmental sciences and archeology.

In addition to its long-term value for tracing the movements of continents and evaluating paleogeography, paleomagnetism has been used to unravel the history of reversals of the geomagnetic field polarity and derive the geomagnetic polarity time scale (GPTS), the chronology of periods of normal and reversed polarity. This in turn is the basis of magnetostratigraphy, the use of magnetic reversals to date rocks and correlate them from place to place. The evaluation of the intensity and the direction of the geomagnetic field from rocks such as lava flows that are rapidly formed and magnetized can be used to resolve secular variation. This latter phenomenon embraces the continuous and ongoing changes in the strength and direction of the geomagnetic field and is thus a signature of behavior of the geodynamo. In addition the direction and the fabric of fossil magnetism in rocks have applications to sedimentology, resolution of paleostress, fluid-migration through rocks, and to mineral exploration.

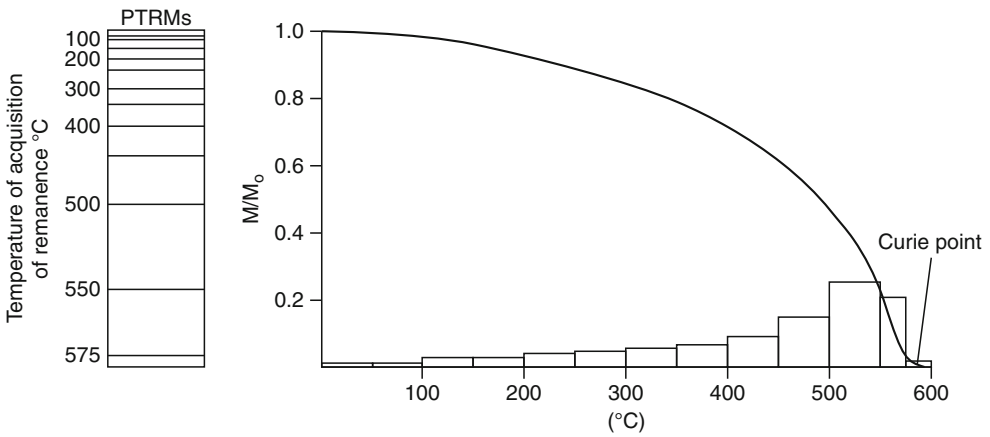
Rocks, like all natural materials, possess a weak *diamagnetism* in applied magnetic fields as the electron spins align with, and oppose, the field. Minerals containing transition elements such as iron and manganese also possess a stronger *paramagnetism* as unbalanced electron orbits align with and reinforce the field. However, these effects are transient and only survive as long as the applied field is present. Although important to the interpretation of the magnetic field over the ground, they are not relevant to paleomagnetism. Instead this relies on a memory of the magnetic field aligned with the ambient magnetic field lines at the time of rock formation, or at some later time period that can be determined from the geological history of the rock. In igneous and metamorphic rocks this magnetic memory resides in grains with ferromagnetic properties, which are usually iron-titanium oxides

(titanomagnetites) comprising a few percent of the total rock volume. Although the intensity of the Earth's magnetic field is quite weak (~0.5 T), as these rocks cool through the Curie point they are capable of acquiring a strong and stable magnetism from this field. The ferromagnetic minerals only have a transient weak diamagnetism and paramagnetism in the ambient field at high temperatures, but they acquire a permanent magnetization (ferromagnetism) as they cool through a critical temperature called the Curie point (generally <680°C). This is much lower than the temperatures at which many igneous rocks start to crystallize (~1,200°C) and is lower than the higher grades of metamorphism. If an igneous rock such as a dyke or lava flow cools rapidly at, or near to, the surface it will cool rapidly to this temperature over period of only days or weeks. It will then record a field direction that is instantaneous in geological terms and part of the secular variation rather than a long-term average of the field representative of the Earth's rotation axis. In contrast igneous rocks that cool slowly deep within the Earth's crust will integrate a record of the magnetic field over long periods of time; they may even record reversals of the geomagnetic field.

Since rocks contain magnetic grains with a wide diversity of size and magnetic domain properties, their magnetizations are acquired over a broad range of temperature below the Curie point. This trapping of remanence over a range from the Curie point down to ambient temperatures defines a blocking temperature spectrum (Figure 1). In igneous rocks the total spectrum is a Thermal Remanent Magnetization (TRM). This can be considered to be the sum of Partial Thermal Remanences (PTRMs), which obey the Law of Additivity: the remanence fraction acquired over any particular interval is discrete and independent of the fraction acquired over any other interval (Figure 1).

Many sedimentary rocks include grains of ferromagnetic minerals derived from the erosion of older rocks. When these grains are deposited they can be geomagnetically aligned if the influence of the Earth's magnetic field is stronger than gravitational and hydraulic forces. This is a Depositional Remanent Magnetism (DRM). Usually sediments undergo a protracted history of compaction and cementing (lithification) and sometimes chemical alteration (diagenesis), and may well also be disturbed by the burrowing of organisms (bioturbation). As these processes operate, the ferromagnetic grains usually have the opportunity to become more precisely aligned along the ambient geomagnetic field to give the rock a Post-Depositional Remanent Magnetism (PDRM). Chemical changes resulting in the precipitation of new ferromagnetic minerals produce a Chemical Remanent Magnetism (CRM). In sedimentary rocks such as limestone ferromagnetic grains may be formed directly by diagenetic processes or may incorporate magnetite grains precipitated by magnetotactic bacteria. Although these secondary processes are often still poorly understood, they operate over long time periods and therefore always tend to integrate the geomagnetic field to provide information that is of paleomagnetic rather than of short-term geomagnetic significance.

Paleomagnetism is a time relaxation phenomenon since all natural magnetizations decay with time and relax into the field prevailing at the present day. Nevertheless this relaxation time is very long, and longer than the age of the Earth for many small ferromagnetic grains. It is the very great length of time required for some natural magnetizations to decay, which means that a record of permanent magnetism is still preserved, even from the oldest rocks on Earth. Nevertheless, it also means that if a rock is metamorphosed at an elevated temperature below the Curie



Paleomagnetism, Measurement Techniques and Instrumentation, Figure 1 A blocking temperature spectrum of Thermal Remanent Magnetization (TRM) showing the progressive fraction reduction of an initial magnetization M_o. Each segment of the TRM is referred to as a Partial Thermal Remanence (PTRM) and these segments obey a Law of Additivity: TRM = PTRM (550–500°) + PTRM (500–450°) + PTRM (450–400°) +.

point and this temperature is sustained for many millions of years, it is capable of being totally demagnetized. It will then acquire a new magnetism recording the time of metamorphism.

In the late 1950s it was also recognized that the rocks commonly acquire secondary magnetizations by a range of processes that include not only later temperature elevation, but also chemical change. The total magnetism in a rock, referred to as the Natural Remanent Magnetization (NRM), is then the vector resultant of the primary magnetization acquired when the rock was formed and the secondary magnetization (which may include several components) acquired during subsequent geological times, or even during coring and sampling. The chief goal of paleomagnetic investigations is to isolate the various components of remanence, and ascribe an origin, age, and reliability to them. The resolved directions from each component may then be used to compute geomagnetic (short-term) or paleomagnetic (long-term) pole positions. This calculation usually assumes that the source of the field can be approximated by an axial geocentric dipole (GAD) constrained to the Earth's rotation axis. The following sections describe the laboratory procedures and techniques of analysis employed in paleomagnetism to achieve these goals.

Techniques of sampling, measurement, and analysis

Field sampling

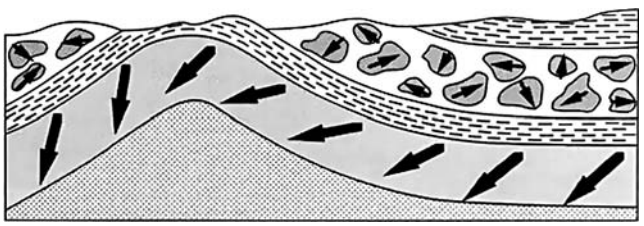
Paleomagnetic sampling procedures follow a hierarchical approach to average out the secular variation and compensate for other inaccuracies. The sampling strategy will be adapted to the time scale of the scientific objective. Typically oriented samples are collected from a number of different localities termed *sites*, each of which can be considered to represent a single time-instant record of the geomagnetic field. A number of such sites will be sampled within a single rock formation. The rock formation may comprise, for example, a suite of lava flows recording a volcanic episode, an igneous intrusion, or a succession of sedimentary beds. The sites clearly need to be in situ, cover a significant regional area, and comprise a representative distribution through the time interval represented by the rock formation. This will ensure that although units such as lavas may record an instantaneous record of the field, the time interval covered by the collective survey is adequate ($\sim 10^5$–10^6 years) to average the effects of secular variation and any longer term excursions of the geomagnetic field. It can then derive a long time-average assessment of the paleomagnetic field direction. Each site will constitute a number of separately oriented samples (N) to compensate for orientation and instrumental errors, and for possible failures to completely separate magnetic components during laboratory treatment. Sample orientation procedures typically have an uncertainty of $\sim 3^\circ$ and precision is gained in proportion to $N^{1/2}$ where N is the number of samples; hence to improve the precision from 3° to 1°, nine or ten individually oriented samples are required. In the case of magnetostratigraphy, where determination of the polarity of a rock level is the major objective, three samples are usually regarded as sufficient to determine the polarity unambiguously.

The samples are taken as oriented cores drilled in situ using portable drilling equipment (Figure 2), or are extracted in the laboratory from oriented block samples; the former method is preferred both because initial accuracy is higher, and because accuracy is lost when block samples are reoriented for laboratory drilling. There are several orientation schemes currently in use but each one employs an orientation device designed to provide an in situ geographic orientation using magnetic and Sun compasses (Figure 2). Since the magnetism of the rock outcrop can influence the reading of a magnetic compass, Sun-compass observations are preferred and use the direction of the shadow, together with a record of the time and sampling location; the azimuth is determined with the aid of the Astronomical Almanac or a simple algorithm. A new orientation device developed by C. Constable and F. Vernon at the Scripps Institution of Oceanography (Lawrence et al., 2009) uses the Global Positioning System (GPS) and laser technology to produce similar accuracies to the Sun compass, although its application requires at least an additional half hour. Nonetheless, achieving sun-compass accuracy is a major breakthrough for paleomagnetic field procedures in regions where siting of the Sun may not be possible. The paleomagnetic specimens recovered for laboratory analysis are typically cylinders $\sim$1 in. long and 1 in. in diameter (Figure 2).

A paleomagnetic field sample should be integrated with geological evidence so that it incorporates widely separated localities linked to defined rock formations for which the tectonic setting is properly understood; this will enable any effects of tectonic deformation on the recovered directions of magnetization to be accounted for. The survey will also need to accommodate any effects of regional or local heating by metamorphism or igneous activity, and also the possible influence of any chemical alteration, for example, by hydrothermal activity. It is also best to avoid elevated exposed locations because these are liable to have been struck by lightning. Lightning is a line current and the associated magnetic fields can induce a strong Isothermal Remanent Magnetization (IRM), which is capable of largely or completely replacing an ancient remanent magnetism in the rock.

A number of paleomagnetic field tests exist, which can be applied to help constrain the age of the magnetization in a rock formation. The fold test compares the directions of magnetization derived from two limbs of a fold before and after adjusting for the tilting of the beds (Figure 3); an improved grouping after adjustment indicates that the magnetization was acquired before the folding episode occurred. Sometimes the optimum grouping of directions from the limbs is attained at an intermediate stage of untilting, which would indicate that the magnetization was acquired during the tectonic event responsible for

Paleomagnetism, Measurement Techniques and Instrumentation, Figure 2 Collecting and orienting rock samples for paleomagnetic study. The cores are drilled (**a**) with a portable motorized coring machine. (**b**) An orienting tool is inserted around the core to determine the inclination and the azimuth by Sun and magnetic compasses. (**c**) The core is extracted and coded. (**d**) A line on the core is used to fix the extracted core with respect to the orientation readings using a directed line on the top or the bottom.



Paleomagnetism, Measurement Techniques and Instrumentation, Figure 3 Schematic cross section of ground showing a fold and conglomerate layer with magnetic components (*black arrows*) illustrating the concepts behind the fold and conglomerate tests used in paleomagnetism.

the folding. The conglomerate test compares the directions of magnetization in a number of pebbles in a boulder bed, preferably when these pebbles can be linked to a source in an underlying bed (Figure 3). If the directions are random it demonstrates that pebble bed has not been remagnetized since formation and that the magnetization in the bed that they were derived from is likely to be primary. If the directions of magnetization in the pebbles are uniformly directed, it is apparent that a later remagnetization event has influenced the formations including the pebble bed. The reversal test compares directions of opposite polarity to evaluate whether they are statistically antiparallel and therefore likely to be the record of a reversing dipolar field. Each of these tests can be placed on a sound statistical basis using methods described in textbooks on paleomagnetism (e.g., Tauxe, 1998).

Measurement and treatment

The Natural Remanent Magnetization (NRM) of a rock sample is the vector sum of the remanent magnetic moments of all the magnetized particles contained within it. These particles are likely to have a wide range of sizes and magnetic domain structures, so the record of primary to secondary magnetization (or even no preserved remanent magnetism at all) will vary greatly. Magnetometers measure the magnetization of rock samples in three mutually perpendicular directions (X, Y, and Z) and the values are compounded to yield the magnitude and direction of the resultant vector. Using the field orientation parameters for the sample (usually as noted above, in the form of a short cylindrical core) the data from the specimen coordinate system can be transformed into geographic coordinates and displayed graphically on a stereonet. The geographic vector can be considered in terms of declination (D), inclination (I), and magnitude (M); M does not change during the transformation of coordinate systems.

The observed NRM of rocks may consist of any combination of primary and secondary components and it is essential to separate these to unravel the magnetic history of the rock. This separation is achieved by partial demagnetization procedures referred to as "cleaning," which rely

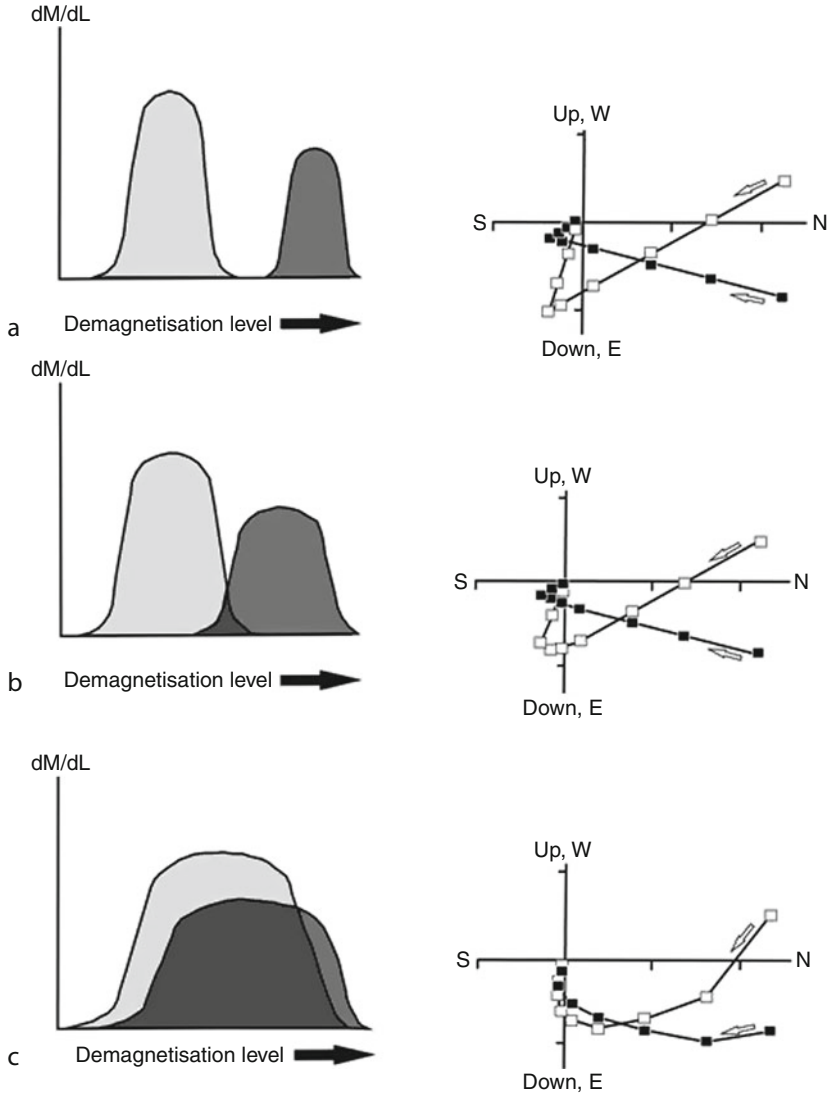
on the observation that magnetic components of different ages and modes of acquisition usually have contrasting thermal and magnetic blocking temperature spectra. Alternating field and thermal demagnetization techniques are now a routine part of paleomagnetic investigations and depend on the relationship of relaxation time to coercivity and temperature to subtract (demagnetize) the lower coercivity or blocking temperature components. Alternating field (AF) demagnetization selectively removes components with short relaxation times that typically have low coercivities. The basis of the thermal demagnetization technique is that these grains also have low unblocking temperatures: the lower the relaxation time, τ, the more likely the grain will carry a later secondary magnetization. All ferromagnetic grains are subject to a number of competing internal forces (see, e.g., Butler, 1992 and Tauxe, 1998) and larger grains tend to subdivide into zones of homogeneous magnetization in order to minimize the total magnetic energy. These *domains* are separated by narrow domain walls typically up to 100 atoms thick where the magnetic dipoles are canted around from one orientation to the other. In magnetite (Fe_3O_4)-bearing rocks, the secondary NRM tends to be carried mostly by multidomain (MD) grains with lower relaxation times while the primary characteristic remanence (ChRM) tends to be retained by single domain (SD) or pseudo-single domain (PSD) grains; the latter have sizes larger than SD but have shape or defect properties that enable them to retain remanence like SD grains. MD grains have coercivities predominantly $<$20 mT, whereas SD and PSD grains have higher coercivities. AF demagnetization can thus selectively remove secondary NRM and leave the primary magnetization unaffected. Rocks containing the less common ferromagnetic minerals pyrrhotite ($Fe_{1-x}S$) and maghemite have properties that are fairly comparable to magnetite but hematite (αFe_2O_3), which tends to be the commonest ferromagnetic constituent in sedimentary rocks, has much higher coercivity and cannot be effectively subtracted by AF demagnetization. Since hematite has a higher Curie point ($\sim$670°C) than the other common ferromagnetic minerals, it requires demagnetization to higher temperatures. Since hematite can also be an oxidation product of magnetite in igneous and metamorphic, and sometimes in sedimentary rocks, its magnetic remanence may only be subtracted by thermal demagnetization after the magnetite remanence has been removed. In this specific case, the primary magnetization in magnetite may be unblocked below a secondary magnetization residing in hematite.

AF demagnetization is unable to completely subtract the magnetism in any rock – the applied alternating fields required to do this are simply unattainable in the laboratory and would in any case, present their own problems. Thermal demagnetization, however, is capable of fully subtracting magnetization if it is continued to the Curie point. It is applied by heating the sample to a specified temperature and cooling the sample in a zero magnetic field so that all magnetization blocked below the assigned temperature is then unblocked and removed. Demagnetization proceeds in increasing temperature steps until the Curie point is passed and all remanence in the rock has been subtracted. The experimental procedure is best conducted in a field-free space so that magnetization cannot readily be imparted from the ambient field before the treated samples have been measured.

Analysis

A stereographic plot of magnetization directions at each stage of demagnetization treatment, although generally used in older studies (see Collinson, 1983 for details of plotting), is not a satisfactory way of resolving the components of remanence because these directions will likely record the resultant of two or more components. Instead it is usual to illustrate the vector (with three-dimensional properties) in two-dimensional space by projecting it onto two planes at right angles. These projections can be combined together on a single projection and are described as "orthogonal plots." They were first applied to a paleomagnetic study by Wilson (1961) and the display technique was formalized by Zijderveld (1967). The *D*, *I*, and *M* data of the progressive demagnetization experiments may be resolved into north, east, and downward (or upward) components to construct a vector component diagram (Figure 4). The base of the NRM vector is placed at the origin of the Cartesian coordinate system and the tip of the vector is projected onto two orthogonal planes. The distance of each data point from the origin is proportional to the intensity of the NRM vector projected onto the plane. The projections of the NRM vectors onto the horizontal plane are constructed by plotting N versus E and the second projection plots progressive NRM data onto a vertical plane. A vertical component of the NRM vector at each demagnetization level is plotted against the north (or east) component. It is a common practice to denote projections on the horizontal plane and vertical planes by solid and open symbols, respectively.

An important observation is that a linear trend of projected points defines the progressive removal of one component only (Figure 4a). If this line migrates toward the origin of the projection it demonstrates that only the high unblocking temperature or coercivity magnetization is left to be removed. Thus the successive removal of the lower unblocking temperature or coercivity magnetizations, and the isolation of the highest unblocking temperature/coercivity component, can be achieved by the recognition of successive linear trajectories. The last component to be resolved is often designated as the Characteristic Remanent Magnetization (ChRM) and is usually the most interesting component for interpretive purposes, although the secondary components removed below this may be of interest for resolving the later history of the rock, especially if they can be linked to a later event such as a thermal pulse or a diagenetic episode. The direction of any component represented by a straight-line segment on an orthogonal plot can be recovered by extracting that



Paleomagnetism, Measurement Techniques and Instrumentation, Figure 4 The orthogonal projection of demagnetization data during progressive demagnetization. When (**a**) there is no overlap between the blocking temperature/coercivity spectra of the sample, the lower and higher components are defined by discrete straight-line segments as they are unblocked by thermal demagnetization or subtracted by AF demagnetization. As the overlap of the spectra increases (**b, c**), the orthogonal projection becomes progressively more curved and linear segments are no longer observed when both spectra are being subtracted at the same time (**c**). Note that declinations can be read directly from the horizontal projection of the vector although a direct reading of inclination is not possible. (After Collinson, 1983.)

component from the data above and below and treating it as a unique component.

The shape of the trajectory of vector projections depends upon the degree of overlap of the coercivity or blocking temperature spectra of the components (Figure 4). When the spectra do not overlap, the vector diagram displays successive linear trajectories, each one recording the subtraction of a component. If overlap is only partial, straight-line trajectories are still identified but intermediate points corresponding to the overlap display a curved trajectory (Figure 4b). If there is complete overlap between the spectra the vector component diagram does not possess linear segments (Figure 4c) but the trajectories may sometimes be used to isolate the components if they define a great circle, a "circle of remagnetization." A single great circle on its own provides insufficient information to estimate the ChRM direction, but if more than one circle is available from different specimens and these circles converge, then the convergence point, or its antiparallel equivalent, can yield an estimate of the ChRM. Halls (1979) reports a procedure for calculating the best-fitting convergence point of several great circles. The majority of convincing paleomagnetic results come from one and two component structures, although paleomagnetic workers

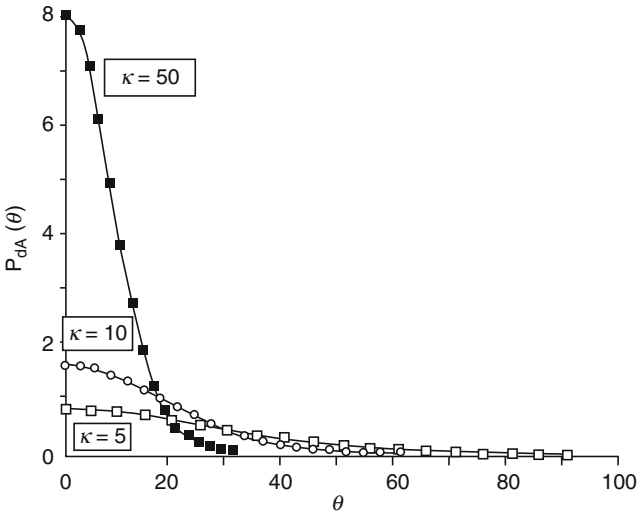
have ventured into the study of rocks with complex magnetic histories comprising multicomponent NRMs, and combinations of AF and thermal demagnetization have been used to isolate complex remanence structures.

The procedure for separating and resolving the directions of individual components is Principal Component Analysis (PCA) and comprises a method for determining the direction of the lines or planes that best fit the data described by the orthogonal projections (Kirschvink, 1980); essentially a component segment is isolated from the orthogonal projection and isolated as though it is the only component present in the rock. An advantage of this method is that it compensates for errors due to signal to noise ratio, and provides a quantitative measure of the precision of the direction. The precision of the direction determined from the best-fit line is described in terms of a Maximum Angular Deviation (MAD) and MAD values <15 are usually considered to be acceptable.

Statistics

As with all experimental data, the directions resolved by PCA will be associated with a number of random errors arising from factors such as instrument noise, inaccuracies in sample orientation, imperfections in demagnetization, and secular variation of the recorded geomagnetic field. The net result of all these factors is a dispersion from which the researcher requires the best estimate of the mean direction. Determining a mean and confidence limit from any set of values by a statistical method is based on a probability density function. A function applicable to paleomagnetic directions was developed by Fisher (1953) and provides a simple procedure for treating paleomagnetic data. Fisher's analysis treats each accepted direction in the population as having a unit weight and represents it as a point on a sphere of unit radius. The dispersion of these points on the sphere is assumed to have a circular Gaussian distribution (Figure 5) and the mean of the set of directions is calculated by vector addition: each unit vector is resolved into three orthogonal components in Cartesian coordinates and the components are summed to obtain the length of the resultant vector (R) and the mean direction (D and I). D and I define the direction of the best estimate of the mean while R is used to assess confidence in the result.

Two simple parameters, k and α_{95}, are used to illustrate the quality of the calculated mean in paleomagnetic studies. The probability of finding a direction at an angle (θ) from the calculated mean is expressed in terms of a probability function involving a precision parameter κ. Fisherian statistics use an approximation for κ defined by a parameter denoted by $k = (N-1)/(N-R)$ where N is the number of directions and R is the length of the resultant vector. The α_{95} value in degrees defines the angular radius of the cone of confidence projected from the center of the unit sphere within which we have 95% confidence that the true direction actually lies. It is simply calculated from $\alpha_{95} = 140/(k \times N)^{1/2}$ and is comparable to the standard



Paleomagnetism, Measurement Techniques and Instrumentation, Figure 5 The Fisher circular distribution derived from the Gaussian probability function. $P_{dA}(\theta)$ is the probability per unit angular area of finding a direction within an angular area, dA, centered at an angle θ from the true mean. Distributions are shown for values of the precision parameter κ, of 50, 10, and 5.

estimate error of the mean in Gaussian statistics; 5% of the observed directions are expected to fall outside of this circle. High reliability in paleomagnetic analysis is therefore expressed by large k and small α_{95}. As with all statistical analysis, precision estimates are improved by increasing the sample population although in terms of experimental effort the increase becomes progressively less worthwhile for large values of N. Sample populations from a single unit such as a lava flow are generally considered satisfactory if N is 7 or more, while k values <10 are unsatisfactory.

Determination of paleomagnetic poles

The resolved mean direction of magnetization is used directly in many analyses and interpretations – in regional, small- to medium-scale tectonic problems. However, for larger scale analysis such as the determination of past global geometries and analysis of the geomagnetic field it is necessary to convert it into an equivalent pole position. This also provides a meaningful way of comparing ChRM directions from rock formations of same age, but widely separated on the same continent or tectonic plate; for this purpose a common reference system is necessary and is provided by the paleomagnetic pole. The pole position is the position where the geomagnetic field axis responsible for the observed paleomagnetic field direction meets the surface of the Earth. A model is required to perform this calculation. Since more than 90% of the present geomagnetic field can be explained in terms of an axial geocentric dipole source, this is the model normally used to calculate paleomagnetic poles and the position is

expressed in the present geographic longitude – latitude grid. For a paleomagnetic field of inclination I, the ancient latitude (λ) of the site is given by the simple formula Tan $I = 2$ Tan λ. The equations required to calculate the position of the paleomagnetic pole require the declination D and the coordinates of the sample locality in addition to the inclination. They are given in all paleomagnetic text books (see, e.g., Tauxe, 1998).

The geomagnetic pole calculated from a single geomagnetic field observation (e.g., from a single lava flow) yields a Virtual Geomagnetic Pole (VGP) and will include components of the non-dipole field incorporating the secular variation. A paleomagnetic pole is calculated from a number of VGPs derived, for example, from several lava flows or levels through a sedimentary section and is anticipated to have averaged non-dipole field components so that it represents the Earth's rotation axis at the time that the ChRM was acquired by the rocks. The angular dispersion of the VGPs constituting the collection can be used to test whether this is effectively achieved by comparison with the dispersions achieved by secular variation during historic and recent geological time (McFadden et al., 1991). The expected angular dispersion is a function of latitude although the comparison is presently of limited value because the comparative database is sparse in the southern hemisphere and temporal coverage is poor. Confidence limits for a paleomagnetic pole are derived from α_{95}, the radius of the cone of 95% confidence about the mean direction. These translate into an oval of confidence within which we have 95% confidence that the pole actually lies. The maximum and minimum radii of this oval are given by $dp = 2\ \alpha_{95}\ (1/(1+ 3\text{Cos}^2 I))$ and $dm = \alpha_{95}$ (Sin p/Cos I) where p is the magnetic colatitude or the great-circle distance from the sampling location to the pole.

Instrumentation

The magnetization intensity can vary by several orders of magnitude between different rock types. Strong magnetizations can be rapidly measured with instruments of low sensitivity while weak magnetizations require more sensitive instruments and the measurements may need to be integrated over longer periods of time. The magnetometer developed by Blackett (1952) utilized the deflection of a suspended astatic magnetic system. Since this system is sensitive to vibration and measurements are relatively slow, it is now obsolete. Spinner magnetometers are used to measure magnetizations of strong to intermediate intensities, including those found in many igneous and metamorphic rocks. These magnetometers are less suitable for sedimentary rocks, which can often be weakly magnetized, and superconducting magnetometers (superconducting quantum interference devices-SQUIDs) are then required.

Spinner magnetometers

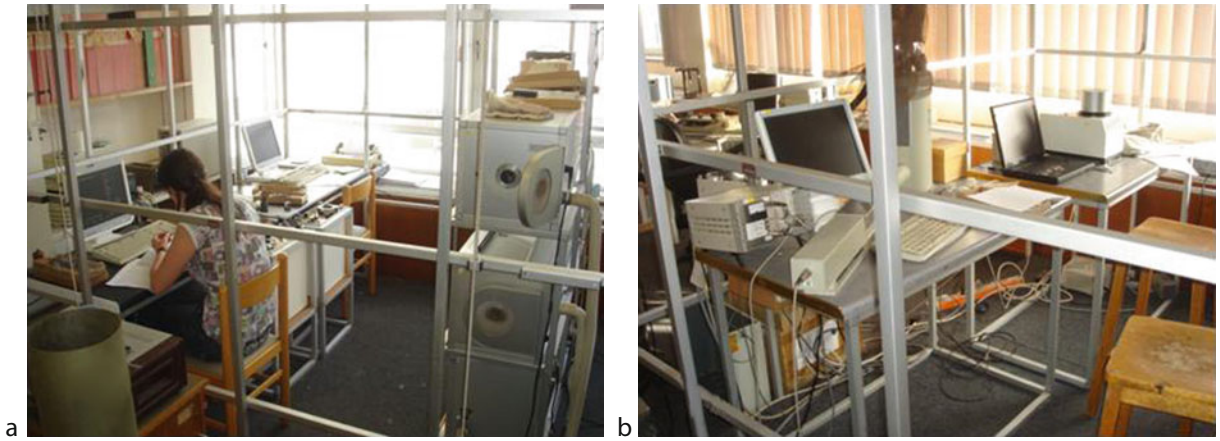
In the spinner magnetometer the sample is spun adjacent to a detecting coil. A fluctuating electromotive force (emf) is induced proportional to the magnetization with the amplitude of the output voltage proportional to the component of magnetic moment perpendicular to the rotation axis ($P_\perp$). The phase of the voltage is utilized to relate the direction of the measured component to a reference direction in the sample. The sample is rotated successively about three axes to obtain average values of the NRM components and reduce the effect of inhomogeneity to $\sim 2^\circ$ in direction and $\sim \pm 5\%$ in intensity. During the early development of spinner magnetometers, measurement of weak magnetizations was achieved by increasing the rotation speed and by careful design of the pick-up coil system. Since electric motors operating at a frequency of ≈ 100 Hz are required for rotating the samples, a challenge to the design has been the requirement to shield this unavoidable source of electromagnetic noise from the measuring assembly. This problem can be reduced by using an air turbine, and in modern magnetometers, the frequencies used are in the range 5–80 HZ with signal discrimination circuits employed to reduce the effects of noise.

The usual procedure for NRM measurement is to spin each sample in six mutually orthogonal orientations so that any two of X, Y, and Z are in turn perpendicular to the spin axis. Since two orthogonal components of magnetization are obtained from each spin, the orientations are chosen to provide four determinations each of X, Y, and Z, usually two of each sign for each component. This procedure tends to average out any random noise on the signals, the effects of inhomogeneity of NRM, and any remanent magnetism in that part of the sample holder permanently attached to the shaft. The six sample orientation used should ideally satisfy the following: (a) provide two values each of $+x$, $+y$, and $+z$, (b) distribute the displaced dipole caused by inhomogeneity of the NRM as uniformly as possible in space, (c) average the possible error arising from imperfect calibration of the NRM direction of the magnetometer, and (d) measure each NRM component in each "channel" of the magnetometer, that is, the sample is placed in the holder with X, Y, and Z each in turn parallel and perpendicular to the reference direction.

Some commercial spinner magnetometers (Figure 6b) can now achieve sensitivities close to those attainable with SQUID magnetometers, but since they need to be run at high speed they are unsuitable for friable or poorly shaped samples.

Cryogenic magnetometers

Developments in cryogenic technology have had important implications for paleomagnetic analysis. Firstly they have permitted an increase in sensitivity by one to three orders of magnitude and the rapidity of measurement has permitted study of a wider range of rock types than has been possible hitherto. Secondly it is possible for magnetism in a sample to be measured by a single insertion into the sensing coils because only one axis of the sample needs to be presented to the magnetometer. Magnetic



Paleomagnetism, Measurement Techniques and Instrumentation, Figure 6 Instrumentation in a typical paleomagnetic laboratory. Measurements and demagnetization procedures take place within large sets of Rubens coils designed to continuously nullify the ambient geomagnetic field. In (**a**), two thermal demagnetizers are stacked on the right; these are designed for the treatment of trays of multiple cores to specified temperatures for programmed time intervals. On the left, the operator is using one of two interfaced spinner magnetometers for measurement of relatively strongly magnetized rock samples. (**b**) Two magnetometers used for the measurement of weakly magnetized rock samples are an FIT hydrogen SQUID magnetometer on the left and a JR6 high speed spinner magnetometer on the right.

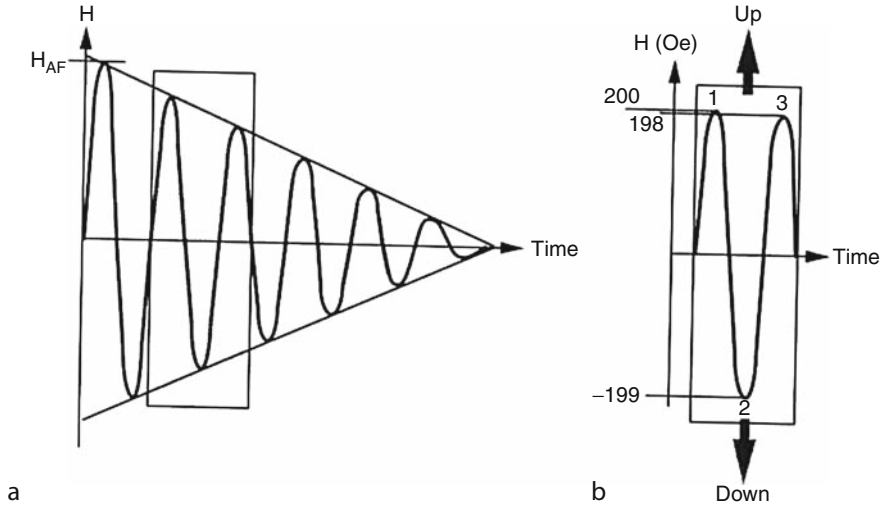
and thermal cleaning can then potentially be automated with the measuring procedure. Cryogenic magnetometers use the properties of a superconducting ring made of a metal with perfect diamagnetism, which causes the expulsion of all magnetic flux from the ring below a certain temperature. When an external field is applied in the superconducting state all flux is expelled from the interior of the metal but not from the hole within the ring; this means that induced currents flow in one direction on the outer face of the ring and in the opposite direction on the inner face. If the applied field is removed the currents on the outside of the ring disappear but the current on the inside face persists to maintain a field within the ring.

The ability of the ring to trap and maintain magnetic fields to better than one part in 10^9 can be used in magnetometry by trapping a zero field within the loop and then using this volume for measurement. These magnetometers are operated at the superconducting: resistive boundary. The ring current is quantized so any flux entering the ring at this transition state is also quantized and is measured in terms of a feed-back current. The magnetometer is surrounded by an insulating evacuated space cooled by liquid helium or nitrogen and pick-up coils connected to SQUID detectors are employed to measure the NRM along one axis or simultaneously along two or three mutually perpendicular axes. Mumetal shields are employed to blanket external field variations and some automated systems allow complete demagnetization within this field-free environment. Cryogenic magnetometers are faster and typically more sensitive than spinner magnetometers but cost much more to purchase and operate; the liquid nitrogen represents an ongoing cost, and while helium can achieve lower temperatures it is more expensive and less commonly used these days. For further details on these magnetometers the reader is referred to Collinson (1983) and Tauxe (1998).

Alternating field demagnetizers

In AF demagnetization, an oscillating field is applied to a paleomagnetic specimen in a nullified magnetic field environment. The equipment comprises a coil, a capacity bank for tuning the coil, a power supply and current reduction mechanism, a coil system or magnetic shield for providing a low ambient field over the sample, and a sample tumbler. The rock specimens are rotated within the center of the coil in a tumbling system incorporating simultaneous rotation about horizontal and vertical axes. The ratio of the two tumbling rates is designed to expose as many configurations of the sample as possible to the applied field acting along the axis. The applied current is then smoothly and gradually reduced to zero as the specimen continues to rotate. Most modern AF demagnetizers are designed to achieve initial fields in excess of 100 mT. It is essential to ensure that the geomagnetic field within the demagnetizer is completely cancelled or the specimen may acquire an Anhysteretic Remanent Magnetization (ARM), and an alternating field with a very pure waveform is essential for the best results than avoid rotational and gyro magnetic effects.

The theory of AF demagnetization is explained by Butler (1992). The waveform of the applied alternate magnetic field is a sinusoid with linear decrease in magnitude with time (Figure 7). The expanded figure (b) emphasizes the smooth decrease required. All the grain moments with coercivities below the peak AF will track the field. These entrained moments will become stranded, however, as the peak field drops below the coercivities of individual magnetic domains. The directions of stranding will be



Paleomagnetism, Measurement Techniques and Instrumentation, Figure 7 (**a**) The generalized wave form of the magnetic field used during AF demagnetization and (**b**) expanded version of a small portion of AF demagnetization between two successive peaks. (After Butler, 1992.)

effectively random if the tumbler has been appropriately designed to present as many configurations as possible of the sample to the declining applied field. By progressively increasing the applied field, the domains of successively higher coercive force are randomized or subtracted from the total measured remanence. The coercive spectra of the sample can then be resolved up to the maximum field of demagnetization and contributions of successive components to the NRM resolved by application of orthogonal projections and PCA. As noted above, AF demagnetization is unable to fully subtract the NRM in a rock and is ineffectual with hematite-bearing rocks due to the high coercivity of this mineral.

Thermal demagnetizers

Thermal demagnetizers can serve the purpose of both demagnetizing the NRM in paleomagnetic specimens and applying a partial thermal remanent magnetization (pTRM); the latter procedure is used for investigating the record of the ancient intensity (paleointensity) preserved in some rocks. If the temperature of the rock is raised until the relaxation time of some fraction of the particles carrying NRM has been reduced to a few minutes or less, the NRM of these particles will be unblocked. Their contribution to the NRM of the rock will be lost if they are cooled to the ambient temperature in a zero field. In this way, the contribution of lower stability grains to the NRM can be randomized. Alternatively, if a DC field is applied during cooling, the grains whose unblocking temperatures have been exceeded will be realigned into the new applied field direction; they will then have acquired a pTRM. If the rock sample is subjected to successively higher temperature treatments, the lower blocking temperature components are progressively removed. The components contributing to the NRM are then progressively subtracted and can be resolved by projecting the successive magnetizations on an orthogonal projection and applying PCA.

Thermal demagnetizers comprise a magnetic shielding system to exclude the ambient field, a furnace, a thermocouple, and a temperature control unit. A prime concern is the control external magnetic field, while the furnace should be of sufficient size to hold a reasonable number of samples (10–50) and to give a uniform temperature through them ($<\pm 10^{\circ}$C at 600°C). The winding of the heating coil onto the curved surface of a silica tube forming the inside wall of the furnace is noninductive to reduce stray magnetic fields. It is required to be capable of heating samples to a temperature ($\sim$700°C) in excess of the highest Curie points of the ferromagnetic minerals over a reasonable time ($\sim$1 h). The available maximum voltage is usually that of the electrical mains. The maximum current is dictated by that of the temperature controller being used, or of a variable transformer, and thus the resistance of the furnace winding is determined. Between 20°C and 700°C the resistance of the winding will increase (by 5–10%). If the resistance determined from the available voltage and current is the cold resistance, the current (and heating rate) decreases at high temperatures, whereas if it is the hot resistance the cold current will be slightly above that calculated. The wire diameter depends on the size of the furnace and the number of turns with a minimum separation of the turns set by the requirement for sufficient insulation between them; a wide separation is detrimental to temperature uniformity.

Although large furnaces are useful for dealing with many samples at once, their thermal inertia is considerable and may cause problems with temperature control. A suitable thermocouple material is Pt/13%Pt-Rh, with an output of $\sim$6.7 mV at 700°C. The thermocouple tip may be covered with a standard rock sample to the center of which an axial hole has been drilled. It is then likely that

the rock samples will be uniformly heated throughout the indicated temperature, by ensuring that the thermal inertia of rock and thermocouple are similar. Several commercial temperature controllers are available for automatically raising and holding the furnace temperature to a preselected value. The controller senses the furnace temperature via a thermocouple placed close to the furnace winding.

Some furnaces are designed so that it is possible to conduct heating/cooling cycles in an inert atmosphere or in a vacuum to minimize oxidation. Although thermal demagnetizers are in principle, capable of completely removing the NRM in a rock and thereby recovering the entire component structure in a way that may not be resolvable by AF demagnetization, this is seldom fully achievable in practice. The limitations are the difficulty of completely canceling the ambient magnetic field without excessive shielding and the possibility of chemical alteration during the successive heating and cooling cycles. To surmount the latter problem, microwave systems are in the process of development. They have been successively applied to paleointensity studies by a few laboratories (Walton et al., 1992) but have not yet been produced with a capability for routine NRM demagnetization.

Summary

A record of the past magnetic field, known as magnetic remanence, is preserved by ferromagnetic mineral grains in rocks and other natural materials. These minerals are mostly oxides of iron (magnetite, hematite, maghemite) and some types of sulfide (pyrrhotite). Although they usually comprise no more than a few percent of the total rock volume, they provide key information for geology and geophysics, and for the wider fields of the environmental sciences and archeology. The material for study is collected in the form of small oriented samples in sufficient numbers and distributed amounts to be representative of a sample of the ancient magnetic field for short periods of time (study of secular variation and short-term geomagnetic field behavior, archeological applications) or long periods of time (geology and geophysics). The magnetizations are measured using magnetometers designed for determining a range of strong to weak natural magnetizations that comprise spinner magnetometers based on the current induced in a detector system as the sample is spun near to a coil, or SQUID magnetometers based on cryogenic technology.

The magnetizations in most materials are composite effects comprising the vector sum of a primary magnetization imparted when the rock was formed and secondary magnetizations acquired during later events that may have influenced the rock. These component magnetizations are separated and resolved by demagnetization procedures that progressively subtract the magnetization resolved either in terms of the coercivity spectrum of the magnetization (AF demagnetization) or the blocking temperature spectrum (thermal demagnetization). The magnetizations are resolved by projecting the successive values of the vector during this progressive treatment onto a two-dimensional orthogonal plot and identifying the constituent components from their contributions to the coercivity/blocking temperature spectra. The populations of magnetic directions resolved from a paleomagnetic study are statistically analyzed using a system of (Fisherian) statistics assuming that each direction is of equal importance and the population can be represented in terms of a circular Gaussian distribution. Two simple parameters comprising a precision parameter k, and a cone of 95% confidence α_{95} are used to express the quality of the result. The mean direction can be converted into an equivalent geomagnetic of paleomagnetic pole by assuming that the source of the field direction resolved at the surface of the Earth originates in a geocentric dipole source in the Earth's core. In the long-term application, this dipole may also be assumed to be axial and constrained to the rotation axis of the Earth.

Bibliography

Blackett, P. M. S., 1952. A negative experiment relating to magnetism and the Earth's rotation. *Philosphical Transactions of the Royal Society London*, **A245**, 309–370.

Butler, R. F., 1992. *Paleomagnetism: Magnetic Domains to Geologic Terranes*. Cambridge: Blackwell, p. 319.

Collinson, D. W., 1983. *Methods in Rock Magnetism and Paleomagnetism*. London: Chapman and Hall, p. 500.

Fisher, R. A., 1953. Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A*, **217**, 295–305.

Halls, H. C., 1979. Separation of multicomponent NRM: combined use of difference and resultant magnetization vectors. *Earth and Planetary Science Letters*, **43**, 303–308.

Kirschvink, J. L., 1980. The least-squares line and plane and the analysis of paleomagnetic data. *Geophysical Journal of Royal Astronomical Society*, **62**, 699–718.

Lawrence, K. P., Tauxe, L., Staudigel, H., Constable, C., Koppers, A., McIntosh, W. C., and Johnson, C. L., 2009. Paleomagnetic field properties at high southern latitude. *Geochemistry Geophysics Geosystems*, **10**, QO1005, doi:10.1029/2008GC002072.

McFadden, P. L., Merrill, R. T., McElhinny, M. W., and Lee, S., 1991. Reversals of the Earth's magnetic field and temporal variations of the dynamo families. *Journal of Geophysical Research*, **96**, 3023–3933.

Tauxe, L., 1998. *Palaeomagnetic Principles and Practise*. Boston: Kluwer. p. 299.

Walton, D., Shaw, J., Share, J. A., and Hakes, J., 1992. Microwave demagnetization. *Journal of Applied Physics*, **71**, 1549–1551.

Wilson, R. L., 1961. Palaeomagnetism in Northern Ireland. Part 1. The thermal demagnetization of natural magnetic moments in rocks. *Geophysical Journal of Royal Astronomical Society*, **5**, 45–58.

Zijderveld, J. D. A., 1967. A.C. demagnetization of rocks: analysis of results. In Collinson, D., Creer, K., and Runcorn, S. (eds.), *Methods in Paleomagnetism*. Amsterdam: Elsevier, pp. 254–286.

Cross-references

Archaeomagnetism
Curie Temperature
Geodynamics
Geomagnetic Field, Polarity Reversals
Plate Tectonics, Precambrian
Plates and Paleoreconstructions

PALEOMAGNETISM, POLAR WANDER

Jean Besse[1], Vincent Courtillot[2], Marianne Greff[2]
[1]Laboratoire de Paléomagnétisme, Institut de Physique du Globe de Paris, Paris, France
[2]Institut de Physique du Globe de Paris, Paris, France

Definition and introduction

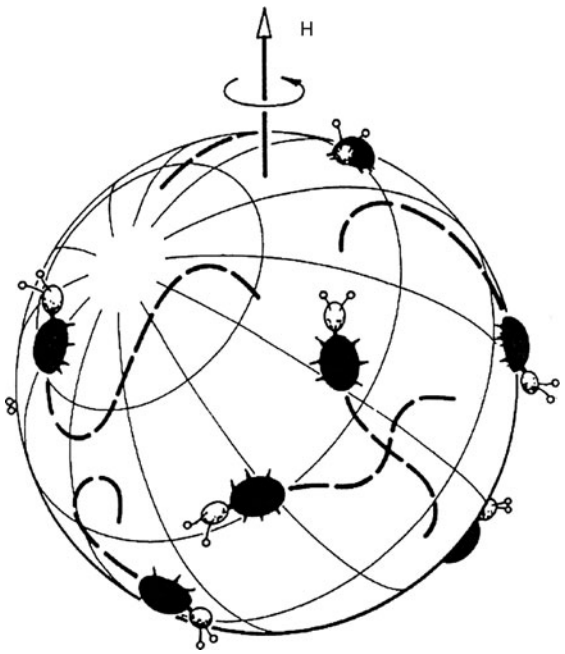
For more than two centuries, geoscientists have suggested, based on geological field observations (reconstruction of paleoclimate belts based on fossils or certain rock types), that the past Earth's equator (and Equatorial conditions) must have at some time been located far from its present position. In the early 1950s, paleomagnetists such as Runcorn (1956) provided quantitative evidence that the instantaneous geographic or rotation pole had moved with respect to certain continents. The paths followed by the poles in the geological past were termed "apparent" polar wander paths (APWPs), because it was not clear whether it was the pole or the continent that had moved. Since then, we have learned that oceanic and continental plates have moved with respect to each other and that a significant part of APW was actually due to these relative motions. A remaining fraction in polar wander, which would be a characteristic of "Earth as a whole" and which would not be accounted for by plate tectonics, is called "true" polar wander (TPW). True polar wander refers thus to the large-scale motions of the Earth's rotation axis through geological time. External forces driving shorter-term changes such as precession are not taken into account here.

True polar wander could result from centrifugal forces acting on mass anomalies distributed on or inside the Earth: excess mass, for instance, will lead to slow deformation of the Earth and change its rotation axis, causing the mass to move towards the equator without being displaced with respect to the solid Earth. The problem is to define precisely in a deformable Earth which part wanders with respect to which reference frame. What is defined (e.g., Steinberger and O'Connell, 2002) is the axis of rotation of a reference frame (RF) relative to an inertial frame of reference (i.e., determined by the angular momentum vector of Earth). The reference frame can be selected as the one that has zero net rotation when motions are integrated over the entire mantle, called the "mean mantle" reference frame. Or one can also define a "mean lithosphere" frame, in which the lithospheric plates have no net rotation. These frames in general do not coincide (Ricard et al., 1991; O'Connell et al., 1991). The "mean mantle" frame can be computed in numerical simulations in which the density distribution is given over the entire mantle. However, if one attempts to constrain it with actual paleomagnetic observations, one has to assume that the "mean mantle" frame is identical to some frame based on observations: the "hotspot reference frame" will most often be the one involved. Steinberger and O'Connell (2000) show that the approximation (discussed later in this paper) is quite good even if hotspots move in the convecting mantle.

The crust and mantle actually deform actively at the typical velocities of plate tectonics (~10–100 mm/year). Thus, polar wander involves (relative) motions of the rotation axis with respect to the Earth's mantle.

Goldreich and Toomre (1969) hypothesized that large-scale polar wandering and continental drift shared a common explanation through mantle convection and the resulting redistribution of masses within the mantle. Goldreich and Toomre (1969) started their analysis of the motions of the rotation axis of a quasi-rigid Earth with the classical figure, where a colony of beetles (the continents or other mass anomalies) moves slowly at the surface of a rigid rotating sphere (Figure 1). They simulated very large polar wander, with particularly intense and rapid swings.

The physics of TPW is explained among others by Ricard et al. (1993a): TPW is seen as the result of conservation of angular momentum of a rotating, deformable body, whose inertia tensor slowly changes with time. The reasons for the changes are mass redistributions linked to mantle convection. In simple terms, the inertia tensor component that affects polar wander is related to subducting slabs, upwelling



Paleomagnetism, Polar Wander, Figure 1 Diagram illustrating polar wander on a quasi-rigid planet (After Goldreich and Toomre, 1969). The beetles represent mass anomalies that change with time (drifting plates or mantle heterogeneities), inducing changes in the principal non-hydrostatic moments of inertia. The coordinate system attached to the (slowly deforming) solid planet moves with respect to the rotation (or spin) axis.

plumes, and more generally thermal and chemical mass heterogeneities resulting from a long history of mantle convection. With estimates of degree 2 components of the geopotential derived from satellite data, it can be shown that the non-hydrostatic part of the Earth's inertia ellipsoid is triaxial, the present rotation axis coinciding with the largest of the non-hydrostatic moments of inertia. The rotation axis tracks the maximum principal inertia axis (PIA) of the non-hydrostatic inertia tensor as fast as adjustment of the hydrostatic bulge allows.

Measure of true polar wander

The position of the rotation axis depends on the time scale. According to astronomers and geodesists, polar wandering is occurring at present. For instance, a determination by Gross and Vondrak (1999) uses the star catalog established from satellites during the past century. The result is a mean drift of the Earth's rotation pole over the period 1900–1992 of ~107 mm/year along the direction of the 79°W meridian.

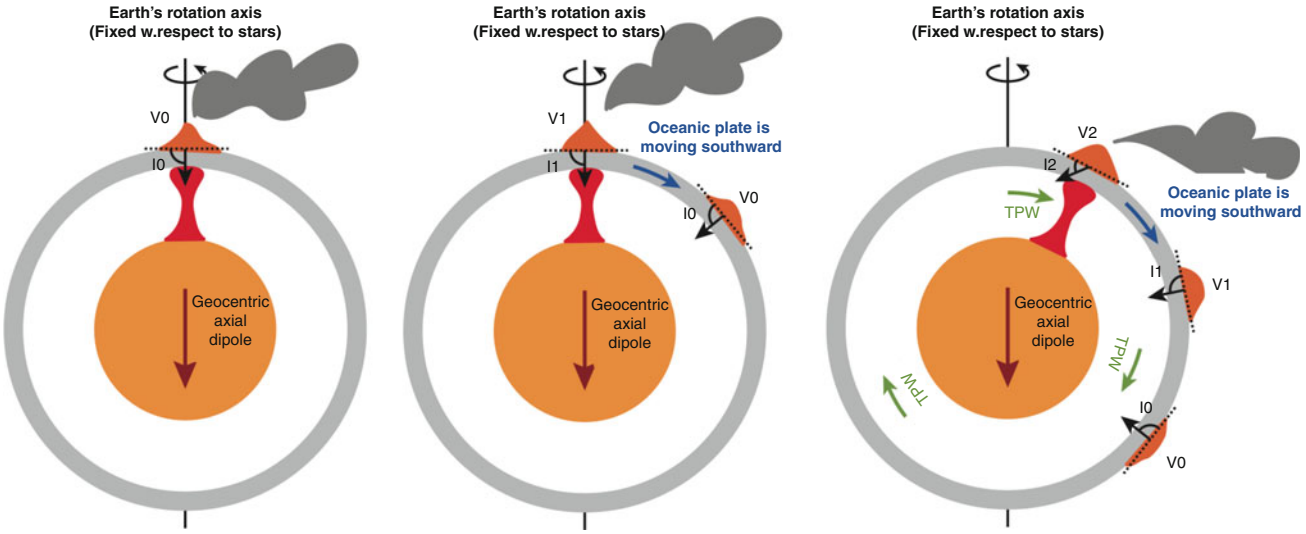
It has also been proposed that this long-term polar wander is largely due to post-glacial rebound, i.e., lithospheric and mantle response to melting of ice (e.g., Sabadini and Peltier, 1981; Mitrovica and Milne, 1998). The pole is moving towards Hudson Bay, where a large ice mass melted and the lithosphere is currently rising.

On time scales longer than 10^6 years, TPW most likely reflects convection of the whole mantle expressed at the surface by horizontal plate motions Steinberger and O'Connell (2002), and paleomagnetism has become the primary source of information. The apparent polar wander paths from all the major plates are being defined with increasing accuracy, and their differences reflect continental drift and plate tectonics. This demonstration of the reality of "apparent" polar wander seen from any plate has led some to propose that the concept of "true" polar wandering might be irrelevant or unnecessary. Are remaining fractions in polar wander, which would be a characteristic of "Earth as a whole," and which would not be accounted for by plate tectonics, actually resolvable? The concept is worthwhile if a reference frame external to the plates themselves can be defined, in which there was significant motion of the pole. Most studies performed in the last 2 decades followed an approach, based on the possibility, originally suggested by Morgan (1972, 1981), that hotspots could provide a valid reference frame for the mantle. Its meaning and existence has been the subject of continued debate, up to this day.

Hotspots are active volcanoes apparently unrelated to plate boundaries, which have been active for tens of millions of years. Hotspots are supposed to be the surface manifestation of plumes, which have left traces in the form of nearly linear chains of extinct volcanoes on the plate(s), which passed over them. Morgan (1972, 1981) proposed that these plumes have a "deep" mantle origin: these convective instabilities may be anchored at least down to the transition zone and possibly all the way down to the core-mantle boundary. As a plate moves over such a plume, the relative motion can be described by a rotation about a pole. If hotspots are indeed fixed with respect to each other and are fixed within the mantle, then hotspot apparent polar wander describes the wander of the Earth's rotation axis with respect to the mantle. It is this motion of the pole with respect to the hotspots, assumed to act as a reference frame for the mantle, which has been taken as an estimate of true polar wander. Determining TPW requires the best possible knowledge of APWPs of all plates (i.e., the best possible data base of paleomagnetic poles to determine the paleolatitudes and orientations of plates with respect to the rotation axis), of their relative motions (determined by marine magnetic anomalies and azimuths of transform faults), and finally the positions and ages of individual volcanoes along hotspot tracks on plates. Paleolatitudes of these seamounts are used for comparison to test model predictions.

APWPs and kinematic models were developed from the 1960s to the 1980s under the key assumption that when averaged over a sufficient amount of time, in excess of a few thousand years, the Earth's magnetic field could be described accurately by a geocentric axial dipole (so-called GAD hypothesis). Besse and Courtillot (1991, 2002), among others, proposed to blend two data sets: the paleomagnetic poles used to derive the APWPs and the oceanic data used to derive plate kinematic models. They used then available paleomagnetic and kinematic models and produced a "synthetic" APWP, which was defined with respect to South Africa and then transferred to all other plates. Motion of hotspots with respect to the plates (Müller et al., 1993) was then integrated to derive an estimate of TPW (Figure 2).

The resulting TPW curve is shown in Figure 3 with a 20 Myr time resolution (each point corresponds to the successive positions with respect to time of a hotspot presently situated at the North geographic pole). Points are shown with their 95% confidence circles every 10 Myr and every other point is statistically independent (estimates at 5 Myr intervals are also available in that study). The first points, corresponding to the period 8–59 Ma, are all in the same quadrant, between 3.3° and 5.9° away from the present rotation pole, with 95% uncertainties ranging from 2° to 3.8°. They are not statistically distinct from each other and therefore could correspond to a standstill. A mean position can be calculated from all data points in that time window: It is found to lie significantly displaced from the pole. The youngest mean pole at ~5 Ma is also significantly displaced from the pole but not statistically different from the 55 Ma overall mean. It seems that TPW may have been negligible for an extended period of about 50 Myr, but accelerated a few million years ago, with a velocity on the order of 100 km/Myr. Prior to that, the path displays a succession of a standstill at 160–130 Ma, a quasi-circular track from 130 to 70 Ma, then the standstill at 50–10 Ma and the



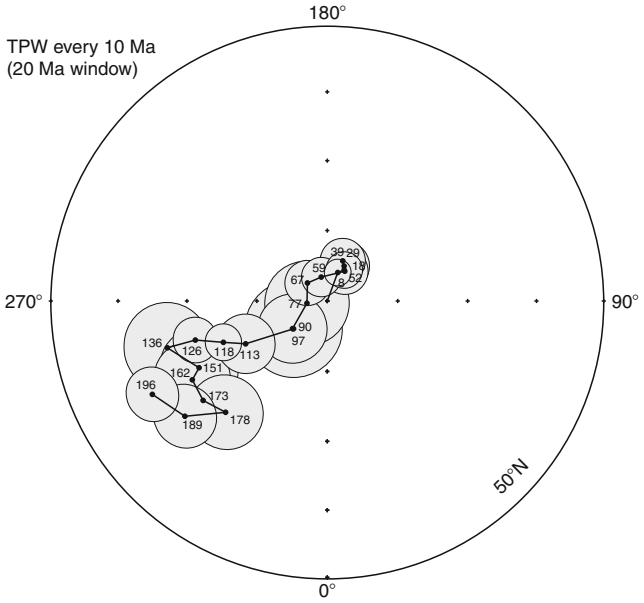
Paleomagnetism, Polar Wander, Figure 2 How mantle TPW is determined: At Time T0, a deep plume (located arbitrarily under the North Geographic Pole) produces a volcano V0 on an oceanic plate. In the hypothesis of a Geocentric Axial Dipole field, a magnetization of inclination I0 = 90° is fossilized in the lavas (the inclination (I) depends on the colatitude (l) such as TgITgθ = 2). Time T1: the oceanic plate has moved southward and a hotspot trace is formed, but the Plume remains at the same position (i.e., no mantle TPW). A new volcano (V1) is fed and records the same 90° inclination (I1) as in the volcano V0. Time T2: an episode of TPW occurs: the plume is tilted southward (with the rest of the mantle). The third volcano V2 will fossilize a lower inclination than V0 and V1. Later on, a paleomagnetist will measure past inclinations in lavas: as lithosphere only moves on the mantle, volcanoes will record the same inclination; the plume remains stationary. In case of TPW, the difference of paleolatitudes deduced from measured inclination will indicate the amount of motion.

faster motion up to the present. The true polar wander rate between 130 and 70 Ma averages 30 km/Myr. True polar wander appears episodic in nature, with periods of (quasi) standstill alternating with periods of faster TPW. The typical duration of these standstill periods is on the order of a few tens of millions of years (50 Myr). Because of all the uncertainties in models of hotspot kinematics prior to ~130 Ma, it is not safe to place too much weight on behavior prior to that time. The major event since then is therefore the end of the 130–60 Ma period of relatively fast polar wander, with a standstill (i.e., no or little TPW) from 50 Ma (actually because of larger 95% confidence circles, possibly 80 Ma) to 10 Ma.

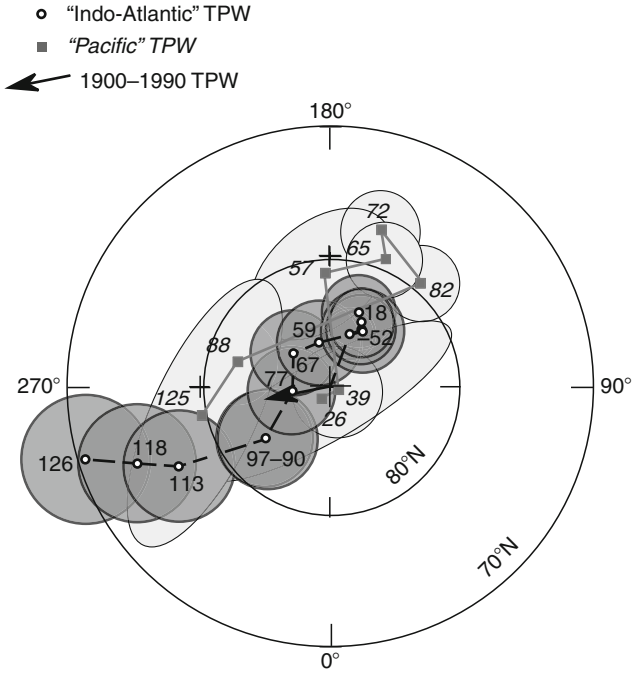
However uncertain, evidence for the fact that Earth emerged from a long standstill to enter a new period of faster polar wander in a different direction 10 Ma ago (~5 Ma at the higher resolution) is particularly interesting. That period would then still be going on. The youngest pole at ~5 Ma, taken at face value, would imply a TPW rate of 130 km/Myr in the direction ~15°W (with a large azimuthal uncertainty from 10°E to 40°W, due to proximity of the 5 Ma pole to the present rotation axis). The reasonable agreement with recent astronomical determinations (107 km/Myr in the direction 79°W, with a possible range from 65°W to 85°W) has reinforced the idea that this might be a valid estimate and that TPW direction and rate might have been rather stable over the last 5 million years corresponding to significant, large-scale changes in ice cover. Ice is not considered to have played a significant role before. It was absent prior to ~30 Ma.

A legitimate concern regarding the above conclusions is due to the fact that the analysis is not truly global, in that it fails to encompass the Pacific plate. Petronotis and Gordon (1999) have compiled an APW path (nine poles from 125 to 26 Ma) for the Pacific plate, based on the analysis of skewness of ocean crust magnetic anomalies and on seamount magnetic anomaly modeling. Using the Pacific plate versus hotspot kinematic model of Engebretson et al. (1985), based on dating of volcanoes that are part of hotspot tracks such as Hawaii, Louisville, or MacDonald, Besse and Courtillot (2002) have determined a corresponding 125–26 Ma "Pacific hotspot only" TPW curve. The "Pacific" and "Indo-Atlantic" TPW curves are compared in Figure 4. This comparison is interesting because the data sets they are based on are entirely different and independent.

Despite some significant differences, it is worth emphasizing that the two curves are altogether similar in shape (tracks, amplitudes, azimuths), particularly the 300° longitude trending track from 130 to 70 Ma, though the two are offset (in the same general direction) by about 7°. More precisely, the confidence intervals intersect (though means are not in the intersection) near 125, 90, 60, 40, and 30 Ma. The main differences occur between 80 and 70 Ma: the 82



Paleomagnetism, Polar Wander, Figure 3 True polar wander paths deduced from the hotspot model of Müller et al. (1993) going from the Present back to 130 Ma, and that of Morgan (1981) going from 130 back to 200 Ma, with associated 95% confidence ellipses (shaded *light gray*); 20 Ma sliding window. Ages are in Ma and correspond to the actual mean age of the data in the corresponding window, which is why they are not regularly distributed when the original time windows were. Note that the external circle is at 50° latitude (After Besse and Courtillot, 2002).



Paleomagnetism, Polar Wander, Figure 4 Comparison between the "Indo-Atlantic" (*open dots* and *dashed line*) and "Pacific" (*squares* and *solid line*) TPW Paths. The *arrow* shows the mean pole motion between 1900 and 1990 (After Besse and Courtillot, 2002).

and 65–72 Ma poles derived from Petronotis and Gordon (1999) and the 77 and 67 Ma poles derived by Besse and Courtillot (2002) are clearly distinct. Tarduno and Cottrell (1997) and Tarduno et al. (2003), however, expressed doubts on the reliability of APWPs for an oceanic plate, such as the Pacific, when they are based solely from inversions of magnetic surveys over seamounts and/or analysis of skewness of marine magnetic profiles. They have determined the paleolatitude (based on inclination-only data from cores) of the 81 Ma old Detroit Seamount, which is part of the Emperor chain, and noticed that the paleolatitude is distinct from that based on the 81 Ma pole of the Pacific APWP (Sager and Pringle, 1988) based on seamount anomaly poles. Tarduno and Cottrell (1997) exclude the possibilities of inadequate sampling of secular variation, bias due to unremoved overprints or off-vertical drilling. Di Venere and Kent (1999) and Cottrell and Tarduno (2000) argue that the reliability of the Pacific paleopoles based on either modeling of seamount magnetic anomalies or determination of skewness of marine magnetic anomalies should be considered suspect. Both are prone to numerous biases and could yield errors in excess of 10° in the position of the mean poles derived from them. This debate is not closed.

Is a reference frame based on hotspots valid?

Another question discussed in the last decade is Pacific hotspot fixity with respect to one another and to the "Indo-Atlantic" hotspots. Norton (1995) has suggested that the famous 43 Ma Hawaiian bend in the Hawaiian–Emperor hotspot track was actually a "nonevent," i.e., indicated a change in motion of the Hawaiian hotspot with respect to the mantle rather than a change in Pacific plate motion. Koppers et al. (2001) find that the 0–43 Ma Hawaiian and Foundation seamount trail pair is the only one compatible with the fixed hotspot hypothesis. The 43–80 Ma Emperor/Line pair shows particularly large discrepancies, requiring motions of at least 30 km/Myr. Koppers et al. (2004) studied 40Ar/39Ar age progressions of the Hawaiian and Louisville hotspot tracks. They showed that the progression is nonlinear and can be better fitted with a moving hotspot model rather than a fixed hotspot frame of reference. The primary Hawaiian and Louisville hotspots did not move in concert, showing significant inter-hotspot motion between 80 and 47 Ma. While the Hawaiian hotspot experienced a substantial 15° southward motion prior to 47 Ma (Tarduno et al., 2003), the Louisville hotspot is predicted to have experienced an eastward motion of 5° during the same geological period, followed by a minor 2° latitudinal shift over the last 30 Myr. Clouard and Bonneville (2001) showed

that Pacific seamounts are actually created by different processes, many being short-lived. Only the Hawaii and Louisville chains qualify as long-lived hotspots that can robustly be tested for fixity. Their preferred explanation for this is southward drift of Atlantic hotspots prior to ~95 Ma. Courtillot et al. (2003) reanalyzed the characteristics of the world catalog of some 50 hotspots. They suggest that surface hotspots on Earth may have three distinct origins with only seven "primary" hotspots (Hawaii, Easter, Louisville, Iceland, Afar, Reunion, and Tristan da Cunha) originating from the deepest part of the lower mantle, possibly anchored on chemical heterogeneities deposited in the D″ layer. Courtillot et al. (2003) found no evidence for inter-hotspot motion significantly larger than 5 km/Myr either within the Pacific hemisphere or the Indo-Atlantic hemisphere hotspots. Such rms velocities of 5 km/Myr or less, i.e., an order of magnitude less than rms plate velocities, are to first order "small." So, some primary hotspots indeed seem to provide a quasi-fixed frame in each hemisphere over the past 80–100 Myr.

Was there then any motion between the two hotspot ensembles? This raises the well-known difficulty of establishing a reliable kinematic connection between the two hemispheres through Antarctica. This has been addressed by Raymond et al. (2000). Based on updated kinematics (assuming that there is no missing plate boundary or unaccounted for motion between East and West Antarctica), these authors predict the location of the Hawaiian hotspot back in time, under the hypothesis that Reunion and Hawaii have remained fixed with respect to each other; for this, they use the dated tracks left on the African and Indian plates by the Reunion hotspot since it started as the Deccan traps 65 Ma ago. The plot of misfit between the predicted and observed positions for Hawaii as a function of time indicates that the two hotspots have actually drifted slowly, at ~10 km/Myr, for the last 45 Myr, but at a much faster rate (~50 km/Myr) prior to that, vindicating conclusions of Pacific hotspot motion reached by Tarduno and Cottrell (1997) or Tarduno (2007). Courtillot et al. (2003) conclude that the primary hotspots may form two distinct subsets in each one of the two geodynamically distinct hemispheres. Each subset would deform an order of magnitude slower than typical plate velocities. The two subsets would have been in slow motion for the last 45 Myr, but in faster motion prior to that.

Using their paleomagnetic results from Marie Byrd Land in West Antarctica, Di Venere and Kent (1999) demonstrated that some motion must have taken place since 100 Ma between West (Marie Byrd Land) and East Antarctica; they concluded that this motion cannot account for more than 20%, and possibly as little as 4% of the 14.5° offset between the observed and predicted positions of the 65 Ma Suiko seamount on the Emperor continuation of the Hawaiian hotspot track. They also discussed the integrity of the Pacific plate and the role of missing plate boundaries and errors in kinematic plate circuits, and find that they play a small role. Di Venere and Kent (1999) concluded that most of the apparent motion between the two main groups of hotspots is real, with an average drift of about 25 km/Myr since 65 Ma. Steinberger et al. (2004) have combined hotspot motion in the large-scale flow field of the mantle and intraplate deformation to account for most of the post-65 Ma misfits between computed and observed hotspot tracks. Only a combination of the two effects (hotspot motion and intraplate deformation) successfully predicts the geometry of the four selected seamount chains in the Pacific and Indo-Atlantic hemispheres. If correct, this analysis at the same time invalidates the notion of a reference frame or of separate hemispheres and replaces it with limited, predictable, inter-hotspot motion. But TPW can still be calculated. The Hawaiian hotspot would have moved faster between 80 and 50 Ma, and at the same time TPW would have been faster prior to 50 Ma. TPW alone would account for paleolatitudinal changes of the Reunion hotspot, whereas the two effects would be required for the Hawaiian hotspot.

The estimate of true polar wander (TPW) over the last 200 Myr is thus still a matter of strong debates, as recalled above, being based only on hotspots from the Indo-Atlantic hemisphere. The TPW curve estimated using Pacific-only data is similar to the Indo-Atlantic TPW, seemingly validating to first order the concept that TPW is a global phenomenon. On the other hand, Besse and Courtillot (2002) suggest that TPW pole positions for the two hemispheres are significantly displaced between ~50 and ~90 Ma. Hotspot tracks become fewer and more uncertain as one goes back in the past and pre-130 Ma TPW estimates should be regarded with caution. However, there are strong indications (Besse and Courtillot, 2002) that the phase of true polar wander, which ended around 50 Ma, may have started ~130 Ma ago.

Modeling of true polar wander

In an early attempt, Ricard et al. (1993b) and Richards et al. (1997) derived a model of mantle density heterogeneities based on plate motions recorded by the ocean floor, including slabs crossing the upper/lower mantle transition zone slowing down by a factor between 2 and 5. Based on this, they computed a synthetic geoid, which agreed rather well with observations, provided viscosity increased by a factor 40 going from the upper to the lower mantle. Changes in the degree 2 geoid should reflect the history of the inertia tensor and therefore TPW as revealed by paleomagnetism. They predicted TPW velocities on the order of 300 km/Myr in the direction 130°W, much larger than observed rates. Their three-layer model could not account simultaneously for the present-day geoid and observed TPW.

In other studies, seismic tomography has been used to infer 3D maps of density heterogeneities that drive flow in the viscous mantle. Steinberger and O'Connell (1997, 2002) have developed an algorithm to calculate changes in the Earth's rotation axis for rather general viscoelastic Earth models. Their model combines hotspot motion, polar wander, and plate motions in a single dynamically consistent model, allowing all effects to be simultaneously

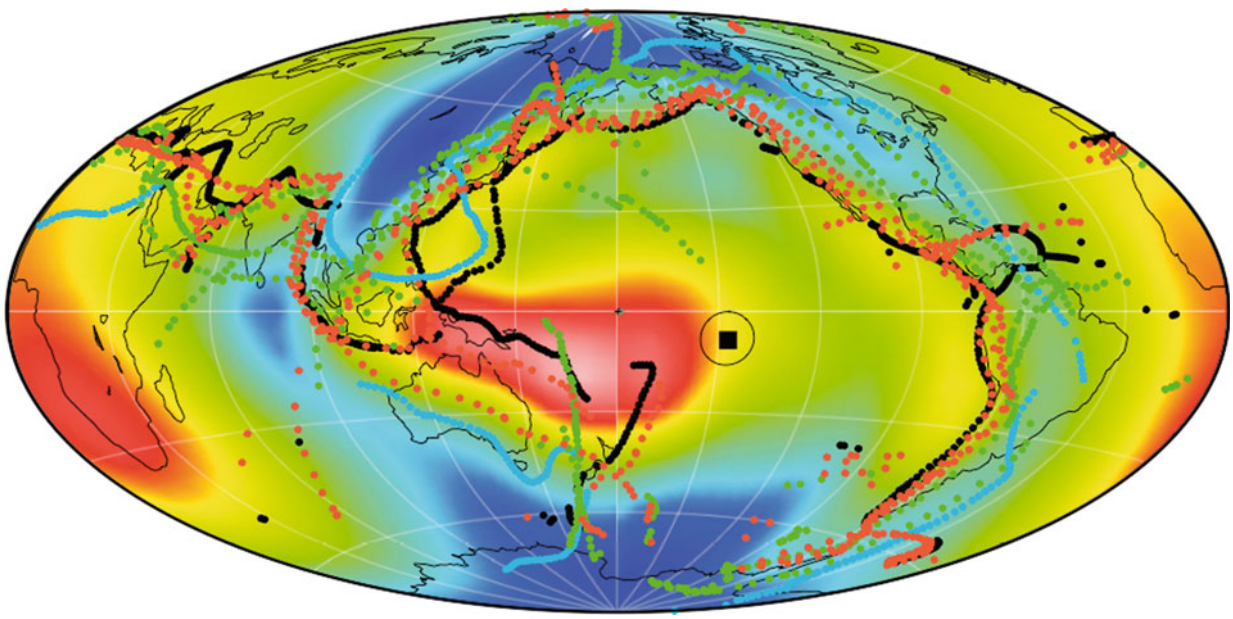
determined and their relative effects sorted out. Changes in rotation are due to advection of "realistic" density anomaly distributions inferred from a number of seismic tomography models in a "realistic" flow field constrained by observed plate kinematics at the surface. Steinberger and O'Connell (2002) compare their model results for a number of tomographic models and discuss the effects of compressibility, deep mantle viscosity, thickness of a high viscosity layer in the deep mantle, and chemical boundaries on rates of TPW. Their TPW estimates are in qualitative agreement with observations for the last 100 Ma. Steinberger and Torsvik (2010) modeled the present-day non-hydrostatic inertia tensor by combining contributions due to large low-shear velocity provinces in the lowermost mantle and due to subduction and found a spin axis at 67°N, 96°E (north Siberia). The distribution of recent subduction, with largest amounts in the northwest Pacific (beneath East Asia) and the southeast Pacific (beneath South America), adds a secondary contribution that moves the spin axis toward the observed poles.

Rouby et al. (2010) also investigated the effect of internal mass redistributions on the position of the Earth's rotational pole for the last 120 Myr. They also used a geodynamic model based on plate reconstructions and estimated the location and rate of subducted slabs under the assumption that they sink vertically into the mantle (Figure 5). The new model takes into account the effect of large-scale mantle upwellings derived from tomography (one is situated under Africa, the other, nearly antipodal under the Pacific Ocean). The temporal variation of the mantle density heterogeneities was essentially found to reflect changes in subduction history. Because the tomographic images seem to show that subducted plates do not systematically reach the CMB, they investigated the influence of the depth down to which these plates retain significant density contrast with respect to the surrounding mantle on the present-day geoid and on the temporal evolution of the Earth's PIA (we recall the rotational axis is aligned to the maximum PIA). They proposed a model, which yields to a present-day geoid in good agreement with the observed one. The associated minimum principal inertia axis seems to be relatively stable since 120 Ma and close to Africa, within the upwelling (Figure 6). Their model displays that some of the features of the computed TPW, which are directly linked to the temporal evolution of the subductions, are reminiscent of observed phases of slow and fast true polar wander, with similar peak velocities. Although their computed TPW is almost in the same meridian as the observed one, the poles are still drifting in the opposite direction, as was already the case for Ricard et al. (1997).

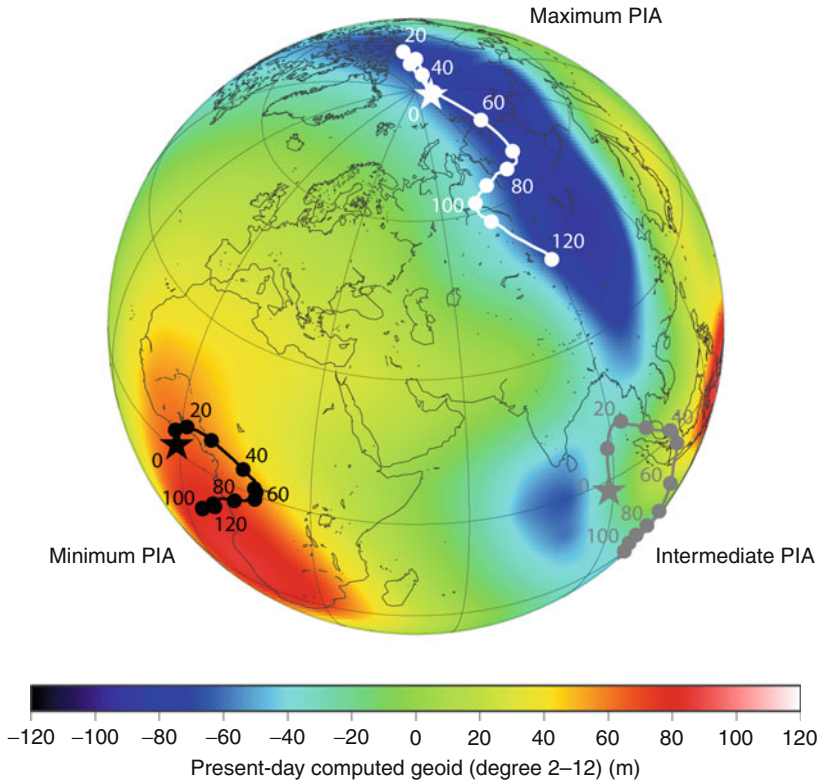
An integrated explanation for polar wander?

Let us now put this study in the broader context of the history of the Earth's rotation.

Paleomagnetic indications that would support rather fast true polar wander in pre-Mesozoic time have been put forward by Van der Voo (1994; with rates between 70 and 110 km/Myr) and for very fast true polar wander by Kirschvink et al. (1997; with rates in excess of



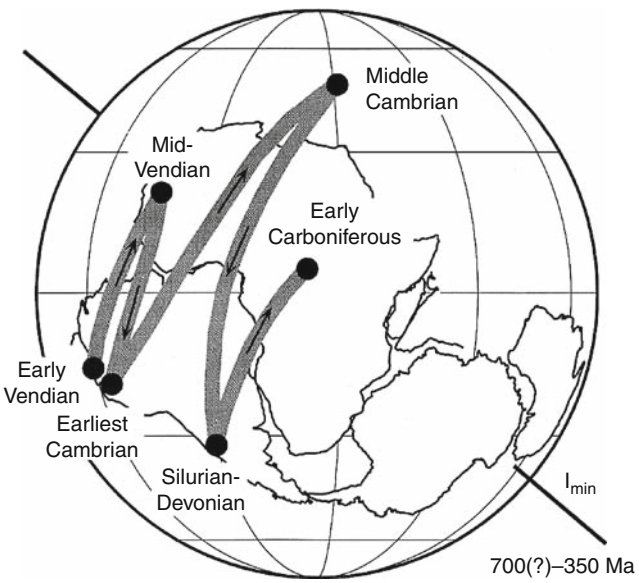
Paleomagnetism, Polar Wander, Figure 5 Subduction zones at the Earth surface, every 40 Myr, since 200 Myr, superimposed to the present-day final computed geoid (degree 2–12). The *square* represents the mean center of small circles fitting the subductions zones (After Rouby et al., 2010).



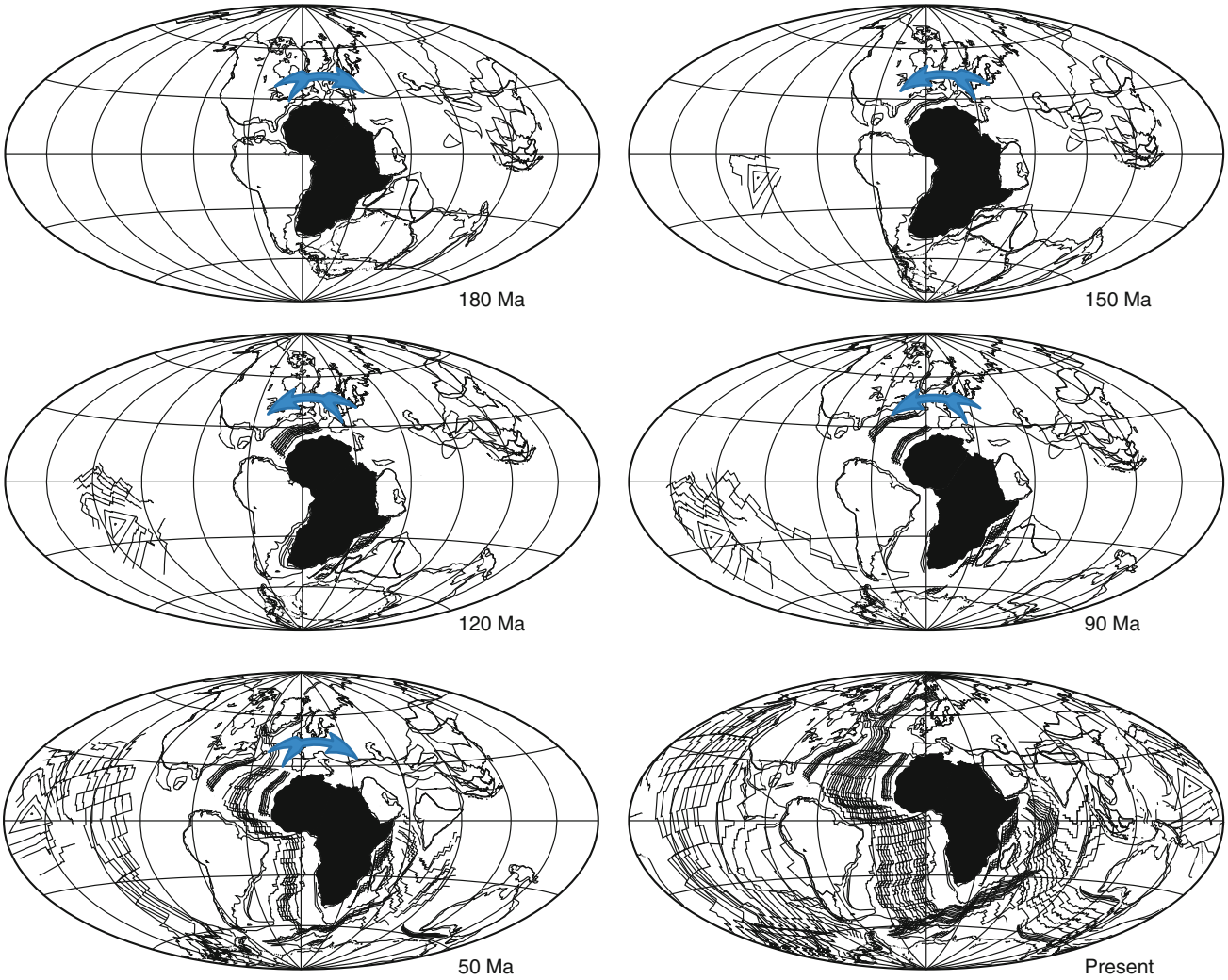
Paleomagnetism, Polar Wander, Figure 6 Principal Inertia Axes since 120 Myr (dots every 10 Myr, stars denote present-day values) for our mantle density heterogeneities model. The corresponding present-day geoid (degree 2–12) is also shown (After Rouby et al., 2010).

600 km/Myr in the early Cambrian). Recall that Steinberger and O'Connell's (2002) numerical models place an upper limit of about 100 km/Myr on TPW velocity. For pre-Jurassic time, ocean floor has been destroyed (together with evidence of plate kinematics and most hotspot traces) and TPW cannot be estimated in the way indicated above. Evans (2003) proposes to obtain a crude estimate of TPW in the case when TPW velocity would have been much faster than between-plate motions. Kirschvink et al. (1997) and Evans (2003) discussed the fact that the Earth may have in the past catastrophically exchanged two of its principle axes of inertia (inertial interchange true polar wander or IITPW), i.e., the whole lithosphere (and mantle) may have rotated by some 90° in only a few million years, bringing for instance parts of the Earth that were close to the poles at the Equator. Evans (2003) interpreted the large rotations as TPW about a common, long-lived, minimum inertial axis (Imin) near eastern Australia, which may also be regarded as the prolate axis of Earth's figure, inherited from Rodinia (Figure 7).

Satolli et al. (2007) on the basis of a high definition magnetostratigraphy in the Apennines argued that a major trend of the "mean" plate motion since at least 200 Ma can be described as large successive clockwise or counter-



Paleomagnetism, Polar Wander, Figure 7 Terminal Proterozoic to late Paleozoic apparent polar wander path for Gondwanaland. Oscillatory APW rotations are interpreted as TPW about a common, long-lived, minimum inertial axis (near eastern Australia (After Evans, 2003).



Paleomagnetism, Polar Wander, Figure 8 Global continent and ocean reconstructions, from 180 Myr to present-day, based on techniques and APWPs described in Besse and Courtillot (2002). Magnetic anomalies are from Royer et al. (1992), the 150 and 130 Myr frames from Satolli et al. (2007). A major trend of the "mean" plate motion since at least 200 Myr can be described as successive clockwise or counter-clockwise rotations (linked to TPW), and roughly centered in Africa (close to the minimum PIA, see Figure 8). The blue *arrow* indicates the direction of rotation until the next reconstruction (After Rouby et al., 2010).

clockwise rotations linked to TPW, and roughly centered on Africa, this last plate being a "central plate" from which most continents have diverged since the break-up of Pangea (Figure 8). An animated sequence derived from Besse and Courtillot (1991) APWP displays visually those features available at http://www.ipgp.fr/~besse/pw. Steinberger and Torsvik (2008) have also found that the global axis of rotation for plates since 250 Ma also roughly centered in Africa, and argued that they may also correspond to maxima of the degree 2 non-hydrostatic geoid, corresponding to the mantle upwellings.

Le Pichon and Huchon (1984) already remarked that during Late Triassic about 200 Ma ago, the outlines of Pangea lay along a great circle passing through the paleopoles of rotation and also noticed that the present geoid also displays hemispheric symmetry with an axis located in the equatorial plane. The stability of subduction zones back to the middle of the Paleozoic in a nearly N-S plane was also recognized by Collins (2003). All these studies pointed out a strong link between pole positions and the geometry of the convection: this coincidence might be explained by a deep-seated, quadrupolar convection as was first advocated by Busse (1983). Rouby et al. (2010) addressed this question by computing a present geoid day (degree 2–12) using the plate model described above (Figures 5 and 6).

The persistent location of subductions in the western Tethyan and Pacific regions for at least 250 Ma (and

possibly earlier according to Collins) have generated a ring of downwelling material inducing hot return flows (the upwellings mentioned above) confined to two main regions (presently Indo-Atlantic and Pacific). This geometry defines the location of PIAs and pins the minimum PIA in the region of upwelling. This large zone is roughly centered on the equator also close to the maxima of the degree 2 non-hydrostatic geoid. The minimum PIA has been relatively stable since 120 Ma and close to Africa. On the contrary, both the maximum and intermediate PIAs have moved in a plane perpendicular to Africa along a meridian corresponding to the large-scale pattern low of the geoid. This meridian seems to be stable with time, similar to the large-scale pattern of subduction around the Pacific and western Tethyan zone of subductions.

The temporal evolution of the rotation axis is that of the maximum PIA. Oscillatory polar wander is attributed to TPW about a common, long-lived minimum inertial axis, which would be the prolate axis of Earth figure, inherited from the past plates/continent configurations and following the quadrupolar convection model of Busse (1983).

This peculiar pattern may constitute a long-lived feature of our planet, at least since 200–250 Ma, which forces continents to be roughly in an EW position, on a single hemisphere bounded by a plane containing the rotation axis, with the Earth's rotational pole not far from the main ring of subduction zones. TPW (and possibly APW) may occur around a nearly equatorial euler pole (axis) of rotation, driven by slight changes of magnitude between the maximum and intermediate PIAs, principally caused by subduction locations and rate changes with time. This view is coherent with the successive CW and CCW rotations of all continents (including the Pacific Ocean) observed by Evans (2003), Satolli et al. (2007), and Steinberger and Torsvik (2008) around a point located somewhere in Africa and close to the equator for these last two authors.

Conclusions

For the time being, the hypothesis that TPW is a go-stop-go phenomenon, with long standstills in the tens of millions of years, and periods of faster, rather uniform polar wander at velocities often not exceeding 30 km/Myr, cannot be discounted. Evidences for superfast events at any time in the past remain to be demonstrated.

The similarity between TPW estimates for the Pacific plate and the rest of the world, which are based on completely different and independent data sets, may be taken as support to the idea that small yet significant TPW, on the order of 10°, occurred since the Cretaceous, though it still requires careful verification.

Most current modeling still fails to some extent to account for the slow values of typical TPW velocity (0–30, and rarely up to 100 km/Myr), and even more so to account for the prolonged (~50 Myr) periods with almost no TPW (standstills). Steinberger and O'Connell (2002) give reasonable velocities but predict smooth evolutions, rather than the alternating episodes, revealed by paleomagnetic data. Rouby et al.'s (2010) model restores the present geoid, displays some of the features of computed TPW that are directly linked to the temporal evolution of subductions, reminiscent of observed phases of slow and fast true polar wander, with similar peak velocities, but the poles are still drifting in the wrong direction, as was also the case for Ricard et al. (1997).

TPW seems to be linked to a pattern of quadrupolar convection permanent (at least on periods of time in excess of 200 Ma). If the two upwellings remain more or less in the same locations, the pole may move on the great circle perpendicular to the axis they form, the system being stabilized by subductions not far from the pole. These concepts remain to be tested for earlier periods of time.

The importance of the Pacific plate and severe limitations on presently available data from that plate, the fact that the global paleomagnetic database still is quite scant for certain continents and time windows, the fact that many hotspot traces are not yet sufficiently constrained in terms of age and paleolatitude, all point to the need for many more direct (paleomagnetic) measurements as opposed to indirect/remote sensing determinations of magnetization direction (i.e., "skewness" or "seamount" data).

The reason for the discrepancies between models and data could be due to the quality of reconstruction of the Pacific Ocean (specially before 130 Ma) and its relationship with bordering continents, and hence, the poor knowledge of the exact amount of subduction at a given time (to which these kind models are so sensitive). Some of this information is unfortunately lost forever, since large parts of the Pacific Ocean have been subducted.

Future research should aim to understand the discrepancies between observation and computations in order to estimate possible uncertainties on the associated mantle mass anomalies by testing other reconstructions. It will be particularly useful to build a single reference frame for both Indo-Atlantic and Pacific hemispheres: this implies that one must be able to use Antarctica to link the two hemispheres. This will not be easy, given the poor knowledge of deformations suffered by the continent and the possibility of existence of active margins on its Pacific border.

Bibliography

Besse, J., and Courtillot, V., 1991. Revised and synthetic polar wander paths of the African, Eurasian, North American and Indian plates, and true polar wander since 200 Ma. *Journal of Geophysical Research*, **96**, 4029–4050.

Besse, J., and Courtillot, V., 2002. Apparent and true polar wander and the geometry of the geomagnetic field in the last 200 million years. *Journal of Geophysical Research*, **107**(2300), 31, doi:10.1029/2000JB000050.

Busse, F. H., 1983. Quadrupole convection in the lower mantle? *Geophysical Research Letters*, **10**(4), 285–288.

Clouard, V., and Bonneville, A., 2001. How many Pacific hotspots are fed by deep-mantle plumes? *Geology*, **21**, 695–698.

Collins, W. J., 2003. Slab pull, mantle convection, and Pangaean assembly and dispersal. *Earth and Planetary Science Letters*, **205**, 225–237.

Cottrell, R. D., and Tarduno, J. A., 2000. Late cretaceous true polar wander: not so fast. *Science*, **288**, 2283a.

Courtillot, V., Davaille, A., Besse, J., and Stock, J., 2003. Three distinct types of hotspots in the Earth's mantle. *Earth and Planetary Science Letters*, **205**, 295–308.

Di Venere, V., and Kent, D. V., 1999. Are the Pacific and Indo-Atlantic hotspots fixed? Testing the plate circuit through Antarctica. *Earth and Planetary Science Letters*, **170**, 105–117.

Engebretson, D. C., Cox, A., and Gordon, R. G., 1985. Relative motions between oceanic and continental plates in the Pacific Basin. *Geological Society of America Special Paper* 206, 59 pp.

Evans, D. A. D., 2003. True polar wander and supercontinents. *Tectonophysics*, **362**, 303–320.

Goldreich, P., and Toomre, A., 1969. Some remarks on polar wandering. *Journal of Geophysical Research*, **74**, 2555–2567.

Gordon, R. G., 1987. Polar wandering and paleomagnetism. *Annual Review of Earth and Planetary Sciences*, **15**, 567–593.

Gordon, R. G., Horner-Johnson, B. C., Petronotis, K., and Acton, G. D., 2004. Apparent polar wander of the Pacific plate and Pacific hotspots: implications for true polar wander and hotspot fixity. *EOS Transactions, AGU*, Spring meeting, Abstract #U52A-04.

Gross, R. S., and Vondrak, J., 1999. Astrometric and space-geodetic observations of polar wander. *Geophysical Research Letters*, **26**, 2085–2088.

Kirschvink, J. L., Ripperdan, R. L., and Evans, D. A., 1997. Evidence for a large-scale early Cambrian reorganization of continental masses by inertial interchange true polar wander. *Science*, **277**, 541–545.

Koppers, A. A. P., Morgan, J. P., Morgan, J. W., and Staudigel, H., 2001. Testing the fixed hotspot hypothesis using Ar-40/ Ar-39 age progressions along seamount trails. *Earth and Planetary Science Letters*, **185**(3–4), 237–252.

Koppers, A., Duncan, R., and Steinberger, B., 2004. Implications of a nonlinear 40Ar/39Ar age progression along the Louisville seamount trail for models of fixed and moving hot spots. *Geochemistry, Geophysics, Geosystems*, **5**(6), Q06L02, doi:10.1029/2003GC000671.

Le Pichon, X., and Huchon, P., 1984. Geoid, Pangea and convection. *Earth and Planetary Science Letters*, **67**, 123–135.

Mitrovica, J. X., and Milne, G. A., 1998. Glaciation-induced perturbations in the Earth's rotation: a new appraisal. *Journal of Geophysical Research*, **103**, 985–1005.

Morgan, W. J., 1972. Plate motions and deep mantle convection. *Geological Society of America Memorials*, **132**, 7–22.

Morgan, J. W., 1981. Hotspot tracks and the opening of the Atlantic and Indian oceans. In Emiliani, C. (ed.), *The Sea*. New York: Wiley, pp. 443–487.

Müller, D. M., Royer, J. Y., and Lawver, L. A., 1993. Revised plate motions relative to the hotspots from combined Atlantic and Indian Ocean hotspot tracks. *Geology*, **21**(3), 275–278.

Norton, I. O., 1995. Plate motion in the North Pacific: the 43 Ma nonevent. *Tectonics*, **14**, 1080–1094.

O'Connell, R. J., Gable, C. W., and Hager, B. H., 1991. Toroidal–poloidal partitioning of lithospheric plate motions. In Sabadini, R., and Lambeck, K. (eds.), *Glacial Isostasy, Sea Level and Mantle Rheology*. Dordrecht: Kluwer, pp. 535–551.

Petronotis, K. E., and Gordon, R. G., 1999. A Maastrichtian paleomagnetic pole for the Pacific plate from a skewness analysis of marine magnetic anomaly 32. *Geophysical Journal International*, **139**, 227–247.

Raymond, C. A., Stock, J. M., and Cande, S. C., 2000. Fast Paleogene motion of the Pacific hotspots from revised global plate circuit constraints. In Richards, M., Gordon, R., and van der Hilst, R. (eds.), *The History and Dynamics of Global Plate Motions*. Washington, DC: American Geophysical Union. Geophysical Monograph, Vol. 121, pp. 359–375.

Ricard, Y., Doglioni, C., and Sabadini, R., 1991. Differential rotation between lithosphere and mantle: a consequence of lateral mantle viscosity variations. *Journal of Geophysical Research*, **96**, 8407–8416.

Ricard, Y., Spada, G., and Sabadini, R., 1993a. Polar wandering of a dynamic Earth. *Geophysical Journal International*, **113**, 284–298.

Ricard, Y., Richards, M., Lithgow-Bertelloni, C., and Le Stunff, Y., 1993b. A geodynamic model of mantle density heterogeneity. *Journal of Geophysical Research*, **98**, 21895–21909.

Richards, M. A., Ricard, Y., Lithgow-Bertelloni, C., Spada, G., and Sabadini, R., 1997. An explanation of the Earth's long-term rotational stability. *Science*, **275**, 372–375.

Rouby, H., Greff, M., and Besse, J., 2010. Mantle dynamics, geoid, inertia and TPW since 120 Myr. *Earth and Planetary Science Letters*, doi:10.1016/j.epsl.2010.01.033.

Runcorn, S. K., 1956. Paleomagnetic comparisons between Europe and North America. *Proceedings. Geological Association of Canada*, **8**, 77–85.

Royer, J. -Y., Müller, R. D., Gahagan, L. M., Lawver, L. A., Mayes, C. L., Nürnberg, D., and Sclater, J. G., 1992. A global isochron chart. Univ. of Texas Inst. for Geophysics Tech. Rep., 117.

Sabadini, R., and Peltier, W. R., 1981. Pleistocenic deglaciation and the Earth's rotation: implications for mantle viscosity. *Geophysical Journal of the Royal Astronomical Society*, **66**, 553–578.

Sager, W. W., and Pringle, S., 1988. Mid cretaceous to early tertiary apparent polar wander path of the Pacific plate. *Journal of Geophysical Research*, **93**, 11753–11771.

Satolli, S., Besse, J., Speranza, F., and Calamita, F., 2007. The 125–150 Ma high-resolution Apparent Polar Wander Path for Adria from magnetostratigraphic sections in Umbria–Marche (Northern Apennines, Italy): Timing and duration of the global Jurassic–Cretaceous hairpin turn. *Earth and Planetary Science Letters*, **257**, 329–342.

Spada, G., Ricard, Y., and Sabadini, R., 1992. True polar wander for a dynamic Earth. *Nature*, **360**, 452–454.

Steinberger, B., 2000. Plumes in a convecting mantle: models and observations for individual hotspots. *Journal of Geophysical Research*, **105**, 11127–11152.

Steinberger, B. M., and O'Connell, R. J., 1997. Changes of the Earth's rotation axis inferred from advection of mantle density heterogeneities. *Nature*, **387**, 169–173.

Steinberger, B., and O'Connell, R. J., 2000. Effects of mantle flow on hotspot motions. In Richards, M., Gordon, R., and van der Hilst, R. (eds.), *The History and Dynamics of Global Plate Motions*. Washington, DC: American Geophysical Union. Geophysical Monograph, Vol. 121, pp. 377–398.

Steinberger, B., and O'Connell, R. J., 2002. The convective mantle flow signal in rates of true polar wander. In Mitrovica, J., and Vermeersen, L. (eds.), *Ice Sheets, Sea-Level and the Dynamic Earth*. Washington, DC: American Geophysical Union. Geodynamics Series, Vol. 29, pp. 233–256.

Steinberger, B., and Torsvik, T. H., 2008. Absolute plate motions and true polar wander in the absence of hotspot tracks. *Nature*, **452**(7187), 620.

Steinberger, B., and Torsvik, T. H., 2010. Toward an explanation for the present and past locations of the poles. *Geochemistry, Geophysics, Geosystems*, **11**, Q06W06, doi:10.1029/2009GC002889.

Steinberger, B., Sutherland, S., and O'Connell, R., 2004. Prediction of Emperor- Hawaii seamount locations from a revised model of plate motion and mantle flow. *Nature*, **430**, 167–173.

Tarduno, J. A., and Cottrell, R. D., 1997. Paleomagnetic evidence for motion of the Hawaiian hotspot during formation of the Emperor Seamounts. *Earth and Planetary Science Letters*, **153**, 171–180.

Tarduno, J. A., and Gee, J., 1995. Large-scale motion between Pacific and Atlantic hotspots. *Nature*, **378**, 477–480.

Tarduno, J. A., Duncan, R. A., Scholl, D. W., Cottrell, R., Steinberger, B., et al., 2003. The Emperor seamounts: southward motion of the Hawaiian hotspot plume in Earth's mantle. *Science*, **301**, 1064–1069.
Tarduno, J.A., 2007. On the motion of Hawaii and other mantle plumes. *Chemical Geology*, **241**, 234–247.
Van der Voo, R., 1994. True polar wander during the mid-Paleozoic? *Earth and Planetary Science Letters*, **122**, 239–243.

Cross-references

Geomagnetic Excursions
Geomagnetic Field, Polarity Reversals
Geomagnetic Field, Secular Variation
Geomagnetic Field, Theory
Paleomagnetism, Magnetostratigraphy
Paleomagnetism, Principles

PALEOMAGNETISM, PRINCIPLES

William Lowrie
Emeritus Professor of Geophysics, Institute of Geophysics, Swiss Federal Institute of Technology, Zurich, Switzerland

Definition

Paleomagnetism is the study of the Earth's ancient magnetic field, based on the measurement and analysis of the *remanent* (i.e., permanent) magnetization acquired by rocks during their formation.

Introduction

Paleomagnetism provides data that are useful in the interpretation of geologic processes. For paleomagnetic results to be reliable for this purpose, several important conditions must be fulfilled. First, a rock under investigation must be able to record accurately the direction and intensity of the weak geomagnetic field during its formation or later alteration. This requires an understanding of the magnetic properties of rocks and minerals. Second, the age of the formation of the rock or its subsequent alteration must be well known. Field tests based on the local and regional geology of a sampling site are valuable in this respect. Third, the remanent magnetization of the rock must possess outstanding stability in order to be able to persist unchanged for geologic eons.

The geocentric axial dipole

The Earth's present magnetic field has been measured and analyzed systematically since Gauss and Weber established the first global network of geomagnetic observatories in 1838. The advent of Earth-orbiting satellites carrying magnetometers has resulted in a vast increase in high quality magnetic data and an improvement of the global coverage. Spherical harmonic analysis of the data shows that more than 99% of the magnetic field measured at the surface of the Earth originates inside the Earth. The potential W_i of the field of internal origin is given by

$$W_i = R\sum_{n=1}^{N}\sum_{m=0}^{n}\left(\frac{R}{r}\right)^{n+1}\left(g_n^m\cos m\phi + h_n^m \sin m\phi\right)P_n^m(\cos\theta)$$

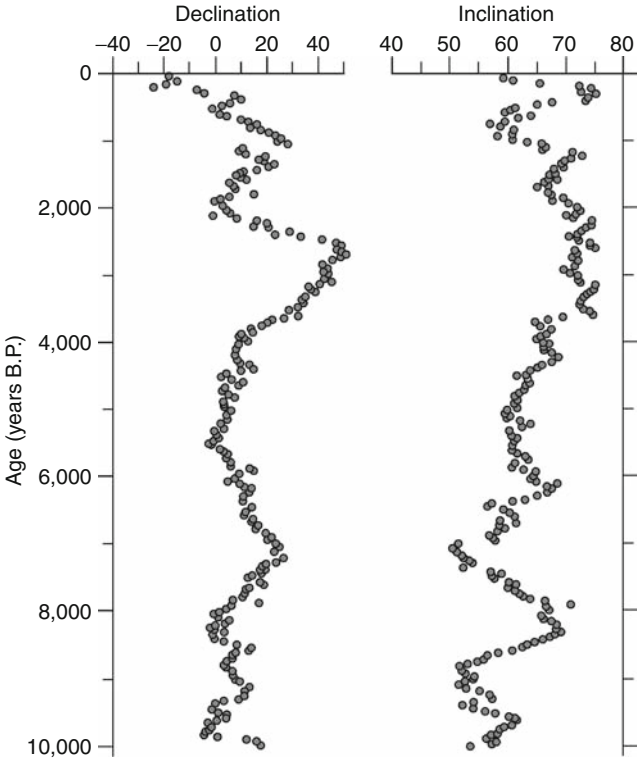
The optimum values of the Gauss coefficients g_n^m and h_n^m are revised every 5 years and define the International Geomagnetic Reference Field (IGRF). In principle, the summation consists of an infinite number of terms, but in practice it is truncated at a particular degree $n = N$ determined by the quality of the data. The recent field model IGRF-11 is defined up to $N = 13$ and contains 195 coefficients. The most important part of the magnetic field is the dipole component, which corresponds to the terms with $n = 1$. All terms with $n \geq 2$ in the spherical harmonic analysis are labeled collectively as the *non-dipole field*. The term g_1^0 is the strongest component of the field, and describes a magnetic dipole located at the center of the Earth and aligned with the rotation axis; the coefficients g_1^1 and h_1^1 correspond to orthogonal equatorial dipole components. Together, the three terms with $n = 1$ describe a geocentric dipole whose axis is inclined to the rotation axis. At present, the axial tilt is about 11°, and the geomagnetic pole does not coincide with the geographic pole. A compass needle aligns with the horizontal component of the magnetic field – the magnetic meridian – and not with the geographic meridian; the angular deviation is the *declination* (D), which varies from place to place on the Earth's surface.

The basic premise of paleomagnetism is the geocentric axial dipole (GAD) hypothesis. This hypothesis postulates that the equatorial dipole components and the non-dipole field average to zero when averaged over sufficiently long intervals of time, leaving only the GAD component as the paleomagnetic record. Consequently, the magnetic pole position calculated in a paleomagnetic investigation defines the location of the paleogeographic axis. Discrepancies from the present axis are interpreted in geologic terms. The *inclination* (I) of the magnetic field is the angle that a magnetic field line makes with the horizontal at Earth's surface. For a GAD field this angle is related to the latitude λ of the observation site by the equation

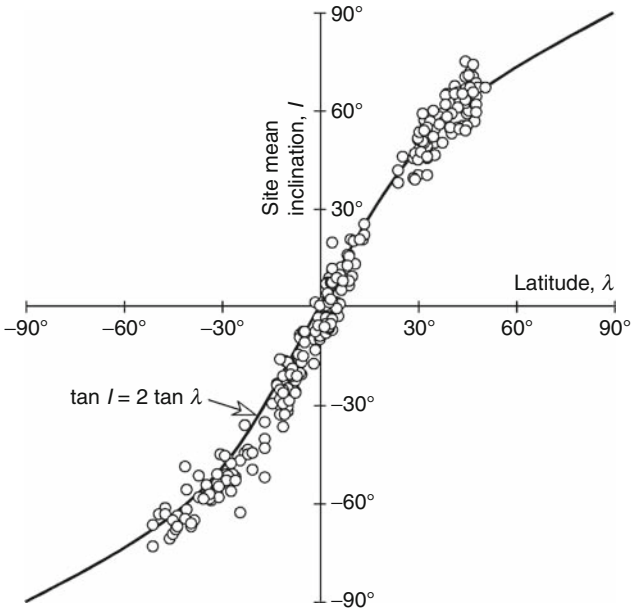
$$\tan I = 2\tan\lambda = 2\cot p \qquad (1)$$

In this equation $p = (90° - \lambda)$ is the colatitude of the observation site, i.e., the angular distance of the magnetic pole from the site.

The dipole and non-dipole fields change slowly with time, a characteristic known as *secular variation*. Time-dependent changes of the Gauss coefficients are observed in repeated analyses of the recent field. Secular variations affect both the direction and the intensity of the field. For example, the geomagnetic field has decayed by about 5% per century since direct measurements began about 160 years ago. It has been possible to reconstruct the



Paleomagnetism, Principles, Figure 1 Master curves of paleosecular variation of inclination and declination for the past 10,000 years, compiled from British lake sediments. (Turner and Thompson, 1981.)



Paleomagnetism, Principles, Figure 2 The mean inclinations of magnetization in marine sediment cores are plotted against the latitude of each sampling site. The continuous curve is the theoretical relationship given by Equation 1 for a geocentric axial dipole (GAD) field. (After Schneider and Kent, 1990.) The small deviations are partly due to paleosecular variation that has not been averaged out, but systematic bias indicates persistent long-term quadrupole and octupole components.

changing strength of the axial dipole from the magnetic observations of mariners since the late sixteenth century; these document a slower, perhaps indeterminate change prior to 1840 (Gubbins et al., 2006).

On a longer timescale, fluctuations in magnetization direction and intensity that are not due to mineralogical variations have been documented in paleomagnetic analyses of sedimentary records. The changes take place over centuries and millennia and are called *paleosecular variations* to distinguish them from changes in the present and historic field. The fluctuations are evident in intensity and direction of the magnetization. For example, oscillations in both inclination and declination measured in cores of sediments from British lakes could be correlated from lake to lake. The sediments were dated using radiocarbon and palynological age determinations. This allowed construction of a master curve of paleosecular variation in Britain for the past 10,000 years (Figure 1).

For the GAD hypothesis to be valid, the paleosecular variations need to average to zero over a few tens of thousands of years, leaving the inclination and declination of the GAD field as the paleomagnetic record. A test of the GAD hypothesis in marine sediments deposited in the last 2.5 Myr compares the measured inclinations with values predicted by Equation 1. The mean inclinations do indeed closely agree with the values expected for a GAD field (Figure 2). However, the agreement is not perfect, and in fact the optimum fit of a theoretical curve to the data is obtained when small geocentric axial quadrupole (g_2^0) and octupole (g_3^0) components are included, each equivalent to a few percent of the dipole. The magnetohydrodynamic theory of the origin of the geomagnetic field gives no rigorous theoretical justification for the GAD hypothesis or explanation of possible persistent zonal quadrupole and octupole components. However, experimental tests of the hypothesis indicate that it is a good first-order model for the time-averaged geomagnetic field.

Rock magnetism

Ferromagnetism is the type of magnetic behavior found in iron and similar metals. It is characterized by a Curie temperature, above which the ferromagnetism disappears. Ferromagnetic metals are strongly magnetic. They have properties such as magnetic hysteresis, saturation magnetization, and remanent magnetization. The important rock-forming minerals exhibit *diamagnetism* or *paramagnetism*, which are very weak forms of magnetism that do not show magnetic hysteresis. Because they cannot acquire a remanent magnetization, they play no part in paleomagnetism. A few minerals, consisting of iron oxides and sulfides, are classified as *ferrimagnetic*, which

is a form of ferromagnetic behavior associated with a crystal structure. Ferrimagnetic minerals are able to acquire a remanent magnetization. Although they occur in tiny concentrations in rocks (e.g., less than 0.01% in some sediments), the paleomagnetic properties of a rock are due to their presence.

The most common ferrimagnetic minerals are members of the titanomagnetite and titanohematite solid solution series, whose end-members are magnetite and hematite, respectively. Magnetite is usually, but not always, a primary magnetic mineral coeval with the origin of igneous rocks or predating the deposition of sediments. Hematite, pyrrhotite, and greigite are other important magnetic minerals, but they originate in the rock after its formation, so the timing of their magnetization can be dubious.

The magnetization of a rock is due to the partial alignment of the magnetic moments of its constituent grains. Within any particular grain the magnetic moment has a direction, conventionally called the "easy" direction of magnetization, which is determined by the magnetic anisotropy of the grain. This, in turn, depends on the intrinsic magnetic properties of the magnetic mineral. For example, magnetite has a cubic structure and a comparatively strong spontaneous magnetization ($M_s = 4.8 \times 10^5$ A m^{-1}), so a grain's shape dominates the anisotropy; the magnetic moment favors the direction of the longest axis of the grain. In contrast, hematite has a rhombohedral structure and a weak spontaneous magnetization ($M_s \sim 2.5 \times 10^3$ A m^{-1}). It has a strong magnetocrystalline anisotropy related to the crystal structure that causes the magnetic moment to lie in the basal plane of the crystal at room temperature.

In very fine grains of ferrimagnetic minerals, the anisotropic magnetic energy of the grain constrains the spontaneous magnetization direction to lie along an easy direction, and the entire grain is uniformly magnetized as a *single domain*. The high spontaneous magnetization of magnetite creates anisotropic magnetostatic energy related to the particle shape. The theoretical range of single domain sizes in magnetite is narrow, from about 0.03–0.1 μm in equant grains and up to about 1 μm in elongate grains. In contrast, hematite has a large magnetocrystalline anisotropy energy and single domain grain sizes may extend up to 15 μm. The magnetization of a single domain particle is very stable, because to change it requires rotating the entire uniform spontaneous magnetization of the grain against the grain anisotropy, which demands a strong magnetic field. The magnetic field needed to reverse the direction of magnetization of a single domain grain is its *coercivity* B_c. In magnetite, this field is commonly in the range 10–100 mT, but may be up to 0.3 T in acicular grains. The magnetic properties of hematite are variable and its maximum coercivity can exceed 2 T. Because of their stable remanent magnetizations, single domain particles play a very important role in paleomagnetism.

When the grain size of a magnetic mineral exceeds a critical value, the magnetostatic energy of the grain becomes too large for the grain to exist as a single domain and the magnetization subdivides into oppositely directed domains. *Multidomain* magnetizations can be changed by merely moving the wall between adjacent domains, so are less stable against change than single domains. As a result, coarse-grained rocks are often unsuitable for paleomagnetism. Laboratory methods have been developed to identify the magnetic minerals in a rock, and to evaluate whether the grain size is single domain, multidomain, or intermediate (the latter category being termed pseudo-single domain).

The magnetic moment of each grain in a rock aligns with a magnetizing field as closely as permitted by the magnetic anisotropy of the grain. Thermal energy acts to destroy the partial alignment of magnetic moments. As a result, a magnetization that originally has the value M_0 decays exponentially with time, so that after time t the remaining magnetization M_r is given by

$$M_r = M_0 \exp(-t/\tau)$$

The rate of decay is governed by the *relaxation time* τ, which is a function of the magnetic properties of the grain and the temperature T. For single domain grains of magnetite

$$\tau = \frac{1}{f} \exp\left(\frac{VB_cM_s}{2kT}\right) \quad (2)$$

Here f is a frequency factor ($\sim 10^9$ s^{-1}) related to the vibrational frequency of the crystal lattice; V is the volume of the grain, B_c its microscopic coercivity, and M_s its spontaneous magnetization; and k is Boltzmann's constant. A single domain grain of magnetite, 30 nm in diameter, has a relaxation time of about 10^9 year at 20°C. Other ferrimagnetic minerals can also possess this extraordinary stability against thermal decay. It allows a rock to retain the remanent magnetization it acquired during rock formation for very long intervals of geologic time. The relaxation time of the magnetization decreases with increasing temperature and with decreasing grain size. Thus, heating during tectonic deformation, or a change of grain size or mineral chemistry as a result of weathering, can modify or destroy a primary remanent magnetization.

Remanent magnetizations

The development of paleomagnetism as a scientific discipline was made possible by the invention in the early 1950s of the astatic magnetometer, a device that was sensitive enough to measure accurately the weak magnetizations of rocks. The instruments commonly used in modern paleomagnetic laboratories are spinner magnetometers and cryogenic magnetometers, which are even more sensitive and are able to measure rock samples much faster. As a result of the low concentrations of ferrimagnetic minerals, rocks have very weak remanent magnetizations, measuring typically $\sim 10^{-1}$ A m^{-1} for a basaltic lava and $\sim 10^{-3}$ A m^{-1} for a pelagic limestone. The intensity of

a remanent magnetization also reflects the mechanism by which the rock becomes magnetized in the Earth's magnetic field.

When an igneous rock forms, its temperature is initially well above the Curie temperature of ferrimagnetic minerals (e.g., 580°C in magnetite, 675°C in hematite). The magnetic behavior is paramagnetic, incapable of retaining a magnetization. As it cools below the Curie point, the magnetite or hematite becomes ferrimagnetic. At high temperature, the relaxation time is short and the magnetic behavior is classified as *superparamagnetic*. With further cooling, the relaxation time increases rapidly, and a few tens of degrees below the Curie point a *blocking temperature* is reached, at which the relaxation time is so large that the magnetization becomes blocked by the anisotropy of the grain. If this ensues in a magnetic field, the ferrimagnetic grains acquire a *thermoremanent magnetization* (TRM) in the direction of the field. Equation 2 controls this process and ensures that the TRM has a high stability. On the other hand, if subsequent chemical alteration of the grain occurs, the grain volume V or spontaneous magnetization M_s may alter, which, again according to Equation 2, can give rise to a new, stable *chemical remanent magnetization* (CRM). This is usually an undesirable secondary component.

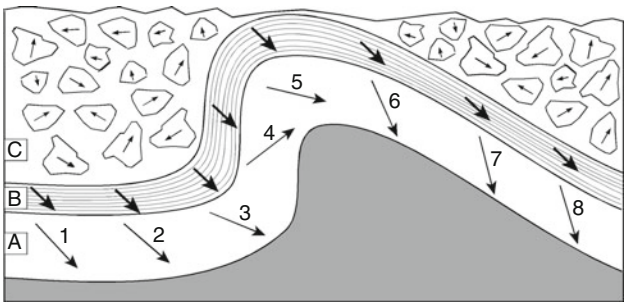
Sedimentary rocks are deposited slowly and can often be excellent archives of the ancient magnetic field. Magnetite grains form a tiny fraction of a sediment (e.g., less than 0.01% of a limestone) and are aligned partially by the magnetic field in which they settle, forming a *depositional remanent magnetization* (DRM). In some very fine-grained sediments, further magnetite alignment takes place in the waterlogged pore spaces between the nonmagnetic matrix minerals. The sedimentation rate of pelagic marine sediments is so slow that paleosecular variations are in principle averaged out within the size of a standard 1-in. paleomagnetic sample. However, in lake sediments the sedimentation rate is much higher and it is possible to study both secular variation (Figure 1) and paleoclimatic effects in these sediments. Shallow-water marine limestones also have a relatively high deposition rate, but the detrital and biogenic fractions that supply the magnetite are diluted in these limestones and they tend to have weak and unstable remanent magnetizations.

In addition to the remanent magnetizations acquired in a geological process, synthetic remanent magnetizations may be imparted to rock samples in the laboratory for the purpose of identifying and describing the carrier of a paleomagnetic signal. If a sample is placed in a magnetic field, the magnetic grains with coercivity less than the field are realigned; when the field is removed, an *isothermal remanent magnetization* (IRM) is left in the sample. If the field is strong enough to magnetically saturate the sample, a *saturation* IRM ensues. Another, important laboratory magnetization is produced by placing a sample in an alternating magnetic field in the presence of a constant bias field. When the alternating field is reduced to zero, an *anhysteretic remanent magnetization* (ARM) is produced in the sample. Both IRM and ARM can be used in various ways to gain information about the magnetic mineral and its grain size.

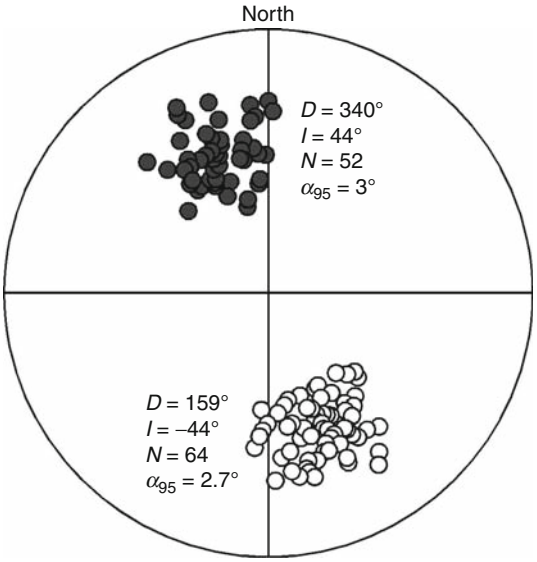
In a paleomagnetic study, special laboratory techniques are used to isolate the components of magnetization present in the rock and to identify the primary component. This is the remanent magnetization acquired during formation of the rock. The remanent magnetization of a rock in situ is called its *natural remanent magnetization* (NRM). It may contain a primary component of geological significance, but this can be accompanied by secondary components of later origin. These components can arise from geological processes, but some may also be induced during sampling. The secondary components can partially – in extreme cases, completely – overprint the primary component and must be understood for the paleomagnetic analysis to have geological significance. By progressively demagnetizing the samples, the structure of the NRM can be analyzed and stable components isolated. There are two main methods of achieving this. When a rock sample is placed in an alternating magnetic field (AF), the magnetizations of all domains with coercivity less than the peak AF field are randomized. If the peak field is slowly reduced to zero and the equipment is in a space shielded from the Earth's field, this part of the rock magnetization is demagnetized. The part that remains has been "magnetically cleaned." If the procedure is carried out in progressively increasing peak fields, the NRM decay shows the components present. Similarly, when a rock sample is heated to a given temperature T, components that have lower blocking temperatures than T are thermally randomized. If the sample is then cooled in field-free space, this part of the NRM remains demagnetized. The progressive destruction of the NRM by AF or thermal demagnetization reveals the components present in the NRM. The stable primary component is identified computationally by principal component analysis of the directions during magnetic cleaning (Kirschvink, 1980) and graphically using vector diagrams (Zijderveld, 1967).

For paleomagnetic results to be a valid record of the ancient magnetic field, it is important to know that the remanent magnetization has the same age as the rock, so identification of the magnetic mineral is important. However, this is not a sufficient criterion and field tests play a crucial role in establishing whether a magnetization component is primary or secondary. The most important are the structural fold test and the magnetic reversals test; less common, but also useful, are the conglomerate test and baked contact test.

The principles underlying the fold and conglomerate tests are illustrated in Figure 3. Layers A and B are folded rocks, and layer C is a conglomerate layer of large cobbles. The *conglomerate test* examines whether the cobble magnetizations have random directions, which would indicate that the magnetization of the parent formation from which the cobbles derive is stable. The *fold test* examines the direction of magnetization around a fold.



Paleomagnetism, Principles, Figure 3 The fold and conglomerate tests of the paleomagnetic stability of magnetization directions. The cobbles in the conglomerate (layer C) have random directions of magnetization, which indicates these are stable. The directions are also stable in layer A, changing around the fold, but in layer B their uniformity indicates they have been remagnetized after folding. (After Lowrie, 2007.)



Paleomagnetism, Principles, Figure 4 Positive reversals test in a section of the *Scaglia cinerea* formation in Umbria, Italy. The mean of 52 normal directions (*solid points*) has declination 340°, inclination 44° (α_{95} = 3.0°), and differs by less than 1° from the antipode of the mean of 64 reversed directions (open points), which has declination 159°, inclination –44° (α_{95} = 2.7°).

If the magnetization of a formation is stable, as in layer A, its direction will change from site to site around the fold; if it is unstable, as in layer B, all sites in the layer will have a common magnetization direction.

If a sedimentary formation is deposited over a long time interval, or if a collection of similarly aged igneous rocks covers enough time so that both normal and reversed intervals of geomagnetic polarity are represented, the *reversals test* may be applied. If secular variation is adequately averaged out, the reversed directions should be antipodal to the normal directions, as in Figure 4. The *baked contact test* is applied to igneous flows and intrusions. The heat of a lava dike can remagnetize the contact zone of a rock it intrudes, or on which it flows. If the paleomagnetic directions in the baked contact zone agree with those of the lava and differ from the rest of the host rock, the lava has a primary magnetic direction. In each of these tests, rigorous statistical methods are used to decide the significance of the magnetization directions.

Statistical testing of paleomagnetic data

Paleomagnetic studies are based on a hierarchical sampling system. A comparatively small number of samples, n, is taken at each of N sampling sites in the geological formation of interest; N and n are each commonly around 10–20. Statistical tests are carried out on the measured data at each hierarchical level. Testing the significance of the results is based on a method developed by Fisher (1953) that represents the paleomagnetic directions as unit vectors. These can be visualized as a distribution of points on a sphere, represented by a stereographic plot. The vector mean of the unit vectors is the best estimate of the population mean direction. Fisher suggested that the probability density P of the deviations θ between the individual sample directions and their mean direction is given by

$$P(\theta, \kappa) = \frac{\kappa}{4\pi \sinh \kappa} \exp(\kappa \cos \theta)$$

The "precision parameter" κ describes the dispersion of the points. If $\kappa = 0$, the directions are uniformly or randomly distributed; if κ is large, the points cluster tightly about their mean direction. The best estimate of κ is k, defined as

$$k = \frac{N - 1}{N - R}$$

in which R is the vector sum of the N unit vectors. In paleomagnetism the level of 95% confidence is used to define a circle with radius α_{95} around the estimated mean direction; there is a 95% probability that the true mean of the distribution lies within this circle. An approximate formula (McElhinny and McFadden, 2000) for the radius of the circle of confidence, especially appropriate for the small numbers of samples or sites in a paleomagnetic study, is

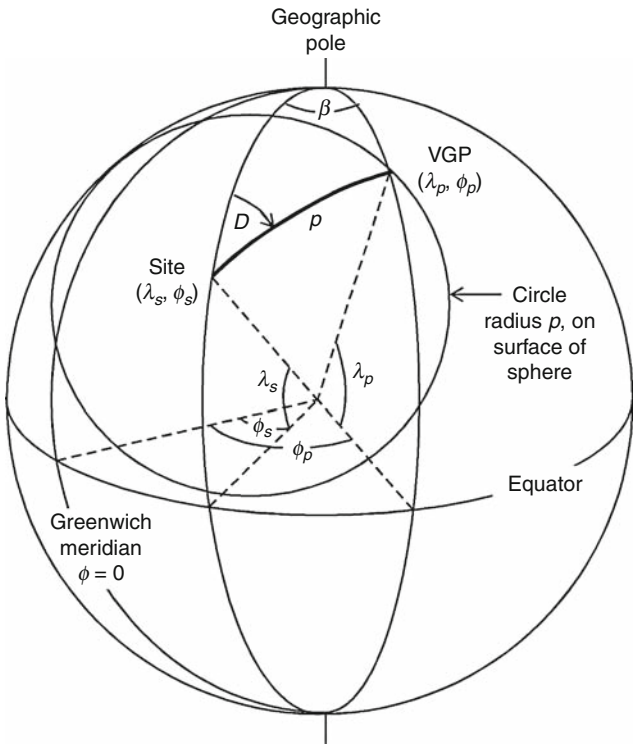
$$\alpha_{95} = \frac{140}{\sqrt{kR}}$$

The statistical definition of a confidence circle around the mean direction allows comparison of a paleomagnetic result with other data of the same age and from the same region. It allows critical evaluation of the directions before and after tectonic corrections (the fold test), as well as the equivalence of antipodal directions in a reversals test.

Statistical analysis lends credence to the results and is an important aspect of all paleomagnetic investigations. The statistical methods can be applied to paleomagnetic directions, and also to distributions of paleomagnetic pole positions.

Virtual geomagnetic poles and apparent polar wander paths

Assuming the GAD hypothesis, the location of the magnetic pole at the time of formation of a given rock formation is coincident with the geographic pole. As a result of subsequent plate tectonic motions, the location of the paleomagnetic pole becomes displaced and rotated. When the magnetizations of the rocks are measured, they appear to have been magnetized by a pole at the displaced location, which is accordingly designated the *virtual geomagnetic pole* (VGP). The principle of locating a VGP from the measured declination and inclination of the stable component of remanent magnetization in a rock is illustrated in Figure 5. First, the inclination is used to calculate the angular distance p to the magnetic pole, employing Equation 1. This defines a circle with radius p centered on the sampling site; it is the locus of possible pole locations. The VGP is located at the intersection of this circle and the great circle defined by the declination; a magnetic pole at this location gives the measured inclination and declination at the sampling site.



Paleomagnetism, Principles, Figure 5 Method for locating the virtual geomagnetic pole in a paleomagnetic study. The distance p to the pole is computed from the paleomagnetic inclination I; the virtual geomagnetic pole (VGP) is located at the intersection of the circle with radius p and the great circle with declination D.

The VGP positions of samples from geologically young rocks cluster near to the geographic pole, corresponding to the GAD hypothesis. Older rocks from the same continent give VGP positions that group well away from the present rotation axis. They plot systematically along a curved path, called an apparent polar wander (APW) path, which appears to trace the motion of the paleomagnetic pole relative to that continent. However, a different APW path is defined by the paleomagnetic data for each continent. Assuming the GAD hypothesis, the paleomagnetic pole at any time lies on the rotation axis. The Earth has only one rotation axis, so it follows that each APW path is a record of the motion of its continent relative to the rotation axis. The APW paths for different continents diverge because the continents have been displaced relative to each other. This motion of the continents was originally called "continental drift" and is the result of plate tectonics.

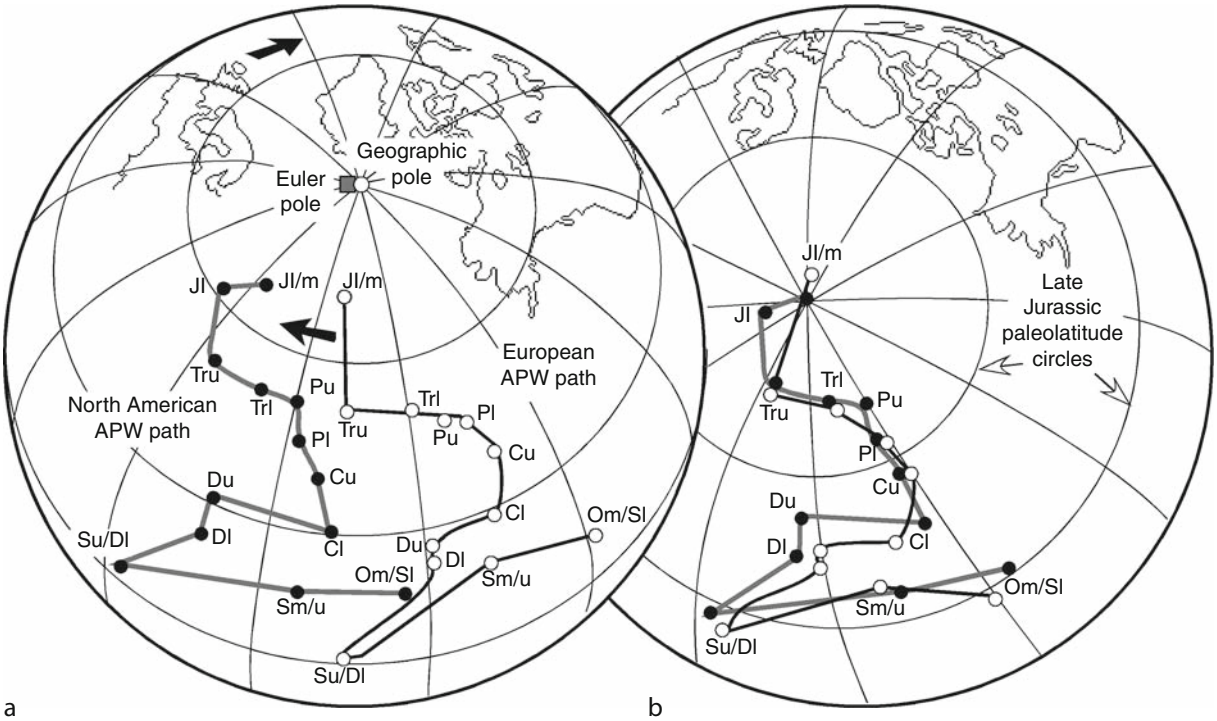
The use of paleomagnetic data for tracing continental movements is illustrated by comparing the APW paths for Europe and North America prior to the Jurassic. Their shapes are similar (Figure 6a), particularly for the Upper Carboniferous (Cu) to the Upper Triassic (Tru). The relative motion of plates on the surface of the spherical Earth is equivalent to a relative rotation about an *Euler pole* of rotation. The APW path of a continent is constrained to move with the continent. The Cu-Tru segments of the two APW paths can be made to overlap by a clockwise rotation of Europe through 38° (Figure 6b) about an Euler pole close to the present-day geographical pole (Figure 6a). Later segments of the APW paths diverge, indicating relative motion between Europe and North America since the Early Jurassic.

Continental reconstruction based on paleomagnetic data requires that the APW paths are securely defined. In order to ensure this, stringent selection criteria have been developed to ensure that the paleomagnetic data accepted for incorporation into the reference database are of satisfactory quality (Table 1).

Magnetic polarity reversals

The geomagnetic field has reversed polarity numerous times, apparently spontaneously and randomly. The record of polarity reversals has been studied in detail in igneous and sedimentary rocks, and in sediments. Beginning in the 1950s, systematic studies of paleomagnetic direction were made in radiometrically dated lavas for the most recent 5 Myr of Earth history. The polarity sequence was confirmed by investigations in slowly deposited marine sediments covering the same time interval, which established that the reversals are a geomagnetic phenomenon and not merely a rock magnetic artifact.

Intervals of constant normal or reverse polarity are called polarity *chrons*; they last typically for 10^5–10^6 year.



Paleomagnetism, Principles, Figure 6 Reconstruction of earlier positions of Europe and North America based on apparent polar wander (APW) paths. (After Van der Voo, 1990a.) In (**a**) the APW paths are shown in present coordinates; in (**b**) the European continent has been rotated by 38° clockwise about the Euler pole shown in (**a**) to bring the Cu-Tru segments of the APW paths into coincidence.

Paleomagnetism, Principles, Table 1 Reliability criteria for paleomagnetic data. Van der Voo (1990b) suggested that acceptable paleomagnetic data should meet at least three of the following criteria

No.	Brief description of reliability criterion
1	Well-determined rock age and a presumption that the magnetization has the same age.
2	Sufficient number of samples ($N \geq 24$), adequate precision of directional grouping ($k \geq 10$), and confidence angle $\alpha_{95} \leq 16°$.
3	Progressive demagnetization that demonstrably includes vector subtraction information.
4	Field tests that constrain the age of magnetization.
5	Structural control and tectonic coherence with the craton or block involved.
6	The presence of reversals.
7	No resemblance to paleomagnetic poles of younger age (by more than a period).

The chrons may be interrupted at irregular intervals by polarity *subchrons* that last on the order of 10^4–10^5 year. The *transition* from one polarity to the opposite one is estimated to take about 3–10 kyr, which is much shorter than a polarity chron or subchron. The polarity transition is accompanied by a decrease in intensity of the paleofield, which may last up to twice as long as the directional change. The polarity record also shows occasional large *excursions*, in which the magnetic pole spends 10^3–10^4 year in equatorial latitudes, but instead of completing a reversal it returns to its initial location on the rotation axis.

The irregular pattern of polarity intervals provides a kind of geological fingerprint, which can be used to date and correlate other magnetic polarity sequences in both igneous and sedimentary rocks, and sediments. The most important application has been the interpretation and dating of marine magnetic anomalies. New oceanic lithosphere forms at oceanic ridge systems and is then transported away from the ridges by the seafloor spreading process. As the basaltic lava of the ocean crust cools, it acquires a TRM parallel to the Earth's magnetic field. Seafloor spreading persists for millions of years at a ridge, during which the magnetic field changes polarity many times. This leaves some blocks of oceanic crust normally magnetized parallel to the field while their neighbors are reversely magnetized. A sequence of positive and negative magnetic anomalies, corresponding to normal and reverse polarities of the oceanic crust, is observed over the ridges, from which the sequence of polarity of the Earth's magnetic field is obtained. The oceanic polarity pattern correlates with the polarity sequence in lavas that have been dated radiometrically, thus the ages of magnetized blocks of the ocean floor can be determined.

Knowing the distances of the polarity changes from the oceanic ridge, the rate of seafloor spreading can be calculated.

Older reversal patterns have been dated by magnetostratigraphy. This entails stratigraphic sampling of a sedimentary section characterized by a uniform lithology and deposited at a uniform rate. Standard paleomagnetic techniques are used to evaluate the paleomagnetic direction of each sample and establish the existence of *magnetozones* of constant polarity. As is the case with paleomagnetic data, strict reliability criteria regulate the acceptance of a magnetic stratigraphy into a global database (Table 2).

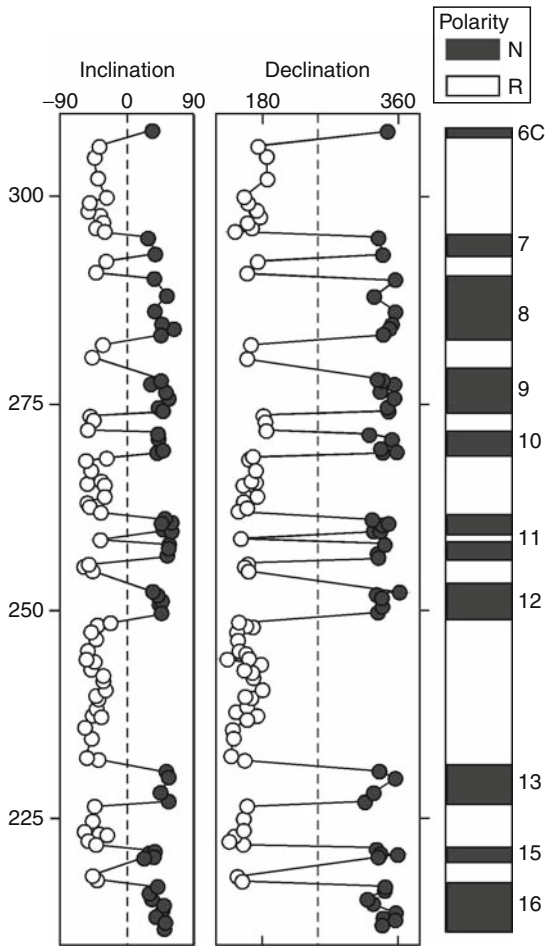
An example of coordinated magnetostratigraphy and biostratigraphy that correlates well with marine magnetic reversal history is shown in Figure 7. The *Scaglia cinerea*, a marly limestone in Umbria, Italy, was deposited in the Late Paleogene and acquired a DRM that preserves the direction and polarity reversals of the geomagnetic field during deposition. Paleontological dating of the limestones enabled correlation of major stage boundaries with the polarity sequence. The absolute ages of many stage boundaries are known independently, so the reversal pattern could be dated. Correlation with marine magnetic anomalies allowed the corresponding regions of the oceanic crust to be dated. The polarity sequence correlates well with the history of geomagnetic polarity derived from marine magnetic anomalies for chrons numbered 6C to 16 (Cande and Kent, 1992).

Magnetic polarity stratigraphy has served in this way to correlate and date the oceanic magnetic record in many paleomagnetic sections, on land and in cores from the ocean bottom. The ages of key parts of the geomagnetic polarity record have thus been determined at many correlation levels. The ages of magnetic reversals between dated tie-points can be computed by interpolation or extrapolation. The dated polarity sequence is called a geomagnetic polarity timescale (GPTS).

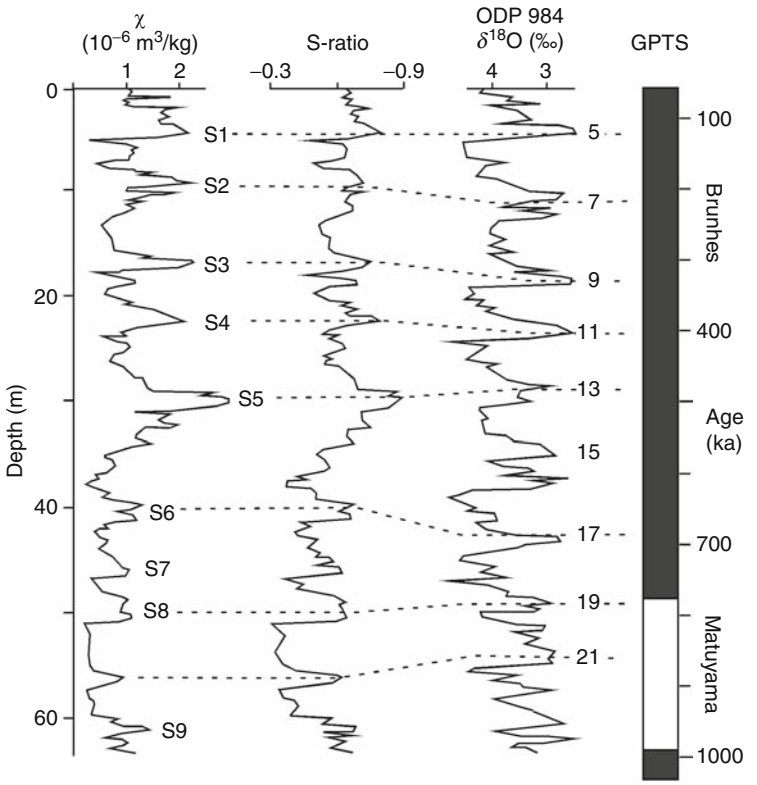
The record of geomagnetic polarity is well established from the present back to the Late Jurassic (Oxfordian), equivalent to the last 160 Myr of geological time. Three distinct episodes of reversal behavior are identified in this time interval. The oldest episode, comprising frequent reversals in the Late Jurassic and Early Cretaceous, is referred to as the M-sequence. The youngest episode, comprising the Late Cretaceous and Cenozoic, consists of frequent reversals known as the C-sequence. The M-sequence and C-sequence are separated by an interval, 38 Ma in length, called the Cretaceous Normal Polarity Superchron (CNPS), in which correlated magnetic anomalies are absent; evidently the Earth's magnetic field did not reverse polarity during this time. The episodes of

Paleomagnetism, Principles, Table 2 Reliability criteria for magnetostratigraphy. Opdyke and Channell (1996) proposed that an acceptable magnetostratigraphic section should pass at least five of the following ten tests

No.	Brief description of reliability criterion
1	Stratigraphic age known to the level of the stage, and associated paleontology presented adequately.
2	Sampling localities placed in a measured stratigraphic section.
3	Complete thermal or alternating field demagnetization performed, and analysis of magnetization components carried out using orthogonal vector projections.
4	Stable paleomagnetic directions determined from principal component analysis.
5	Data published completely as plots of VGP latitudes and/or of declination and inclination against stratigraphic position. The statistical parameters should be fully documented.
6	Magnetic mineralogy determined.
7	Field tests to constrain the age of magnetization undertaken where possible.
8	A positive reversals test should be carried out.
9	Radiometric ages, especially $^{40}A/^{39}Ar$ or U-Pb ages from volcanic ashes or bentonites, available in the stratigraphic section.
10	Multiple overlapping sections confirm the polarity sequence.



Paleomagnetism, Principles, Figure 7 Declination and inclination profiles in the Late Paleogene *Scaglia cinerea* formation, whose directions are shown in Figure 4, correlated to the history of geomagnetic polarity derived from marine magnetic anomalies for chrons numbered 6C to 16.



Paleomagnetism, Principles, Figure 8 Profiles of magnetic susceptibility (χ), hardness of remanent magnetization (S-ratio), and paleomagnetic polarity in a loess section at Sanmenxia, China, correlated to the marine oxygen-isotope ratio ($\delta^{18}O$) at Ocean Drilling Program (ODP) site 984. (Data source: Wang et al., 2006.) The peaks marked S1, S2, etc., correspond to paleosols; the lows between them are loess beds.

reversing behavior are long enough to make tentative interpretations about the reversal process. It is evident from the patterns of reversals that they do not occur cyclically. Making allowance for the time spent in transition between opposite polarity states, statistical analysis indicates that the reversals in the C- and M-sequences occur randomly. There is as yet no convincing evidence to refute identification of the CNPS as an uninterrupted period of constant normal polarity. There are competing interpretations of its origin. Whereas some investigators consider that the CNPS represents a special behavior of the geodynamo, others regard it as an exceptionally long feature in a continuous reversal process.

The breakup of Pangea and the onset of seafloor spreading date from the Early Jurassic, for which magnetostratigraphic confirmation of the marine polarity record is fragmentary. For earlier eras there is no marine polarity record to which a measured magnetostratigraphy can be correlated. Consequently, despite many magnetostratigraphic investigations in ancient rocks, there is no consistent dated polarity record older than Late Triassic. To establish a GPTS for earlier times will require verifying the same polarity sequence in multiple overlapping sections. This presents a large but important task for paleomagnetists and biostratigraphers.

Environmental magnetism

In recent years, knowledge of the magnetic properties of minerals, the grain-size dependence of parameters of magnetic hysteresis, and the use of magnetic stratigraphy have been combined in applications to a wide range of environmental problems. For example, regional environmental pollution by heavy industry and traffic has been investigated by using ferromagnetic properties to trace heavy metal contamination. However, most environmental studies have been made in sedimentary archives. The magnetic susceptibility of a sediment reflects both the concentration and type of the magnetic minerals, whereas variations in grain size may be revealed in the stratigraphic variations of magnetic coercivity and remanent magnetizations produced in the laboratory. The proportion of an IRM carried by grains with coercivity above a certain value (the so-called S-ratio) can be an indicator of the relative size of the fine-grain fraction of the ferromagnetic grains. Lake sediments and wind-blown loess deposits have comparatively high deposition rates. Detailed evaluation of the magnetic signal in these sedimentary archives can reveal variations in environmental conditions during deposition. The magnetic properties of Chinese loess deposits illustrate the way in which magnetic parameters can act as proxy for paleoclimatic indicators.

The boundary between the Brunhes and Matuyama polarity chrons (age 780 ka), and the upper and lower boundaries of the Jaramillo normal polarity subchron (990 and 1,070 ka, respectively), tie the loess section at Sanmenxia, China, to deep-sea sediments cored at Ocean Drilling Program (ODP) Site 694 (Figure 8). The wind-blown loess sediments in central China were deposited in cold climatic conditions. Paleosols formed under warmer, moister conditions were conducive to enhancement of the magnetic susceptibility. Alternations of loess deposits and paleosols are accompanied by fluctuations in magnetic susceptibility, with low values marking the loess and higher values in the paleosols. Oxygen-isotope results at the ODP site reveal cycles of climatic variation, which correlate well with the magnetic susceptibility and S-ratio data from the Sanmenxia section. The results demonstrate the usefulness of magnetic mineralogy parameters as proxies for isotopic data that document paleoclimatic variation.

Summary

Apart from isolated but important pioneering studies, in which the characteristics of the geomagnetic field and the rudiments of rock magnetism were explored, paleomagnetism originated as a scientific discipline in the 1950s. As a result of the continuous development and refinement of instrumentation and analytic techniques, it has made many notable contributions to geological knowledge. Initially it focused on the tectonic problem of "continental drift" and by developing apparent polar wander paths for the different continents it documented incontrovertibly the mobility of the continents. Matching of APW paths led to the reconstruction of supercontinents such as Rodinia, Pangea, and Gondwana. In parallel to this work, paleomagnetic analysis in conjunction with geological dating worked out the history of geomagnetic reversals. On the basis of the established reversal history and the rock magnetic properties of oceanic basalts, the hypothesis of seafloor spreading was developed. This development provided the key to the mechanism of plate tectonics and allowed the documentation and dating of plate motions, and explained how "continental drift" took place. In turn, the pattern of magnetic reversals in the oceanic crust along with magnetostratigraphy established a dated sequence of geomagnetic polarity history for the past 160 Ma. The methods of paleomagnetism and rock magnetism have found wider application in addressing environmental problems and in contributing to an understanding of paleoclimatic history.

Bibliography

Cande, S. C., and Kent, D. V., 1992. A new geomagnetic polarity time scale for the Late Cretaceous and Cenozoic. *Journal of Geophysical Research*, **97**, 13917–13951.

Fisher, R. A., 1953. Dispersion on a sphere. *Proceedings of Royal Society of London Series A*, **217**, 295–305.

Gubbins, D., Jones, A. L., and Finlay, C. C., 2006. Fall in Earth's magnetic field is erratic. *Science*, **312**, 900–901.

Kirschvink, J. L., 1980. The least-squares line and plane and the analysis of paleomagnetic data. *Geophysical Journal of the Royal Astronomical Society*, **62**, 699–718.

Lowrie, W., 2007. *Fundamentals of Geophysics*, 2nd edn. Cambridge: Cambridge University Press.

McElhinny, M. W., and McFadden, P. L., 2000. *Paleomagnetism: Continents and Oceans*. San Diego: Academic Press.

Opdyke, N. D., and Channell, J. E. T., 1996. *Magnetic Stratigraphy*. San Diego: Academic Press.

Schneider, D. A., and Kent, D. V., 1990. The time averaged paleomagnetic field. *Reviews of Geophysics*, **28**, 71–96.

Turner, G. M., and Thompson, R., 1981. Lake sediment record of the geomagnetic secular variation in Britain during Holocene times. *Geophysical Journal of the Royal Astronomical Society*, **65**, 703–725.

Van der Voo, R., 1990a. Phanerozoic paleomagnetic poles from Europe and North America and comparisons with continental reconstructions. *Reviews of Geophysics*, **28**, 167–206.

Van der Voo, R., 1990b. The reliability of paleomagnetic data. *Tectonophysics*, **184**, 1–9.

Wang, X., Yang, Z., Løvlie, R., Sun, Z., and Pei, J., 2006. A magnetostratigraphic reassessment of correlation between Chinese loess and marine oxygen isotope records over the last 1.1 Ma. *Physics of the Earth and Planetary Interiors*, **159**, 109–117.

Zijderveld, J. D. A., 1967. AC demagnetization of rocks. In Collinson, D. W., Creer, K. M., and Runcorn, S. K. (eds.), *Methods in Paleomagnetism*. Amsterdam: Elsevier, pp. 254–286.

Cross-references

Curie Temperature
Geomagnetic Field, IGRF
Geomagnetic Field, Polarity Reversals
Geomagnetic Field, Secular Variation
Magnetic Domains
Paleomagnetism, Magnetostratigraphy
Paleomagnetism, Measurement Techniques and Instrumentation
Paleomagnetism, Polar Wander
Remanent Magnetism

PALEOSEISMOLOGY

Shinji Toda
Disaster Prevention Research Institute, Kyoto University, Uji, Kyoto, Japan

Synonyms

Earthquake archaeology; Earthquake geology

Definition

Paleoseismology. An interdisciplinary field dealing with prehistoric or pre-instrumental earthquakes, in particular, their location, timing, and magnitude.

Paleoseismology

Introduction

Modern instrumental records of earthquakes cover only a little more than a century since 1890s (see *Seismic Instrumentation*). Historic accounts in several countries

such as China, Japan, and Mediterranean regions allow us to extend the records of large earthquakes up to more than a couple of thousand years. However, for most parts of the globe, no earthquake data are available for the time periods of even the past hundreds to thousands of years that are comparable to repeat time of large earthquakes on a fault. Nevertheless, geologic sedimentary strata and geomorphic features preserve evidences for strong shaking, ground breaks, and landscape changes associated with large earthquakes. Although geologic and geomorphic records are incomplete and poorly constrained in time and space, they greatly help to understand faulting process and recurrence of earthquakes that are unavailable to seismologists due to the brevity of historic documents. We have defined this discipline as "paleoseismology" and such individual earthquakes are referred to as "paleoearthquakes."

Paleoseismology is not a subbranch of seismology, but rather it is an interdisciplinary field of research, which borrows many concepts from structural geology and active tectonics. Methodology and techniques in paleoseismology are derived primarily from Quaternary geology and related disciplines such as geomorphology, soil mechanics, sedimentology, archaeology, paleoecology, photogrammetry, radioisotope dating or pedology (Grant, 2007; McCalpin, 2009).

The history of paleoseismology is as short as that of instrumental seismology. Although several pioneer geologists described surface faulting accompanying earthquakes in the late 1800s, geology and seismology at that time were separated (Yeats and Prentice, 1996). But a few leading geologists, e.g., R. E. Wallace, C. R. Allen, and D. B. Slemmons, who studied San Andreas fault and active scarps in Nevada, and radiocarbon dating technique pervasive since about 1960s–1970s, played prominent roles in bridging the gulf between instrumental seismology and geology (Yeats and Prentice, 1996). Since the 1970s, paleoseismic investigations have been yielding information about the paleo-movements on faults, dates of previous earthquakes, recurrence times, average slip rate, and earthquake effects over time intervals ranging from decades to thousands of years (Grant, 2002). Consequently, numerous paleoseismic data (e.g., Sieh, 1978) and fundamental concepts of recurrent large earthquakes (e.g., Schwartz and Coppersmith, 1984) derived from field observations have largely contributed to progress in understanding the mechanics of earthquakes and faulting.

Unlike many subbranches of geology, the uniqueness of paleoseismology lies in its potential to decipher the future from the past and present. The traditional geological doctrine of "uniformitarianism," which is used to reconstruct the past events from phenomena observable in the present, can be stretched to the future in paleoseismology to perform seismic hazard estimates and earthquake forecasting. The final goal and major contribution of paleoseismology is to not only understand long-term faulting processes, but also to provide fundamental data and speculations to mitigate seismic hazard.

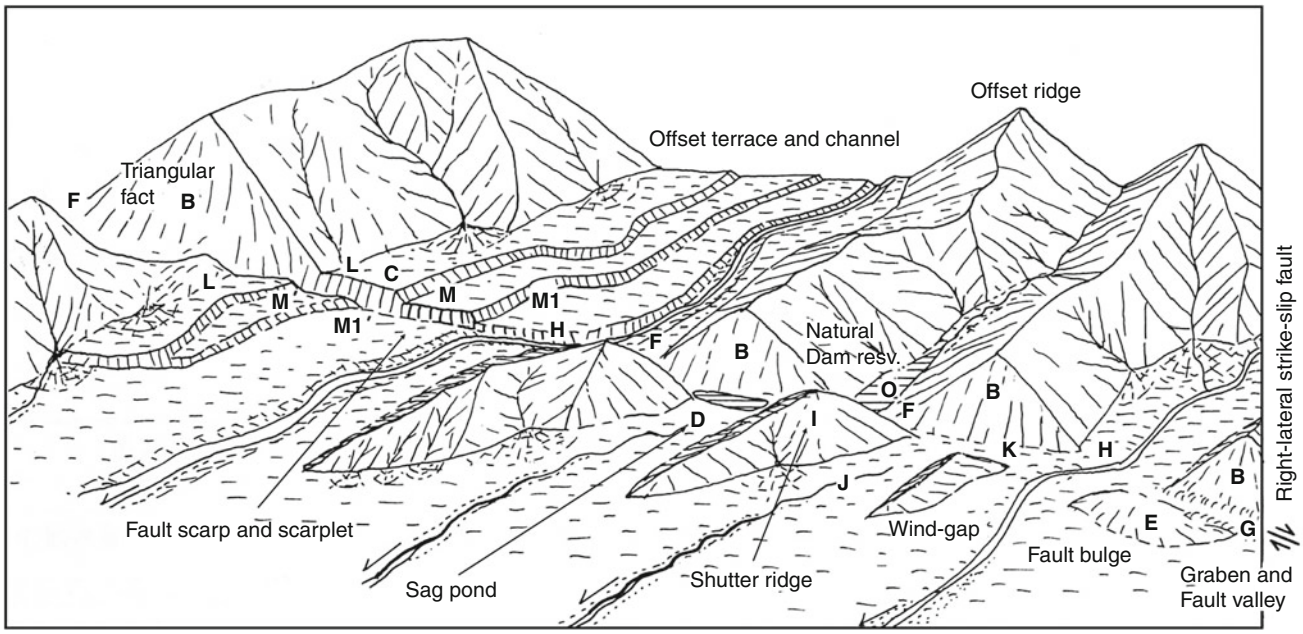
Paleoseismic records

In seismology, estimation of earthquake magnitude depends on the performance of seismometers and on network coverage (see *Seismological Networks*; *Statistical Seismology*). Geomorphic markers, surface soil, and sediments are like poorly performing strong-motion seismograms. If an earthquake is large and shallow enough, the seismogenic faulting normally reaches the Earth's surface. For inland intraplate and interplate earthquakes occurring shallower than ~20 km ("seismogenic layer," see entries on *Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat* and *Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity*), $M \geq 6$ earthquakes may leave tectonic surface breaks and permanent ground offsets (Wells and Coppersmith, 1994). The minimum magnitude of a paleoearthquake we can identify in strata and landscape is about $M \sim 6$ or higher (discussed later). For offshore earthquakes, including subduction associated mega-thrust events, $M \geq 7.5$ earthquakes leave records of significant coastal uplift and subsidence (e.g., Berryman et al., 1989). However, the paleoseismic record is spatially and temporally incomplete. Unlike instrumental seismology, recording processes by sedimentary layers are intermittent, and erosion often destroys the evidence for paleoearthquakes.

Fault identification and slip rate

Identifying an active fault is the first step for paleoseismology. As described above, a large earthquake often leaves surface breaks and warping with significant offset when the faulting reaches the Earth's surface. Detailed excavation studies often reveal evidence for each individual earthquake at an individual site. However, recognizing active traces relies on identification of geomorphic expression due to repetition of surface-faulting earthquakes as "fault landforms" for which deformation rate must be higher than local rates of erosion and deposition. Although the definition of "active fault" partly depends on the purposes (engineering or scientific) and on the usage of the terminology in different countries, it is usually identified by its association with tectonically deformed Quaternary-age materials or surfaces and by its potential to cause a large earthquake in near future.

To identify active faults and associated active landforms, aerial photographs and/or satellite images are commonly used ("aerial photographic interpretation"). High-resolution satellite images may be used to identify faulted landforms, even finding coseismic surface ruptures, around the globe, particularly in hardly accessible areas (e.g., Lin et al., 2006). The most frequently encountered feature is topographic lineaments. But lineaments are often developed by a combination of constructive geologic features and erosion. Therefore, other characteristics associated with faulting such as triangular facets, fault scarps, sag ponds, and bulges should be found along



Paleoseismology, Figure 1 Topographic features formed by right-lateral offset (modified from Okada, 1991 and Research Group for Active Faults of Japan, 1992). B, triangular facet; C, fault scarplet; D, sag pond; E, tectonic bulge; F, fault saddle; G, graben; H, offset stream; I, shutter ridge; J, beheaded stream; K, wind gap; L-L′, offset of piedmont line; M-M′, offset of terrace edge; Q, fault pond.

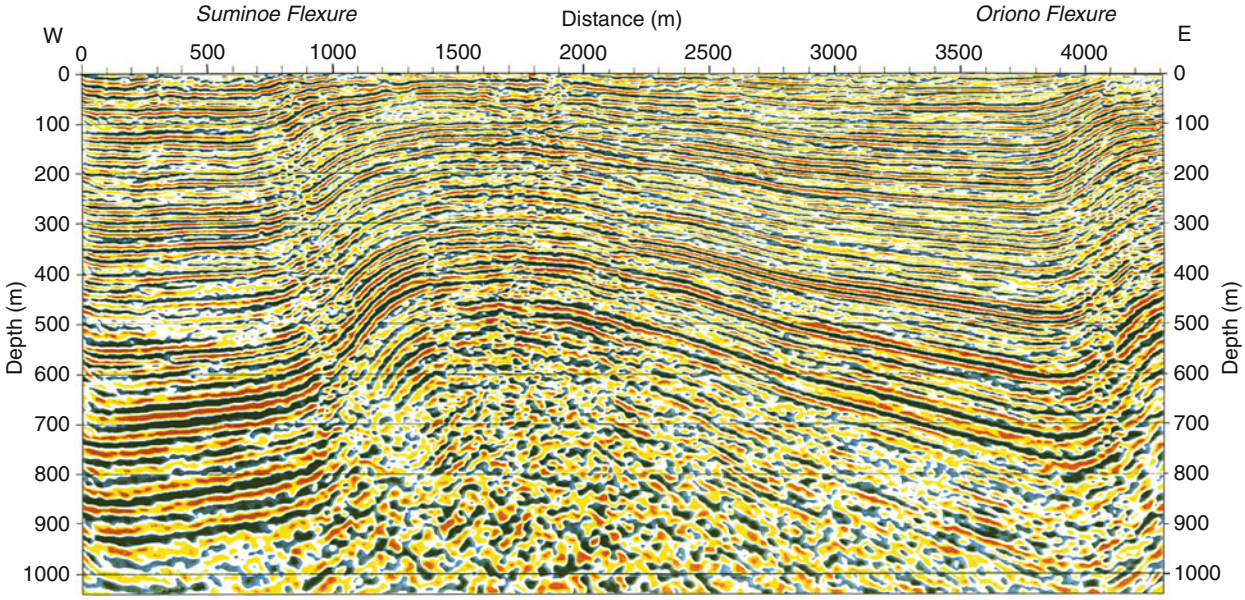
a fault lineament (Figure 1). Strike-slip faults accompany offset streams, shutter ridges, beheaded streams, wind gaps, and offset terrace risers. Multiple and systematic fault landforms along a single trace increase the reliability of its identification as an active fault trace. Most of such traces tend to be roughly along geologic faults juxtaposing different geologic units so that a traditional geologic reconnaissance is also useful to constrain an active fault zone. Fresh fault gouge (clay) and/or fault breccia are often exposed on outcrop along geomorphic traces. Geologic drilling and shallow coring are auxiliary approaches to constrain the location of a fault trace and sometimes help to estimate vertical offsets of stratigraphic units, particularly much deeper than trench excavation (e.g., Dolan et al., 2003). Deeper geologic information down to a few kilometers depth from industry oil wells provides a larger picture of fault structure and associated folds (e.g., Tsutsumi and Yeats, 1999). Several other methods reveal subsurface geologic structures for the areas having shallow water table and urban areas with limited open space. For example, cone penetrometer testing (CPT) and large diameter borings can be applied for these conditions (Grant et al., 1997). The "Geoslicer" (Nakata and Shimazaki, 1997), which consists of a sample tray and a shutter plate, can be inserted into the ground by a weighted vibrator to extract a sliced soil sample of larger than a few meters dimension (Takada and Atwater, 2004). Geoslicer can directly sample a fault itself without losing information of strike, dip, and geologic context of the fault.

Spatial resolution and accuracy in aerial photographic images are limited. Recently, airborne Light Detection And Ranging (LiDAR) surveys, which measure bare topography irradiating a laser beam from an airplane or helicopter, have been widely used to image very detailed fault-zone topography beneath dense vegetation cover (e.g., Sherrod et al., 2004) and urbanized areas (e.g., Kondo et al., 2008). These data are at the appropriate scale (meters) and accuracy (decimeter) to provide useful measurements of the fine features of fault landforms (Arrowsmith and Zielke, 2009). Airborne Laser Swath Mapping (ALSM) is also expected to find hidden active faults.

To locate an active fault trace and to confirm deformed subsurface structure various geophysical imaging techniques such as seismic reflection and refraction, ground-penetrating radar (GPR), aeromagnetic surveys, and gravity surveys are employed. The advantage of these techniques is that these are nondestructive methods for seeking deeper structures (Bond et al., 2007). High-resolution seismic reflection surveys are most pervasive to visualize shallow geologic structure associated with details of fault activity, even to reveal blind fault and fault-related fold (Figure 2).

These geomorphic, geologic, and geophysical explorations for fault identification also provide information about long-term slip rates. The slip rate on a fault, S, is calculated by dividing cumulative displacement (D) across the fault by time interval T (Figure 3),

$$S = D/T. \quad (1)$$



Paleoseismology, Figure 2 An example of fault-related fold revealed by a seismic reflection survey. Depth-converted P-wave seismic reflection profile across two flexures in south Osaka, Japan (Sugiyama et al., 2003). The strata are composed of Plio-Pleistocene non-marine and marine sediments.

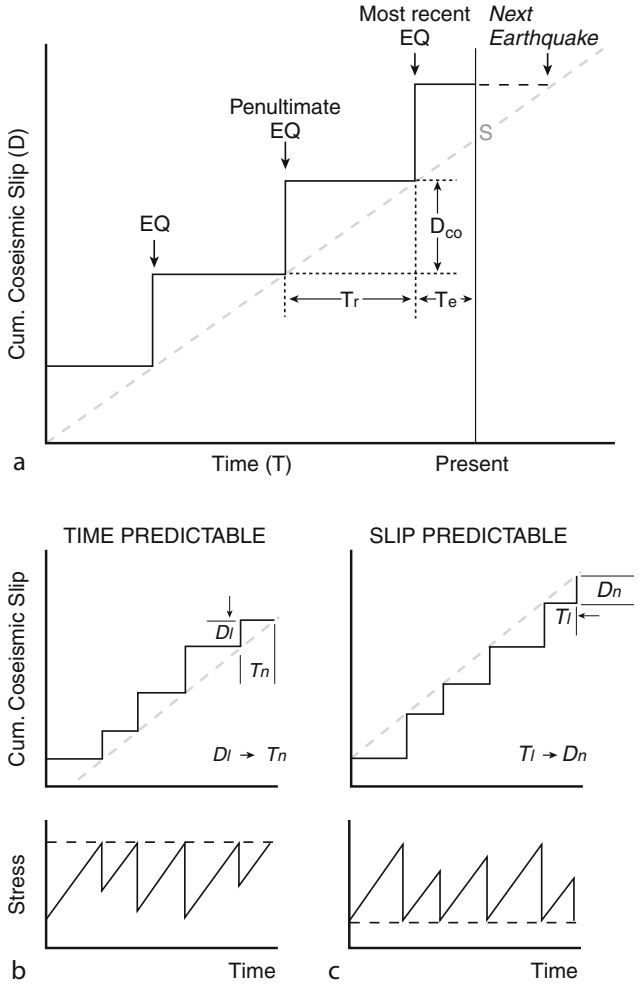
To obtain *S*, we need to find "piercing lines" or "piercing points" which intersect a fault zone as a line or a point. The piercing lines and points should be formed before the subsequent faulting events occurred. Geomorphologically, offset ridges, streams, and gullies become piercing lines, but more often well-reconstructed and dateable terraces and terrace risers are employed as the best piercing lines to estimate slip rates (e.g., L-L', M-M', M1-M1' in Figure 1). If offsets increase with terrace ages, the fault is considered to have been continuously active. Offset bedrock geology is also used as piercing lines and points to roughly estimate the very long-term slip rate. Cultural features such as roads, fences, walls, buried pipes, and rice paddy boundaries across a fault line are often used to measure the amount of coseismic slip for very recent earthquakes and historic events. In coastal and fluvial areas, where uplifted terraces are commonly seen, the vertical component of the slip rate *S*, is more easily obtained than the lateral component. Such slip rate estimates are used for classifying the fault activity (e.g., in Japan, class A: order of m/1,000 years; class B: order of 0.1 m/1,000 years; class C: order of 0.01 m/1,000 years), which is an important index for the estimation of the recurrence time of major earthquakes or for future seismic activity (Research Group for Active Faults of Japan, 1992).

Evidence for paleoearthquake

Paleoseismic evidence is either primary or secondary (McCalpin, 2009). Primary evidence is formed by tectonic deformation resulting from coseismic slip along a source fault plane. Secondary evidence is produced by earthquake shaking. Paleoseismic features are also distinguished as being on or above fault trace (on-fault or near-field features), or away from or far above a fault trace (off-fault or far-field features) (McCalpin, 2009). Because off-fault evidence for a paleoearthquake is mostly associated with strong shaking, the problem with the off-fault evidence is its difficulty in finding the corresponding seismic source. Thus, the majority of paleoseismic studies focus only on "on-fault" evidence.

Encountering a natural exposure of faulted young sediment is extremely rare. We thus excavate trenches across a fault zone (sometimes parallel to the fault zone) to expose faults and young strata that record evidence of paleoearthquakes. The site for trenches is chosen based on the fault location, land usage, sediments preservation, paleoearthquake records, and abundant dating material. The best trench site is considered as the one where: (1) deformation zone is narrow, or fault strand is single and representative of the entire fault zone; (2) frequent but dilute sedimentation processes occur with numerous dateable materials. The trench walls are not long-lived exposures and mostly not reproducible so that detailed observations and objective records by as many observers as possible are required (Grant, 2007). A trench log, which is a geologic sketch, is used for defining and mapping sedimentary units and recording evidence for surface-rupturing earthquakes, which we call an "event horizon."

An event horizon is used as the ground surface at the time of a paleoearthquake (Pantosti et al., 1993). The event horizon may be preserved as scarp-derived colluvial wedge deposits, unconformities that terminate faults, and warping (Figure 4), each of which is defined as an "event



Paleoseismology, Figure 3 (**a**) Schematic diagram showing the cumulative displacement (D) at a site along a fault as a function of time (T). D_{co} coseismic displacement, T_r recurrence interval, T_e elapsed time since the most recent earthquake, S slip rate estimated from D/T. (**b**) Time-predictable model. (**c**) Slip predictable model. After Shimazaki and Nakata (1980).

indicator" (Scharer et al., 2007). It is most likely to conclude that surface rupture and significant deformation such as folding and tilting along active traces are the direct products of seismogenic faulting. There are various kinds of evidences for paleoearthquakes which can be rated for their reliability.

Upward fault termination: This is the most common and easily identified event indicator in which faults and faulted strata are overlain by a non-deformed horizon (Figure 4a). But one tends to misinterpret an event horizon in special situations where fault slip near the ground surface tapers and dies out upward (Bonilla and Lienkaemper, 1990). Thus the quality and confidence level of this indicator depend on the amount of slip and stratigraphic conditions between different exposures at a site.

Fissure fills: Fissures and fissure fills are a sort of open crack associated with faulting and younger material filled during postseismic process (Figure 4b). They are commonly seen as a wide zone of alternating tension gashes and compressional mole tracks immediately after a large surface-rupturing earthquake due to strike-slip faulting.

Colluvial wedge: The colluvial wedge is formed on the footwall side as a result of collapse of the fault scarp exposed at the time of a scarp-forming earthquake (Figure 4c and d). A trench wall across a fault normally exposes a wedge-shaped section tapering toward or away from the fault scarp, often showing upward-coarsening structure, similar to talus deposition. When a bedrock fault plane is exposed at the earthquake, the collapsed sequence of gouge, breccia, and intact bedrock block are sequentially observed in the wedge. When it occurs within the loose sediments, scarp degradation is involved. Colluvial wedges are not only seen along normal faults but also along strike-slip and thrust faults.

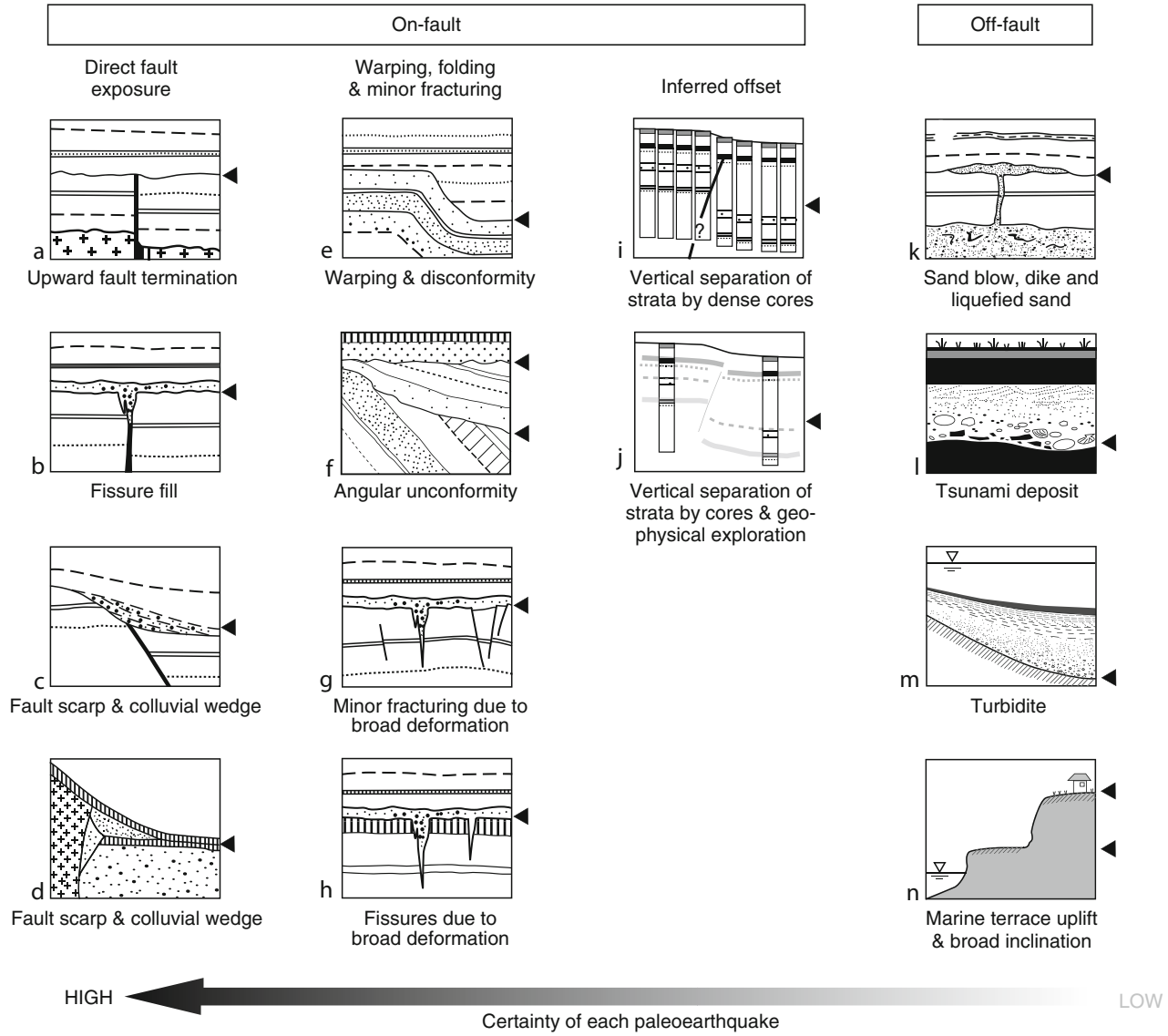
Angular unconformity associated with tilted and folded strata: An angular unconformity is produced by broad deformation of subsurface sediments due to faulting and subsequent erosion and sedimentation (Figure 4e and f). If a fault plane is simple and it is hard to find an upward termination, angular unconformity would be a substitutive criterion to identify an event horizon. But, except for significant inclination of strata, horizontal sedimentation must be guaranteed. Adhesive composition with clay and silt often shows a veneer deposit along a slope, which often mimics crustal deformation.

Minor cracks and faults associated with subtle deformation: Where the amount of displacement is small, and/or surface condition is dry, numerous minor cracks or tension fissures associated with deformation can be produced (Figure 4g and h).

The deeper we expose stratigraphic units, the more events and associated structures are likely to be exposed. Since evidence associated with older events tends to be destroyed and obscured by the most recent event, we must restore the stratigraphy and structures prior to the most recent event and then reinterpret the older events to identify multiple paleoearthquakes at a site ("Retrodeformation analysis," McCalpin, 2009). It means that reliability of the paleoearthquake tends to be reduced when we go farther back in the past. To have higher confidence for a paleoearthquake, multiple exposures of a fault and/or significant deformation (event indicators) in the same stratigraphic horizon should be confirmed at a site (Scharer et al., 2007).

Time constraints of paleoearthquakes

Although there are several methods to directly determine ages of fault activities from fault gouges (e.g., Fukuchi, 2001), timing of each paleoearthquake is commonly constrained by ages of deformed stratigraphic units and overlying undeformed strata, for which many dating



Paleoseismology, Figure 4 Schematic diagrams of typical stratigraphic event indicators of paleoearthquakes in both on-fault and off-fault regions. Solid triangle beside each panel indicates an event horizon. Several on-fault diagrams are based on Allen (1986) and Okada (1991). Note that confidence level of paleoearthquakes increases from the diagrams in off-fault regions to the ones showing direct fault exposure.

approaches using different isotopes have been used (see details in Grant, 2002 and references therein).

Among them, radiocarbon dating is the most powerful and pervasive technique in paleoseismic studies. In this technique, the time since the death of a plant or animal is estimated by measuring the sample's residual ^{14}C content relative to modern standards and the half-life decay constant of ^{14}C (5,568 years). The beta particle counting method conventionally requires 2–200 g of carbon and allows us to date back to ~40,000 years before present (yBP; actually years before 1950 AD). Meanwhile, newly developed Accelerator Mass Spectrometry (AMS) dating requires samples as small as a few milligrams and extends the measurable date back to 50,000–60,000 yBP. Since the ratio of ^{14}C to ^{12}C has varied with time due to variations in number of cosmic rays associated with the strength of the Earth's magnetic shield, measured radiocarbon age needs to be converted to calendar date. From dendrochronological approaches with tree rings, glacial varves, corals that annually record the amount of ^{14}C in the atmosphere, radiocarbon ages are calibrated to corresponding calendar dates back to about 24,000 years before present (Stuiver et al., 1998).

Such dendrochronological approaches to constrain timing of a paleoearthquake have been commonly employed since Jacoby (1989). Time constraint of

a paleoearthquake requires at least two radiocarbon dates that bracket an event horizon, in which probabilistic density is calculated from the calibrated dates of both samples. To have better constraints for paleoearthquakes, there are several calibration programs, one of which "OxCal" (Bronk Ramsey, 2001; Lienkaemper and Bronk Ramsey, 2009) not only calibrates the radiocarbon ages to the calendar years but also implements rigorous means of estimating these event ages and their uncertainties using Bayesian statistics.

As mentioned before, several countries have historical accounts longer than a thousand years or more and thus may cover more than a few seismic cycles of very active faults. However, the longer we go back in the past, the fewer and less reliable the historical accounts are. In addition, most of the historical earthquakes roughly control isoseismal areas from documents of earthquake damage, but not the specific source of each earthquake. Thus a combination of such historical documents and paleoseismic evidence with well-determined dates are used to confirm the source fault for such historical earthquakes (e.g., Armijo et al., 2005).

Archaeoseismology reaches further back in the past to reveal prehistoric earthquakes based on their effects on man-made objects, usually buildings or other monuments (Karcz and Kafri, 1978, see entry on *Archaeoseismology*). The time range depends on history and archaeology in each country. Historic accounts are often available and tied with ruins due to devastating shaking, paleo-liquefactions, and paleo-landslides (rock-slides) at archaeological sites. The advantage of archaeoseismology is narrow time constraints of earthquake occurrence by archaeological and historical knowledge. Characteristic off-fault damage types include horizontal shifting of large building blocks, downward sliding of one or several blocks from masoned arches, collapse of heavy, stably-built walls, chipping of corners of building blocks, and aligned falling of walls and columns (Marco, 2008). In some rare cases, faults intersect with archaeological sites, which enables us to measure sense and size of slip and constrain its occurrence time. Several recent archaeoseismological studies in the Eastern Mediterranean region in particular have been unveiling the detailed slip history and rupture behavior of the Dead Sea fault (e.g., Meghraoui et al., 2003), and large earthquakes associated with mythology along the Aegean Extensional province, Cyprus, and Hellenic Arcs (e.g., Stiros, 2001).

Magnitude and frequency of paleoearthquakes

Like "uniformitarianism," no paleoearthquakes can be inferred without detailed knowledge of recent observed earthquakes that left important clues between subsurface seismic processes and surface deformation. Since paleoseismology relies on paleo-surface information preserved in geology and landforms, seismic and geologic data compilations of recent large shocks associated with surface-ruptures, tilt, warping, and other earthquake-related deformation are necessary.

As seismologists found simple self-similar scaling laws for source faults, dimension and magnitude, surface ruptures compiled from worldwide data have similar scaling relations (Wells and Coppersmith, 1994). For example, the empirical relation between surface rupture length (L in km) and magnitude (M) is expressed as $M = 5.16 + 1.12 \log L$ for strike-slip fault. For seismic hazard assessment, such an empirical relation is increasingly applied to estimate magnitude from the length of an identified fault trace. Similar empirical scaling relations are found between coseismic surface displacement and magnitude. For maximum displacement D_{max} in meter, $M = 6.69 + 0.74 \log D_{max}$. For average displacement, D_{ave} in meters, $M = 6.93 + 0.82 \log D_{ave}$. Besides fault length, coseismic slip, either D_{max} or D_{ave}, is also increasingly used to estimate the magnitude of a paleoearthquake. There are two ways to have such coseismic slip from paleoseismic data. One gives an indirect estimate from slip rate (S) and average recurrence interval (T_r) of paleoearthquakes along a fault as $D = S \times T_r$. The other provides direct measurement from vertical separation of faulted strata on trench walls in particular for reverse faults and three-dimensional reconstruction of lateral offset of piercing points for strike-slip faults.

In terms of magnitude and frequency estimates for an active fault, one of the greatest impacts of geology on seismology is the "characteristic earthquake model" (Schwartz and Coppersmith, 1984, see entry on *Characteristic Earthquakes and Seismic Gaps*). This model illustrates the concept of the repetitions of similar size of large earthquakes with a similar amount of slip, rupture length, and magnitude along a fault. The model also emphasizes that paleoseismic records from a fault suggest more frequent occurrence of large earthquakes compared to the ones extrapolated from the Gutenberg–Richter relation of regional seismicity (Wesnousky, 1994). If only scanty paleoseismic information on a fault is available, one simply uses the characteristic earthquake model to estimate the size of earthquakes potentially caused by the fault.

If the characteristic earthquakes on a fault occur periodically or quasi-periodically with characteristic slip, we would not have any difficulty to approximately forecast the time and magnitude for the next large earthquake (Figure 3a). However, as summarized below, most of the recurrences of earthquakes are not periodic and may be influenced by recent rupture history. Regarding the slip-time relation at a site along a fault, under the assumption of constant long-term slip rate, two conceptual models are proposed (Shimazaki and Nakata, 1980). The "time-predictable model" can predict the time of the next earthquake from a knowledge of the previous slip (Figure 3b), whereas "slip predictable model" cannot predict the time but amount of slip, taking the slip deficit since the last one into account (Figure 3c). In reality, however, at least 30 prehistoric earthquakes in a 6,000-year-long record at Wrightwood along the San Andreas fault, southern California, follow neither time-predictable nor slip-predictable model (Weldon et al., 2004). The

paleoseismic history might prove that slip occurs at a wide range of stress levels and does not return to the same postseismic stress level, keeping quasi-periodic recurrence behaviors (Scharer et al., 2010).

To furnish the magnitude estimate using the length of a fault, one must define the extent of a fault. In nature, we rarely encounter an isolated long single fault trace. Instead, we more often map a group of faults or a network of faults, which require us to define each seismogenic fault. To subdivide faults into portions, the concept of "fault segmentation" is employed. "Fault segment" is a general term for the portions of a fault or fault system or fault zone. There are several types of fault segments ordered in terms of earthquake occurrence; *earthquake segment*, *behavioral segment*, *structural segment*, *geologic segment*, and *geometric segment* (McCalpin, 2009). Identifying and defining earthquake segments that directly determine the size of earthquakes is the ultimate goal. However, numerous studies started to focus on changes in fault orientation and boundaries of fault traces (branches, intersections, separations, and step-overs) to identify each structural and geometric fault segment. There are also varieties of scale of segment and "discontinuous gaps." But since we assess seismogenic fault, kilometer-scale traces are generally taken into account. To identify each seismogenic fault among many complex mapped fault traces, "5-km separation criteria" are now broadly employed for the hazard assessment (e.g., Earthquake Research Committee in Japan (ERC), 2005). Several studies done in the late 1980s and early 1990s compiled the surface ruptures, associated historic earthquakes, and mapped active traces around the globe and then concluded that rupture propagation beyond the 5-km step width is highly unlikely (Barka and Kadinsky-Cade, 1988; Wesnousky, 1988; Wesnousky, 2006). Although these should be case-dependent with many factors, "5 km" is still a standard separation distance to identify each seismogenic fault segment.

A long continuous fault (system) cannot be simply divided by the 5-km separation criterion instead. It generates a variety of rupture patterns and sizes of earthquakes. Several conceptual models are proposed with slip rates and event age constraints at several sites along a single fault or a single fault system. The characteristic earthquake model is one of the several conceptual models. In reality, a space-time diagram of fault rupture patterns estimated from intensive paleoseismic surveys is not so simple to apply the conceptual model. Detailed examples from the southern San Andreas fault (Figure 5, Weldon et al., 2004) and the North Anatolian fault, Turkey (Hartleb et al., 2003) show the complicated rupture patterns on a fault system.

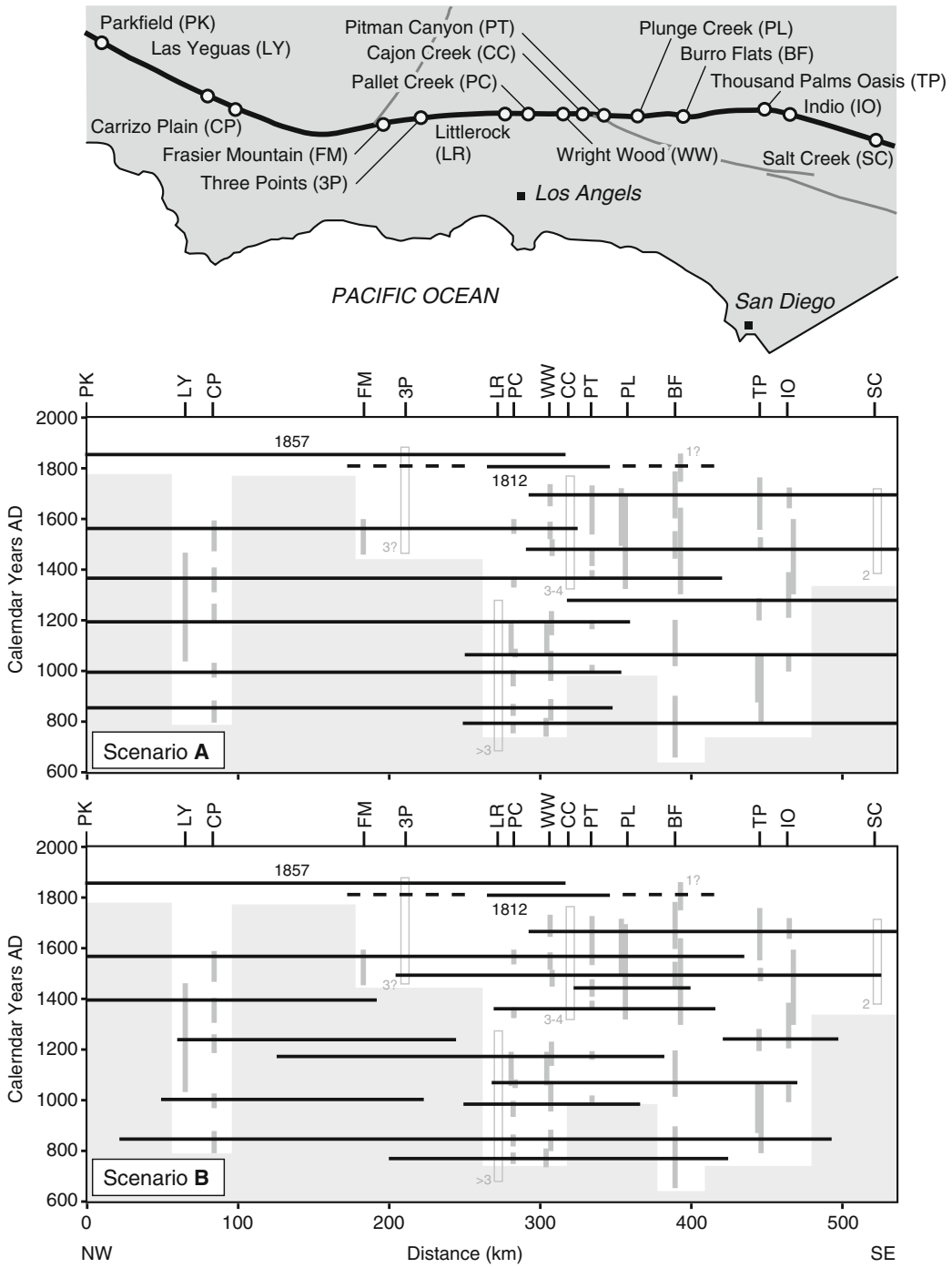
Off-fault paleoseismology

Although a limitation of off-fault evidence is its difficulty in identifying the seismic source, it becomes supportive evidence for paleoearthquakes, together with on-fault evidence. There are mainly two types of off-fault evidence for paleoearthquakes: one is due to strong shaking, and the other is related to features associated with broadly distributed coseismic and postseismic deformation (see *Earthquakes and Crustal Deformation*). The former triggers liquefaction, terrestrial and submarine landslides, and tsunami. The latter leaves uplifted, tilted, and subsided ground surfaces recorded on shorelines and marine and river terraces, where paleo-altitude can be estimated and a flat surface is guaranteed. Besides on-fault geology and geomorphology, preservation of such paleoseismic evidence is necessary. In addition, synchronicity and spatial extent of off-fault events are keys to distinguish paleoearthquakes from other causes of such secondary geologic features.

Liquefaction and associated features such as sand blows, sand boils, sand volcanoes, lateral spreading, clastic dikes, hydraulic fracturing, and rarely gravel blows and gravel dikes are the products of strong ground acceleration due to a large earthquake. Worldwide data on historical earthquakes show that the features having a liquefaction origin can be developed at earthquake magnitudes as low as about 5, but that a magnitude of about 5.5–6 is the lower limit at which liquefaction effects become relatively common (Ambraseys, 1988; Obermeier, 2009). Therefore, paleo-liquefaction is strong evidence for large earthquakes in the past. Although it is difficult to excavate trenches or pits targeted only for paleo-liquefaction, we often encounter evidence for paleo-liquefaction at archaeological sites, which can constrain the earthquake occurrence time much better than ^{14}C date due to cover-up by the non-damaged younger ruins and horizons (e.g., Figure 6, Sangawa, 2010). Since liquefaction can be commonly repeated at the same site where the recent large earthquake occurred, we can sometimes find the paleo-liquefaction on trench walls and confirm the recurrence of strong shaking at the site (e.g., Wesnousky and Leffler, 1992).

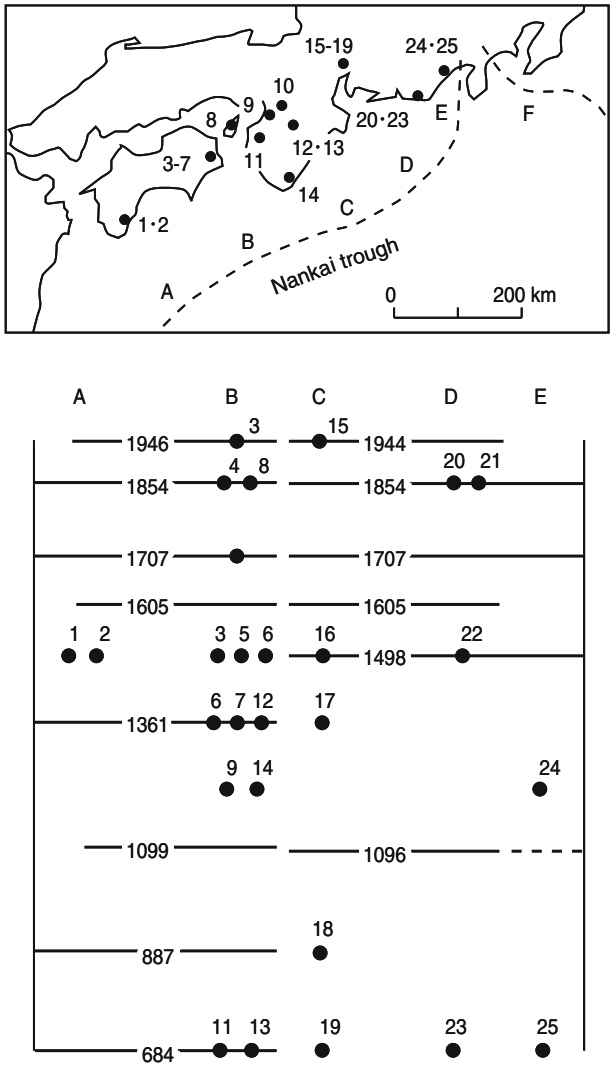
Strong ground motions in mountainous regions often trigger slope failures. In the regions characterized by steep unstable topography, we can identify massive landslides and scarps occurred in the past in aerial photographs. However, interpreting an earthquake-induced landslide or group of landslides excluding other factors like heavy rainfall is particularly hard, and levels of confidence in the resulting interpretation vary widely. Comprehensive geologic surveys with dating of landslides and several criteria (Jibson, 2009) that discriminate landslides from other factors are required.

Widespread thin sediments in coastal regions reveal occurrence of paleo-tsunamis. Sand layer often containing pebble and shell deposit is extensively deposited by tsunami waves and its inundation on the coastal areas (see *Tsunami*). Although there is no single convenient feature to distinguish tsunami deposits from those generated by other events, sedimentary facies analysis with several key features is indispensable for identifying tsunami deposit (Shiki et al., 2008). To better identify paleo-tsunami deposits, there are limited conditions to preserve



Paleoseismology, Figure 5 Two possible rupture sequences on the southern San Andreas fault (Weldon et al., 2004). Vertical grey bars are ranges in age for earthquakes at the sites listed at the lower margin and horizontal bars are rupture lengths. Open boxes represent multiple events age ranges; the individual event ages are unknown. Grey shading indicates regions and times with no data.

such sand layers; paleo-tsunami deposits are distinguished from black humus and mud deposits if the paleo-tsunami reaches the areas in swales, marsh and other calm environments where no sand sources can normally flow. Numerous studies, which employed shallow coring, pits, and trenches, exposed tsunami sand sheets intervened by finer and datable organic soils (e.g., Atwater, 1987). The recent significant contribution of the tsunami deposit studies to the subduction process is to find prehistoric M9 class earthquakes and their recurrence intervals along the 2004



Paleoseismology, Figure 6 History of mega-thrust earthquakes along the Nankai trough, Japan (Sangawa, 2010). The subduction zone is divided into five rupture segments A–E (F is the source of 1923 $M = 7.9$ Kanto earthquake along the Sagami trough). The horizontal lines with number in lower panel indicate the rupture extents and occurrence years of large earthquakes. Dots with number denote archaeological sites with exposed significant liquefactions at specific ruins and their calendar years. Recent archaeo-seismological data filled gaps in long recurrence intervals originally estimated by only historic accounts (e.g., possible rupture on A and B segments at the 1,498 earthquake), and making the Nankai subduction earthquake cycles more regular and frequent. Tsunami deposits recovered from coastal lakes, lagoons, swamps, and lowlands are also filling the gaps (e.g., Takada et al., 2002) and also provide much longer rupture history back to the entire Holocene period (Komatsubara and Fujiwara, 2007).

Sumatra-Andaman (Jankaew et al., 2008), and along the Kuril trench (Nanayama et al., 2003), and relation with the M8 class predecessors of the 1960 M9.5 Chile earthquake (Cisternas et al., 2005), which breaks through the conventional idea on regular occurrences of subduction earthquakes (see *Great Earthquakes*).

Besides paleo-tsunami deposits, submarine turbidites are another geologic clue to reveal large submarine and coastal earthquakes. Goldfinger et al. (2007) demonstrated their validity and significance of sediments associated with earthquake-triggered turbidity currents by analyzing numerous piston cores collected from submarine channels and canyon systems draining the northern California continental margin. By studying synchronous triggering of the turbidity currents, they found 15 turbidites during the last ~2,800 years, equivalent to an average repeat time of ~200 years which is similar to the onshore value of ~230 years along the northern San Andreas fault. Along-strike correlation also suggests the ~320-km rupture extent of the most of the younger events.

Broad deformation of off-fault areas is another indicator for paleoearthquakes (see *Earthquakes and Crustal Deformation*). Vertical deformation can induce changes in local rates of deposition and erosion that provide evidence of a paleoearthquake, particularly in fluvial and coastal environment (McCalpin and Carver, 2009). Subduction associated mega-thrust earthquakes and dip-slip intraplate earthquakes largely change horizontal markers such as shorelines prior to an earthquake and record off-fault coseismic uplift (e.g., Meghraoui et al., 2004) which are preserved as uplifted terraces and shorelines. Coastal areas facing subduction zones, such as southwest Japan, Taiwan, New Zealand, and Crete in Greece, have several steps of elevated terraces and shorelines associated with the mega-thrust earthquakes (see *Seismicity, Subduction Zone*). But significant coseismic uplifts are also inferred to have been formed by shallow near-coast offshore active faults that might be branched from subduction interface (e.g., Maemoku, 2001). Although dating paleo-marine terraces and shorelines is not easy, coral reefs, coral microatolls (e.g., Zachariasen et al., 1999), fossils of sessile organisms inhabit intertidal level and become both paleo-shoreline markers and indicators of timing of earthquake occurrences. For older terraces over 50 ka, other dating techniques and regional tephra fallouts are used to estimate the emergent dates. In Japan, altitudes of uplifted marine terraces formed at marine oxygen isotope stage 5e (~120 ka) often provide us with the best estimates of long-term uplift rates associated with interplate events (Koide and Machida, 2001). In contrast to coseismic uplift, studies for paleoseismic subsidence are much fewer. Subsidence associated with a subduction earthquake normally occurs inland from the rebounded uplift zone. Here, buried soil of subsided marsh and forest records the sudden vertical deformation (e.g., Atwater et al., 1995).

Permanent deformation of landscapes associated with seismogenic faulting is sometimes modified or amplified by postseismic crustal movement. Recent mega-thrust subduction events caused not only coseismic uplift, but also postseismic transient deformation that lasts more than

a few decades (see *Earthquakes and Crustal Deformation*). Although contribution of postseismic deformation associated with paleoearthquakes is hardly measured in the uplifted terraces, Sawai et al. (2004) used fossil diatom assemblages to infer the gradual changes in sedimentary environments corresponding to the transient postseismic deep creep that occurred in the seventeenth century along the Kuril subduction zone.

Seismic hazard assessment based on paleoseismology

Paleoseismic data are directly used for making probabilistic seismic hazard maps (see *Seismic Hazard*) which incorporate information on fault location, dimension, geometry, largest earthquake from such dimension, slip rates, and recurrence intervals (e.g., Working Group on California Earthquake Probabilities, 1995). Two types of statistical models of earthquake occurrence are commonly used to estimate earthquake probability: a stationary Poisson model and a conditional quasi-periodic model (see *Statistical Seismology*). The former is used for active faults or active seismic regions where only frequency of large earthquakes is available. The latter considers fluctuations of repeat times (coefficient of variation in recurrence time) and elapsed time since the most recent event (see Figure 3), which is time-dependent and compatible with the strain accumulation process. Several probability density functions such as normal, lognormal, Weibull, or gamma distributions are used for the recurrence process. Among them, the Brownian Passage Time function (Matthews et al., 2002) is evaluated as so far the best function because of its capability to account for the stress perturbations and has been adopted for the probabilistic seismic hazard studies in California and Japan (Field et al., 2009; ERC, 2005). Time-dependent seismic hazard considering conditional probabilities on most major active faults provides us with more realistic estimates (Fujiwara et al., 2009).

If further detailed paleoseismic information is available for a specific fault, a "scenario earthquake model (seismic hazard maps for specified seismic sources)" is often proposed to show the strong shaking for the areas being assessed when the specific earthquake occurs (ERC, 2005). When providing paleoseismic data to seismic engineers, possible rupture patterns for the next earthquake, slip distribution, and plausible locations of asperities becomes additional information (Somerville et al., 1999).

In addition to the shaking damage, surface rupture and displacement associated with a large earthquake often destroy man-made structures. Structures for human occupancy (e.g., hospital and school) and critical buildings (e.g., nuclear power plant) can be so located to avoid the traces of active faults (e.g., the Alquist-Priolo Earthquake Fault Zoning Act in California, see entry on *Seismic Zonation*). To mitigate direct damage due to surface-faulting, detailed positions of active faults and their paleoseismological properties are progressively been opened to the public through the internet.

Issues in paleoseismology and perspectives

There are fundamental limitations of the geologic approach to infer paleoearthquakes in time and space. The following are major caveats in using paleoseismological information:

1. Surface ruptures are commonly associated with $M > 6$ earthquakes (Wells and Coppersmith, 1994) that occur at shallow depths (mostly <20 km), even though $M \sim 5$ earthquakes may also leave surface ruptures in the case of extremely shallow hypocenter (<5 km) and volcanic environments (Payne et al., 2009, see entry on *Earthquakes, Volcanogenic*). However, we must know that geomorphic and stratigraphic evidence for all $M > 6$ paleoearthquakes is not always produced. It totally depends on regional tectonic setting, depth of hypocenter (thickness of seismogenic crust), fault slip sense, and surface geologic conditions. For example, revisiting the inland earthquakes in Japan since 1923, no more than half of the $M \sim 7$ earthquakes left the surface breaks comparable to the seismic faults, which obviously underestimates the probabilities of even $M \sim 7$ shocks (Toda and Awata, 2008). Even along major strike-slip faults, we experienced enigmatic large earthquakes that did not leave clear surface rupture (e.g., 1989 Loma Prieta earthquake and 2010 Haiti earthquake).
2. Rates of erosion and sedimentation at regional and local scales control preservations of fault landforms and evidence for paleoearthquakes at excavation sites. Human destruction of the surface geology and topography due to urbanization or farming often destroys evidence of very recent paleoseismic events. Furthermore, obtaining better time constraints for paleoearthquakes is rare due to the lack of abundant dating material.
3. Geologic and topographic evidence normally includes interpretation with epistemic uncertainties. Complete agreement is rare in aerial photographic interpretation. "Trench parties" (Grant, 2007) often create indecisive debate that is not visible in the final publication in which simple trench logs are only illustrated. Photos and detailed sketches are objective records, but they are infrequently opened to the public and are hardly accessible. Therefore one should be aware that publications often underestimate or overestimate numbers of paleoearthquakes and need extra effort to clarify quality measurements and other paleoseismic interpretations.
4. Paleoseismic study cannot precisely reveal rupture extent, thus magnitude of an earthquake without any help of historic accounts. Space-time diagrams of paleoseismic events on a fault or fault system may only conclude the cases adjacent segments did not simultaneously rupture through. Displacement data of a paleoearthquake would be useful information to estimate magnitude based on the empirical scaling law (Wells and Coppersmith, 1994). But at least five to ten measurements are stochastically required to

characterize magnitude of a paleoearthquake (Hemphill-Haley and Weldon, 1999).

5. Although we simply apply the characteristic earthquake model or extended segmentation model under the elastic rebound framework (see *Earthquakes and Crustal Deformation*), there may be intrinsic uncertainties of repeated faulting processes, for example, nonlinear strain accumulation and complex rupture process. This might be a subject seismologists have to reveal. But paleoseismology instead can be suggestive with data that also conquer "uniformitarianism." For example, the 2004 $M = 9.2$ Sumatra-Andaman earthquake was an extraordinary earthquake which can be hardly expected from historic accounts. However, intensive detailed studies about coral microatolls (Zachariasen et al., 1999) and tsunami deposits (Jankaew et al., 2008) suggested there were frequent huge events in the past, even if we do not see them in historic ages. This is an important role of paleoseismology.

Summary

Paleoseismology, which is the study of prehistoric or pre-instrumental earthquakes, is an interdisciplinary field of research mainly encompassing seismology, geology and geomorphology. Starting from identifying active faults, various field and laboratory techniques such as aerial photography interpretation, trench excavation, and radiocarbon dating are used to reveal the spatial extent and occurrence times of paleoearthquakes, and associated long-term movements of active faults. Evidence for paleoearthquakes is not only preserved in strata across a fault but also found far from the source as paleo-liquefaction, paleo-landslide, and paleo-tsunami deposits. Empirical relations between rupture length and slip versus earthquake magnitude from recent instrumentally recorded shocks enable us to estimate the size of future earthquakes on a fault or fault zone. Detailed rupture history of a fault also allows us to estimate time-dependent probabilistic seismic hazard. Although there are fundamental limitations and incompleteness in field data, contribution of paleoseismology in understanding the long-term faulting process is enormous and indispensable. Recent progress in paleoseismic data accumulations also allows seismologists and seismic engineers to make better seismic hazard maps to mitigate seismic damage in the future.

Bibliography

Allen, C. R., 1986. Seismological and paleoseismological techniques of research in active tectonics. In Wallace, R. E. (ed.), *Active Tectonics: Studies in Geophysics*. Washington, DC: National Academy Press, pp. 148–154.

Ambraseys, N., 1988. Engineering seismology: earthquake engineering and structural dynamics. *Journal of International Association of Earthquake Engineering*, **17**, 1–105.

Armijo, R., et al., 2005. Submarine fault scarps in the Sea of Marmara pull-apart (north Anatolian Fault): implications for seismic hazard in Istanbul. *Geochemistry, Geophysics, Geosystems*, **6**, Q06009, doi:10.1029/2004GC000896.

Arrowsmith, J. R., and Zielke, O., 2009. Tectonic geomorphology of the San Andreas Fault zone from high resolution topography: An example from the Cholame segment. *Geomorphology*, **113**, 70–81.

Atwater, B. F., 1987. Evidence of great Holocene earthquakes along the outer coast of Washington State. *Science*, **236**, 942–944.

Atwater, B. F., Nelson, A. R., Clague, J. J., Carver, G. A., Bobrowsky, P. T., Bourgeois, J., Darienzo, M. E., Grant, W. C., Hemphill-Haley, E., Kelsey, H. M., Jacoby, G. C., Nishenko, S. P., Palmer, S. P., Peterson, C. D., Reinhart, M. A., and Yamaguchi, D. K., 1995. Summary of coastal geologic evidence for past great earthquakes at the Cascadia subduction zone. *Earthquake Spectra*, **11**, 1–18.

Barka, A., and Kadinsky-Cade, K., 1988. Strike-slip fault geometry in Turkey and its influence on earthquake activity. *Tectonics*, **7**, 663–684.

Berryman, K. R., Ota, Y., and Hull, A. G., 1989. Holocene paleoseismology in the fold and thrust belt of the Hikurangi subduction zone, eastern North Island, New Zealand. *Tectonophysics*, **163**, 185–195.

Bond, C. E., Gibbs, A. D., Shipton, Z. K., and Jones, S., 2007. What do you think this is? "Conceptual uncertainty" in geoscience interpretation. *GSA Today*, **17**, 4–10.

Bonilla, M. G., and Lienkaemper, J. J., 1990. Visibility of fault strands in exploratory trenches and timing of rupture events. *Geology*, **18**, 153–156.

Bronk Ramsey, C., 2001. Development of the radiocarbon calibration program OxCal. *Radiocarbon*, **43**, 355–363.

Cisternas, M., Atwater, B. F., Torrejon, F., Sawai, Y., Machuca, G., Lagos, M., Eipert, A., Youlton, C., Salgado, I., Kamataki, T., Shishikura, M., Rajendran, C. P., Malik, J. K., Rizal, Y., and Husni, M., 2005. Predecessors of the giant 1960 Chile earthquake. *Nature*, **437**, 404–407.

Dolan, J. F., Christofferson, S. A., and Shaw, J. H., 2003. Recognition of paleoearthquakes on the Puente Hills blind thrust fault, California. *Science*, **300**, 115–118.

Earthquake Research Committee, 2005. Report: 'National Seismic Hazard Maps for Japan (2005)', Headquarters for Earthquake Research Promotion, 162 pp. Available from World Wide Web: http://www.jishin.go.jp/main/index-e.html.

Field, E. H., Dawson, T. E., Felzer, K. R., Frankel, A. D., Gupta, V., Jordan, T. H., Parsons, T., Petersen, M. D., Stein, R. S., Weldon, R. J., II, and Wills, C. J., 2009. Uniform California earthquake rupture forecast, version 2 (UCERF 2). *Bulletin of the Seismological Society of America*, **99**, 2053–2107.

Fujiwara, H., Morikawa, N., Ishikawa, Y., Okumura, T., Miyakoshi, J., Nojima, N., and Fukushima, Y., 2009. Statistical comparison of national probabilistic seismic hazard maps and frequency of recorded JMA seismic intensities from the K-net strong-motion observation network in Japan during 1997–2006. *Seismological Research Letters*, **80**, 458–464.

Fukuchi, T., 2001. Assessment of fault activity by ESR dating of fault gouge; an example of the 500 m core samples drilled into the Nojima earthquake fault in Japan. *Quaternary Science Reviews*, **20**, 1005–1008.

Goldfinger, C., Morey, A. E., Nelson, C. H., Gutierrez-Pastor, J., Johnson, J. E., Karabanov, E., Chaytor, J., Eriksson, A., and Shipboard Scientific Party, 2007. Rupture lengths and temporal history of significant earthquakes on the offshore and north coast segments of the Northern San Andreas Fault based on turbidite stratigraphy. *Earth and Planetary Science Letters*, **254**, 9–27.

Grant, L. B., 2002. Paleoseismology. In Lee, W. H., Kanamori, H., Jennings, P. C., and Kisslinger, C. (eds.), *IASPEI International Handbook of Earthquake and Engineering Seismology.*

New York: International Association of Seismology and Physics of the Earth's Interior Press, Vol. 81A, pp. 475–489.
Grant, L. B., 2007. Historical seismicity – paleoseismology. In Schubert, G. (ed.), *Treatise of Geophysics*. Amsterdam: Elsevier, Vol. 4, pp. 567–589.
Grant, L. B., Waggoner, J. T., von Stein, C., and Rockwell, T. K., 1997. Paleoseismicity of the North Branch of the Newport–Inglewood Fault zone in Huntington Beach, California, from Cone penetrometer test data. *Bulletin of the Seismological Society of America*, **87**, 277–293.
Hartleb, R. D., Dolan, J. F., Akyuz, H. S., and Yerli, B., 2003. A 2000-year-long paleoseismologic record of earthquakes along the central north Anatolian Fault, from trenches at Alayurt, Turkey. *Bulletin of the Seismological Society of America*, **93**, 1935–1954.
Hemphill-Haley, M. A., and Weldon, R. J., II, 1999. Estimating prehistoric earthquake magnitude from point measurements of surface rupture. *Bulletin of the Seismological Society of America*, **89**, 1264–1279.
Jacoby, G. C., 1989. Application of tree-ring analysis to paleoseismology. *Geology*, **17**, 226.
Jankaew, K., Atwater, B. F., Sawai, Y., Choowong, M., Charoentitirat, T., Martin, M. E., and Prendergast, A., 2008. Medieval forewarning of the 2004 Indian Ocean tsunami in Thailand. *Nature*, **455**, 1228–1231.
Jibson, R. W., 2009. Using landslides for paleoseismic analysis. In McCalpin, J. P. (ed.), *Paleoseismology.* Academic Press, pp. 565–601.
Karcz, I., and Kafri, U., 1978. Evaluation of supposed archeoseismic damage in Israel. *Journal of Archaeological Science*, **5**, 237–253.
Koide, K., and Machida, H. (eds.), 2001. *Atlas of Quaternary Marine Terraces in the Japanese Islands*. Tokyo: University of Tokyo Press.
Komatsubara and Fujiwara, 2007. Overview of Holocene tsunami deposits along the Nankai, Suruga, and Sagami troughs, southwest Japan. *Pure and Applied Geophysics*, **164**, 493–507.
Kondo, H., Toda, S., Okumura, K., Takada, K., and Chiba, T., 2008. A fault scarp in an urban area identified by LiDAR survey: a case study on the Itoigawa-Shizuoka Tectonic Line, central Japan. *Geomorphology*, **101**, 731–739.
Lienkaemper, J. J., and Bronk Ramsey, C., 2009. OxCal: versatile tool for developing paleoearthquake chronologies – a primer. *Seismological Research Letters*, **80**, 431–434.
Lin, A., Guo, J., Kano, K., and Awata, Y., 2006. Average slip rate and recurrence interval of large-magnitude earthquakes on the western segment of the strike-slip Kunlun Fault, northern Tibet. *Bulletin of the Seismological Society of America*, **96**, 1597–1611.
Maemoku, H., 2001. Reexamination of coseismic uplift of Cape Muroto, southwest Japan, using AMS ^{14}C ages of raised sessile organisms. *Journal of Geography*, **110**, 479–490 (in Japanese).
Marco, S., 2008. Recognition of earthquake-related damage in archaeological sites: examples from the Dead Sea fault zone. *Tectonophysics*, **453**, 148–156.
Matthews, M. V., Ellsworth, W. L., and Reasenberg, P. A., 2002. A Brownian model for recurrent earthquakes. *Bulletin of the Seismological Society of America*, **92**, 2233–2250.
McCalpin, J. P. (ed.), 2009. *Paleoseismology*, 2nd edn. Academic Press. International Geophysics Series, Vol. 95, 613 pp.
McCalpin, J. P., and Carver, G. A., 2009. Paleoseismology of compressional tectonic environments. In McCalpin, J. P. (ed.), *Paleoseismology.* Academic Press, pp. 315–419.
Meghraoui, M., Gomez, F., Sbeinati, R., Van der Woerd, J., Mouty, M. N., Darkal, A., Radwan, Y., Layyous, I., Al-Najjar, H., Darawcheh, R., Hijazi, F., Al-Ghazzi, R., and Barazangi, M., 2003. Evidence for 830 years of seismic quiescence from paleoseismology, archaeoseismology and historical seismicity along the Dead Sea fault in Syria. *Earth and Planetary Science Letters*, **210**, 35–52.
Meghraoui, M., Maouche, S., Chemaa, B., Cakir, Z., Aoudia, A., Harbi, A., Alasset, P. J., Ayadi, A., Bouhadad, Y., and Benhamouda, F., 2004. Coastal uplift and thrust faulting associated with the Mw = 6.8 Zemmouri (Algeria) earthquake of 21 May, 2003. *Geophysical Research Letters*, **31**, L19605, doi:10.1029/2004GL020466.
Nakata, T., and Shimazaki, K., 1997. Geo-slicer, a newly invented soil sample, for high-resolution active fault studies. *Journal of Geography*, **106**, 59–69 (in Japanese).
Nanayama, F., Satake, K., Furukawa, R., Shimokawa, K., Atwater, B. F., Shigeno, K., and Yamaki, S., 2003. Unusually large earthquakes inferred from tsunami deposits along the Kuril trench. *Nature*, **424**, 660–663.
Obermeier, S. F., 2009. Using liquefaction-induced and other soft-sediment features for paleoseismic analysis. In McCalpin, J. P. (ed.), *Paleoseismology.* Academic Press, pp. 497–564.
Okada, A., 1991. Studies of active faults by trenching method. In Yonekura, N., Okada, A., and Moriyama, A. (eds.), *Tectonic Landforms*. Tokyo: Kokon-Shoin, pp. 18–44 (in Japanese).
Pantosti, D., Schwartz, D. P., and Valensise, G., 1993. Paleoseismology along the 1980 surface rupture of the Irpina fault; implications for earthquake recurrence in the southern Apennines, Italy. *Journal of Geophysical Research*, **98**, 6561–6577.
Payne, S. J., Hackett, W. R., and Smith, R. P., 2009. Paleoseismology of volcanic environments. In McCalpin, J. P. (ed.), *Paleoseismology.* pp. 271–314.
Research Group for Active Faults of Japan, 1992. *Maps of Active Faults in Japan with an Explanatory Text*. Tokyo: University of Tokyo Press. 74 pp.
Sangawa, A., 2010. A study of paleoearthquakes at archeological sites – a new interdisciplinary area between paleoseismology and archeology. *Synthesiology*, **2**, 84–94. National Institute of Advanced Industrial Science and Technology (AIST).
Sawai, Y., Satake, K., Kamataki, T., Nasu, H., Shishikura, M., Atwater, B. F., Horton, B. P., Kelsey, H. M., Nagumo, T., and Yamaguchi, M., 2004. Transient uplift after a 17th-century earthquake along the Kuril subduction zone. *Science*, **306**, 1918–1920.
Scharer, K. M., Weldon, R. J., II, Fumal, T. E., and Biasi, G. P., 2007. Paleoearthquakes on the southern San Andreas Fault, Wrightwood, California, 3000 to 1500 B.C.: a new method for evaluating paleoseismic evidence and earthquake horizons. *Bulletin of the Seismological Society of America*, **97**, 1054–1093.
Scharer, K., Biasi, G. P., Weldon, R. J., II, and Fumal, T. E., 2010. Quasi-periodic recurrence of large earthquakes on the southern San Andreas Fault. *Geology*, **38**, 555–558.
Schwartz, D. P., and Coppersmith, K. J., 1984. Fault behavior and characteristic earthquakes: examples from the Wasatch and San Andreas Fault zones. *Journal of Geophysical Research*, **89**, 5681–5698.
Sherrod, B. L., Brocher, T. M., Weaver, C. S., Bucknam, R. C., Blakely, R. J., Kelsey, H. M., Nelson, A. R., and Haugerud, R. A., 2004. Holocene Fault scarps near Tacoma, Washington, USA. *Geology*, **32**, 9–12.
Shiki, T., Tachibana, T., Fujiwara, O., Goto, K., Nanayama, F., and Yamazaki, T., 2008. Characteristic features of tsunamites. In Shiki, T. (ed.), *Tsunamiites – Features and Implications*. Amsterdam: Elsevier, doi:101016/B978-0-444-51552-0.00018-7.
Shimazaki, K., and Nakata, T., 1980. Time-predictable recurrence model for large earthquakes. *Geophysical Research Letters*, **7**, 279–282.

Sieh, K. E., 1978. Pre-historic large earthquakes produced by slip on the San Andreas Fault at Pallett Creek, California. *Journal of Geophysical Research*, **83**, 3970–3939.
Somerville, P. G., Irikura, K., Graves, R., Sawada, S., Wald, D., Abrahamson, N., Iwasaki, Y., Kagawa, T., Smith, N., and Kowada, A., 1999. Characterizing crustal earthquake slip models for the prediction of strong ground motion. *Seismological Research Letters*, **70**, 59–80.
Stiros, S. C., 2001. The AD 365 Crete earthquake and possible seismic clustering during the fourth to sixth centuries AD in the Eastern Mediterranean: a review of historical and archaeological data. *Journal of Structural Geology*, **23**, 545–562.
Stuiver, M., Reimer, P. J., Bard, E., Beck, J. W., Burr, G. S., Hughen, K. A., Kromer, B., McCormac, F. G., van der Plicht, J., and Spurk, J., 1998. INTCAL98 radiocarbon age calibration, 24000-0 cal BP. *Radiocarbon*, **40**, 1041–1083.
Sugiyama, Y., Mizuno, K., Nanayama, F., Sugai, T., Yokota, H., Hosoya, T., Miura, K., Takemura, K., and Kitada, N., 2003. Study of blind thrust faults underlying Tokyo and Osaka urban areas using a combination of high-resolution seismic reflection profiling and continuous coring. *Annals of Geophysics*, **46**, 1071–1085.
Takada, K., and Atwater, B. F., 2004. Evidence for liquefaction identified in peeled slices of Holocene deposits along the lower Columbia River. Washington. *Bulletin of the Seismological Society of America*, **94**, 550–575.
Takada, K., Satake, K., Sangawa, A., Shimokawa, K. Kumagai, H. Goto, K., and Haraguchi, T., 2002. *Survey of tsunami deposits at an archaeological site along the eastern Nankai trough.* Annual Report on Active Fault and Paleoearthquake Researches, Geological Survey of Japan, AIST, Vol. 2, pp. 235–243 (in Japanese).
Toda, S., and Awata, Y., 2008. Does the 2007 Noto Hanto earthquake reveal a weakness in the Japanese national seismic hazard map that could be remedied with geologic data? *Earth Planets Space*, **60**, 1047–1052.
Tsutsumi, H., and Yeats, R. S., 1999. Tectonic setting of the 1971 Sylmar and 1994 Northridge earthquakes in the San Fernando Valley, California. *Bulletin of the Seismological Society of America*, **89**, 1232–1249.
Weldon, R., Fumal, T., and Biasi, G., 2004. Wrightwood and the earthquake cycle: what a long recurrence record tells us about how faults work. *GSA Today*, **14**(9), 4–10.
Wells, D. L., and Coppersmith, K. J., 1994. New empirical relationships among magnitude, rupture length, rupture area, and surface displacement. *Bulletin of the Seismological Society of America*, **84**, 974–1002.
Wesnousky, S. G., 1988. Seismological and structural evolution of strike-slip faults. *Nature*, **335**, 340–343.
Wesnousky, S. G., 1994. The Gutenberg-Richter or characteristic earthquake distribution, which is it? *Bulletin of the Seismological Society of America*, **84**, 1940–1959.
Wesnousky, S. G., 2006. Predicting the endpoints of earthquake ruptures. *Nature*, **444**, 358–360.
Wesnousky, S. G., and Leffler, L. M., 1992. The repeat time of the 1811 and 12 New Madrid earthquakes; a geological perspective. *Bulletin of the Seismological Society of America*, **82**, 1756–1784.
Working Group on California Earthquake Probabilities (WGCEP), 1995. Seismic hazards in southern California: probable earthquakes, 1994–2024. *Bulletin of the Seismological Society of America*, **85**, 379–439.
Yeats, R. S., and Prentice, C. S., 1996. Introduction to special section: paleoseismology. *Journal of Geophysical Research*, **101**, 5847–5853.
Zachariasen, J., Sieh, K., Taylor, F. W., Edwards, R. L., and Hantoro, W. S., 1999. Submergence and uplift associated with the giant 1833 Sumatran subduction earthquake: evidence from coral microatolls. *Journal of Geophysical Research*, **104**, 895–919.

Cross-references

Archaeoseismology
Characteristic Earthquakes and Seismic Gaps
Earthquake, Focal Mechanism
Earthquakes and Crustal Deformation
Earthquakes, Volcanogenic
Great Earthquakes
Seismic Hazard
Seismic Instrumentation
Seismic Zonation
Seismicity, Subduction Zone
Seismological Networks
Statistical Seismology
Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat
Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity
Tsunami

PLATE DRIVING FORCES

Alessandro M. Forte
GEOTOP Département des Sciences de la Terre et de l'Atmosphére, Université du Québec à Montréal, Montréal, QC, Canada

Definition

Plate driving forces. The forces exerted on Earth's tectonic plates that produce the observable horizontal motions of these plates.

Introduction

Since the advent of the theory of plate tectonics in the 1960s, it has been postulated that the forces that drive the observed horizontal motions of Earth's tectonic plates are produced by the process of thermal convection in the mantle. The proposition that such motions could be produced and sustained over very long geological time scales in a rocky shell, comprising the lithosphere and underlying deep mantle, that also behaves as a solid through which elastic earthquake waves travel, was initially viewed as a serious obstacle to the acceptance of the mantle convection hypothesis of plate motions. This obstacle was definitively eliminated with the recognition, amply confirmed by high-temperature deformation experiments on minerals (Evans and Kohlstedt, 1995), that the rheology of the mantle allows steady-state creeping flow when subjected to long-term shearing stresses, provided the ambient temperature is sufficiently high (Poirier, 1985).

The relationship between the mantle shear stress and deformation (measured in terms of strain rate), is

characterized by an effective viscosity whose value is strongly dependent on temperature and, if the stress is sufficiently high, also on the stress magnitude (Karato and Wu, 1993). The temperature dependence of mantle viscosity gives rise to a lithospheric layer that is stiffer and less deformable than the much hotter and hence less viscous upper mantle. Indeed, the very existence of the tectonic plates as effectively rigid bodies, with little surface deformation except for that occurring along their boundaries, is generally explained in terms of the strong temperature dependence of mantle viscosity. The creation of weak plate boundaries, characterized by very low effective viscosity, requires other mechanisms such as strain (or stress) weakening (e.g., Tackley, 2000). The rheological contrast between the lithosphere and the underlying mantle is often used as the basis for models of plate driving forces that treat the lithospheric plates and their descending limbs (subducted slabs) under the oceanic trenches (Figure 1a) separately from the mantle (Elsasser, 1969; McKenzie, 1969; Solomon and Sleep, 1974; Harper, 1975; Forsyth and Uyeda, 1975; Richardson et al., 1976; Chapple and Tullis, 1977; Richter, 1977; Davies, 1978; Bird et al., 2008).

The vertical viscosity gradient in the lithospheric plates, from an effectively undeformable, high-viscosity surface to a more deformable, low-viscosity region at the base of the plates is a consequence of the thermal structure of the lithosphere. The oceanic lithosphere can be viewed as the upper thermal boundary layer of mantle convection, in which the heat transported across the deep mantle by rising and sinking convective flow is finally transmitted to the Earth's surface by vertical conduction of heat across the lithosphere (Turcotte and Oxburgh, 1967). This boundary-layer interpretation provides an alternative context for understanding the plate driving forces, in which the horizontal motions of Earth's oceanic lithospheric plates (Figure 1a) are regarded as an integral component of the mantle convective circulation (Turcotte and Oxburgh, 1967; Richter, 1973; Jarvis and Peltier, 1982; Davies, 1988). According to this view, the horizontal plate motions are assumed to be coupled to the underlying thermal convective circulation and they are maintained by the dynamic stresses in the mantle generated by descending and ascending thermal plumes (Hager and O'Connell, 1981; Forte and Peltier, 1987, 1994; Ricard and Vigny, 1989; Gable et al., 1991; Vigny et al., 1991; Lithgow-Bertelloni and Richards, 1998; Becker and O'Connell, 2001; Lowman et al., 2008; Forte et al., 2009).

Balance between buoyancy forces and viscous dissipation

There are thus two distinct approaches that have been taken to develop models of plate driving forces where, on the one hand, the plate forces are analyzed separately from the dynamics of the underlying mantle and, on the other, the plate motions are modeled as an integral part of the convecting mantle. Both approaches must satisfy a fundamental constraint that expresses how work done by the body forces throughout the mantle are dissipated by the viscous deformation associated with the tectonic plate motions and flow in the underlying mantle.

A mathematical expression of the balance between buoyancy and dissipation (e.g., Malvern, 1969) over the entire volume, V_m, of the mantle (including the lithosphere) is:

$$-\int_{V_m} u_r\,\delta\rho\,g_0\,dV = \int_{V_m} \tau_{ij}\epsilon_{ij}\,dV - \int_{S_d} t_i\,\Delta u_i\,dS \qquad (1)$$

in which τ_{ij} and ϵ_{ij} are the deviatoric stress and strain-rate tensors, respectively, t_i is a traction vector acting on surfaces S_d across which there is a discontinuity in motion Δu_i, u_r is the vertical (radial) component of mantle flow, $\delta\rho$ are lateral perturbations in density, and g_0 is the radial gravitational field. All indices i and j represent components in each of the three Cartesian coordinate directions, and repeated indices (in the terms on the right-hand side of Equation 1) denote summation over all values (1–3) of the indices. The traction vector t_i acting on the discontinuity surface S_d depends on the deviatoric stress as follows: $t_i = \tau_{ij}n_j$, in which the n_j are the Cartesian components of the local normal vector on S_d. Although the motion along S_d is discontinuous, the traction t_i must be continuous by Newton's third law (equal and opposite actions and reactions).

The left-hand side of Equation 1 is the (positive) rate of work done by the buoyancy forces $\delta\rho g_0$ inside the regions of the mantle where there is vertical flow, as in hot plumes where $\delta\rho < 0$ and $u_r > 0$, and in cold descending slabs where $\delta\rho > 0$ and $u_r < 0$ (Figure 1a). The first term on the right-hand side of Equation 1 is the rate at which the energy released by the buoyancy forces is dissipated by the internal stresses as they continuously deform the mantle. Since the mantle rheology on plate-tectonic timescales is viscous, this energy dissipation is simply the viscous friction that converts mechanical work into heat. The second term on the right-hand side represents the rate of frictional dissipation of energy on internal discontinuity surfaces or faults and hence this term is only relevant in the upper mantle, and especially the crust, where significant earthquake faulting occurs. Since the frictional tractions t_i will always oppose the discontinuous fault motion Δu_i, the integral over the fault surfaces S_d must be negative (e.g., Malvern, 1969).

As noted above, a large number of models have evaluated the plate driving forces by specifically focusing on the lithosphere itself, distinct from the mantle. In this context, a buoyancy-dissipation balance analogous to Equation 1 can be developed:

$$\begin{aligned} -\int_{V_l} u_r\,\delta\rho\,g_0\,dV - \int_{S_l} [\delta P - \rho_0\delta\phi]\,u_i n_i\,dS \\ + \int_{S_l} u_j\tau_{ji}n_i\,dS = \int_{V_l} \tau_{ij}\epsilon_{ij}\,dV - \int_{S_d} t_i\,\Delta u_i\,dS \end{aligned} \qquad (2)$$

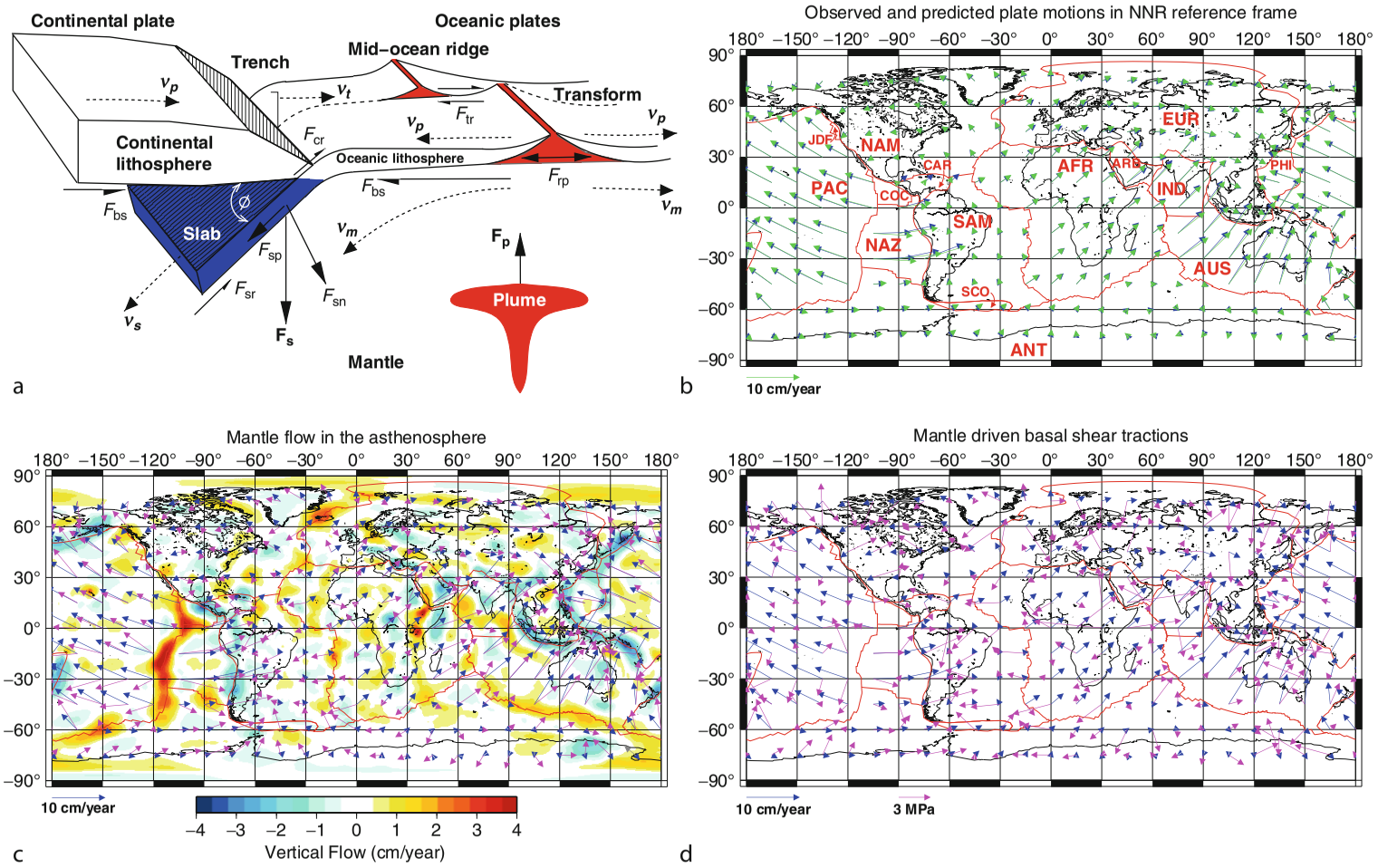


Plate Driving Forces, Figure 1 (**a**) Schematic illustration of forces acting on lithospheric plates. All plate driving forces ultimately arise from the buoyancy forces in the convecting mantle, namely, the force $\mathbf{F_p}$ due to hotter and less dense upwelling plumes and the force $\mathbf{F_s}$ due to the colder and denser sinking slabs. These sublithospheric buoyancy forces generate the plate driving forces $\mathbf{F_{sp}}$ (slab pull) and $\mathbf{F_{rp}}$ (ridge push). Whether the force $\mathbf{F_{bs}}$ (basal shear) is a plate driving or resisting force depends on the relationship of the surface plate velocity $\boldsymbol{v_p}$ to the underlying mantle-flow velocity $\boldsymbol{v_m}$. The purely resistive plate forces are $\mathbf{F_{sr}}$ (slab resistance), $\mathbf{F_{cr}}$ (collision resistance), and $\mathbf{F_{tr}}$ (transform resistance). The balance between the slab-normal force $\mathbf{F_{sn}}$ and the opposing normal force generated by mantle flow will control the trench rollback or retreat velocity $\boldsymbol{v_t}$. (Figure adapted and modified from Forsyth and Uyeda, 1975.) (**b**) The observed (*green arrows*) and predicted (*blue arrows*) present-day plate velocities, where the latter are obtained from a tomography-based model of mantle convection (Forte et al., 2009). The plate motions are with respect to a global No-Net-Rotation (NNR) frame of reference. The plate names are shown by the abbreviations in red font: AFR = Africa, ANT = Antarctica, ARB = Arabia, AUS = Australia, CAR = Carribean, COC = Cocos, EUR = Eurasia, IND = India, JDF = Juan de Fuca, NAM = North America, NAZ = Nazca, PAC = Pacific, PHI = Philippine, SAM = South America, and SCO = Scotia. (**c**) The mantle flow predicted at 200 km depth using a tomography-based convection model (Forte et al., 2009). The magenta arrows show the horizontal component of the predicted flow vector and the color contours show the vertical component (*scale at bottom*). The blue arrows show the predicted surface plate motions from (**b**). (**d**) The horizontal traction vectors (*magenta arrows*) at 30 km depth (just below the mean depth of the crust) exerted by the mantle flow predicted by a tomography-based convection model (Forte et al., 2009). The blue arrows show the predicted surface plate motions from (**b**). Scale bar for vectors is at bottom.

in which V_l represents the volume occupied by the lithospheric layer, S_l defines the lower surface of the lithosphere (where the local normal vector n_i is approximately pointing in the negative radial direction $-\hat{r}$), δP and $\rho_0 \delta\phi$ are, respectively, the dynamic pressure field and self-gravitation force acting on the base of the lithosphere, and $\tau_{ij} n_j = t_j$ are the shear stresses exerted on the base of the lithosphere by the underlying mantle flow.

If the lithosphere has a sufficiently high viscosity, owing to its relatively colder temperatures compared to the mantle, the deformation rate ϵ_{ij} in plate interiors may be assumed to be negligible. If this approximation is accepted, the first term on the right-hand side of (2) may be ignored, and the primary frictional dissipation of energy occurs on the fault surfaces S_d that define the boundaries of the plates. The vertical flow u_r in the lithospheric layer is weak, owing to the proximity of the upper bounding surface, and thus the work done by the tangential shear stresses exerted by mantle flow (third term on the left-hand side of Equation 2) is approximately balanced by the frictional dissipation of energy on plate-boundary faults. Since the dissipation on the right-hand side of Equation 2) is always positive, this implies the tangential stresses on the base of the lithosphere are, on average, acting in the direction of the flow. These approximations illustrate how plate motions may be governed by a balance between basal-shear stresses, acting as plate driving forces, and resistance to motion along plate boundaries (e.g., Hanks, 1977; Davies, 1978).

Expressions (1) and (2) are obtained on the assumption that no external tangential stresses are applied to the upper and lower bounding surfaces of the mantle and that the local vertical flow vanishes on these bounding surfaces. Some models of plate driving forces have employed prescribed plate velocities as a boundary condition, thereby implying corresponding externally applied tangential stresses on the lithosphere. This external forcing is sometimes referred to as the "hand of God" effect (e.g., Karpychev and Fleitout, 1996). To avoid the unphysical possibility that such external stresses do work on the lithosphere and underlying mantle, the models should be formulated so that compensating surface stresses, generated by the flowing mantle, are able to precisely cancel the work of externally applied stresses (e.g., Hager and O'Connell, 1981).

The principal idea that emerges from these general considerations of the plate driving forces is that plate motions are produced by the buoyancy forces $\delta\rho g_0$ in the mantle (left-hand side of Equation 1). Even when the focus is placed on the lithosphere itself, the contribution of the driving forces in the mantle is still present, but "hidden" in the basal stresses that are exerted on the lithospheric plate (left-hand side of Equation 2).

Plate driving forces: lithospheric models

Formulating the detailed physical mechanisms by which the vertically directed buoyancy forces in the mantle ($\mathbf{F}_s$ and $\mathbf{F}_p$ in Figure 1a) are ultimately expressed as horizontal forces acting on and driving the plates is the central challenge in developing models that generate realistic plate motions. The earliest models explored the dynamics of the lithosphere and its descent into the mantle under deep ocean trenches (e.g., Elsasser, 1969, 1971; McKenzie, 1969; Richter, 1973), and they suggested that the negative buoyancy forces due to the cold lithospheric slabs entering the mantle ($\mathbf{F}_s$ in Figure 1a) are the most efficient drivers of the observed tectonic plate motions.

These initial efforts set the stage for the subsequent development of *parametrized plate-force models* in which the forces driving the plate motions are represented by a limited number of discrete force vectors acting on the lithosphere (Figure 1a). One of the most influential studies of the parametrized plate-force models was carried out by Forsyth and Uyeda (1975). The forces they considered are identified in Figure 1a and they include: (1) the **slab-pull force** $\mathbf{F_{sp}} = \mathbf{F_s} \sin\phi$ acting in the down-dip direction where ϕ is the dip angle of the slab; (2) the **slab-resistance force** $\mathbf{F_{sr}}$ representing the opposing viscous shear stress acting on the sides of the descending slab; (3) the collision-resistance force $\mathbf{F_{cr}}$ representing the frictional stress acting on the upper surface of the slab as it slides past the fault surface of the shallow Benioff zone; (4) the **basal-shear force** $\mathbf{F_{bs}}$ representing the tangential shear stress generated by the flowing mantle acting on the base of the lithospheric plates; (5) the **transform-resistance force** $\mathbf{F_{tr}}$ representing the frictional stresses acting on segments of the plate boundaries where there is a component of relative motion that is tangential to the boundary, such as transform faults; and (6) the ridge-push force $\mathbf{F_{rp}}$ corresponding to the lateral force exerted on the divergent plate boundaries due to the dynamic pressure field that produces the topographic elevation of the mid-ocean ridges.

The orthogonal complement to the slab-pull force is the slab-normal force $\mathbf{F}_{sn} = \mathbf{F}_s \cos\phi$ (Figure 1a). This force component was not considered in the evaluation of the plate driving forces by Forsyth and Uyeda (1975), but the balance between this force and the opposing normal stress in the mantle will contribute to the retrograde motion (or "roll back") of the trench (e.g., Elsasser, 1971; Schellart, 2008; Funiciello et al., 2008) with the relative velocity $\boldsymbol{v_t}$ illustrated in Figure 1a.

The results obtained by Forsyth and Uyeda (1975), and earlier by Richter (1973), suggested that the balance of forces on subducting slabs are the primary control on plate motions and this view has been reiterated in a whole generation of subsequently published models (e.g., Lithgow-Bertelloni and Richards, 1998), up to the present day (e.g., Conrad and Lithgow-Bertelloni, 2004; Billen, 2008; Funiciello et al., 2008; Schellart, 2008).

A major source of uncertainty in the plate-force models is the unknown relationship between the motion of the lithospheric plates $\boldsymbol{v_p}$ and the underlying mantle flow $\boldsymbol{v_m}$ (Figure 1a), and hence the unknown local magnitude and

direction of the basal shear force $\mathbf{F_{bs}}$. The majority of models assumed the mantle is passive, and that the flow below the lithosphere is entirely generated by the subducting slabs. This is equivalent to assuming the ambient mantle temperature outside the sinking lithosphere is isothermal or adiabatic (e.g., McKenzie, 1969; Hager and O'Connell, 1981). For these passive-mantle models, the basal shear force acts in the opposite direction to the plate velocity and therefore opposes the driving force provided by slab pull.

The conclusions obtained from the parametrized plate-force models are nonunique and depend on which forces are assumed to be known. For example, whether basal shear forces contribute to driving or resisting the plate motions will depend on what forces (in Figure 1a) are prescribed as known at the outset. In the study by Forsyth and Uyeda (1975), it was assumed that basal shear forces oppose the plate motions and have a direction that is antiparallel to the plate velocity. In contrast, Davies (1978) showed that if the frictional force $\mathbf{F_{tr}}$ acting on transform boundaries is assumed to be known in advance, on the basis of earthquake stress-drop data (Hanks, 1977), then the basal shear force may act as a plate driving force. The force balance considered by Davies also showed that the collision-resistance force $\mathbf{F_{cr}}$ strongly opposes the slab-pull force $\mathbf{F_{sp}}$, greatly reducing its contribution as a plate driving force. The significance of basal shear stresses as a plate driving force has again been highlighted by Bird et al. (2008), who employed a finite-element model of the lithosphere that incorporates stress- and temperature-dependent viscosities as well as an explicit treatment of plate-boundary faults.

Plate driving forces: whole-mantle models

The popularity of parametrized plate-force models that focused on the lithosphere was based on the use of readily observed input variables derived from surface tectonics and observed plate velocities. Since the observed plate velocities were employed as an input, these models could not explain how the plate motions were generated by the internal dynamics of the mantle, nor how they would evolve over time.

Hager and O'Connell (1981) introduced 3-D spherical, whole-mantle flow models that explicitly included the interaction of the moving lithospheric plates with the underlying mantle. Although these models employed the present-day plate velocities as input, they predicted the 3-D flow throughout the mantle driven by a global distribution of subducted slabs derived from seismic catalogs of earthquake hypocenters. With the advent, in the 1980s, of seismic tomographic imaging of lateral heterogeneity in the mantle, it became feasible to apply these whole-mantle flow calculations to explore the detailed connection between surface plate motions and the 3-D flow in the mantle generated by thermal convection (e.g., Forte and Peltier, 1987; Ricard and Vigny, 1989).

Seismic tomography provides important information regarding the 3-D temperature structure of the convecting mantle and it thus constitutes a basis for developing realistic models of mantle flow dynamics that are no longer limited by the 2-D view of Earth dynamics based on surface observations of plate tectonics. The mantle flow modeling by Hager and O'Connell (1981) was therefore further extended in subsequent studies demonstrating that present-day plate motions can be successfully predicted using the 3-D distribution of buoyancy forces derived from seismic tomography (e.g., Vigny et al., 1991; Forte and Peltier, 1994; Becker and O'Connell, 2001). This confirmation that plate motions can be predicted on the basis of stresses generated in the underlying mantle by a realistic 3-D pattern of convection is a major step toward developing a complete understanding of mantle and lithosphere dynamics and how they may change over geologic time (e.g., Lithgow-Bertelloni and Richards, 1998; Conrad and Lithgow-Bertelloni, 2004; Lowman et al., 2008; Forte et al., 2009).

A recent tomography-based model of mantle convection (Forte et al., 2009) illustrates the detailed relationship between the convective circulation in the mantle and surface plate velocities. The plate velocities predicted by this model match the observed motions very well (Figure 1b). Below the lithosphere, in the underlying low-viscosity portion of the upper mantle referred to as the "asthenosphere," the predicted mantle flow velocities show significant deviations from the overlying plate motions, for example, under the western half of the North American plate and under the African plate (Figure 1c). The lack of parallelism between the surface plate velocities and the upper-mantle flow circulation in the asthenosphere is a consequence of the complex 3-D distribution of buoyancy forces in the deep mantle that are not correlated with the surface plate geometry (e.g., subduction zones and mid-ocean ridges) in any straightforward manner. It is also a consequence of the strong reduction in viscosity in the asthenosphere relative to the overlying lithosphere, thereby allowing a partial decoupling of the horizontal flow vectors in these two layers. This decoupling of the pattern of horizontal motions in the lithosphere and asthenosphere, represented by the vectors $\boldsymbol{v}_{\mathbf{p}}$ and $\boldsymbol{v}_{\mathbf{m}}$ in Figure 1a, will directly influence the local basal shear force $\mathbf{F_{bs}}$ (Figure 1a).

The basal shear force F_{bs} mapped in Figure 1d is the horizontal component of the mantle traction vector $\boldsymbol{t} = -\boldsymbol{\tau} \cdot \hat{\boldsymbol{r}}$, where τ is deviatoric stress tensor. The complex local variations in basal shear force relative to the surface plate velocities show that a simple correlation between these two, as assumed in the parametrized plate-force models (e.g., Forsyth and Uyeda, 1975), is not appropriate. It is only by considering the integrated effect of the basal shear force, measured in terms of the total torque acting on each plate, that the relationship to the surface motions becomes clearer. For example, the integrated basal torque acting on the North American plate, calculated on the basis of the tractions in Figure 1d, has

a 95% correlation to the corresponding North American plate-rotation vector.

The whole-mantle flow models thus suggest that it is the basal shear forces generated by a fully 3-D pattern of mantle convection that drive the motions of the overlying plates (e.g., Vigny et al., 1991; Forte and Peltier, 1994; Becker and O'Connell, 2001; Forte et al., 2009). The role of the plates is not entirely passive, however, because their intrinsic rigidity allows them to partially resist (and hence mask) a significant fraction of the deeper mantle flow patterns (Ricard and Vigny, 1989; Forte and Peltier, 1994). Furthermore, the buoyancy forces associated with subducted lithosphere do contribute to the basal shear forces that drive the plates. In this regard, the plates are not entirely passive because their continuous descent into the mantle, over extended periods of geologic time, ultimately generates the global distribution of subducted slabs that provide the slab buoyancy force $\mathbf{F_s}$ (Figure 1a).

Summary

Much progress has been achieved in our understanding of how mantle convection drives the surface plate motions since the earliest parametric force models were published over 3 decades ago. This progress would not have been possible without the information on 3-D mantle structure that has been obtained through seismic tomography. The interpretation of the seismic tomography models is however nonunique, owing to the imperfect sampling of the mantle by seismic waves, and also because there are other effects, in addition to lateral temperature variations, that may give rise to mantle heterogeneity.

This complexity has led to the emergence of two different views on how thermal convection in the mantle drives the tectonic plate motions. The most long held and commonly accepted view is that the buoyancy forces due to subducted slabs are the primary drivers of surface plate motions (e.g., Richter, 1973; Conrad and Lithgow-Bertelloni, 2004). According to this view the thermal evolution of the mantle is controlled almost entirely by cooling from above (i.e., subducting lithosphere), and this would arise if the mantle is almost entirely heated internally, with little or no heat entering the mantle across the core–mantle boundary.

The second view, supported by the earliest mantle convection models (e.g., Turcotte and Oxburgh, 1967) and some subsequent models (e.g., Jarvis and Peltier, 1982; Forte et al., 2009), is that a combination of active hot upwellings ("plumes") and subducted slabs is required to explain the surface plate motions. The existence of such active hot plumes requires a significant heat flux across the core-mantle boundary and it would be compatible with a mainly thermal origin for the seismic anomalies evident in the tomography models (e.g., Forte et al., 2009). The current debate on which of these two contrasting views of mantle dynamics is most compatible with the available data is not yet settled, but a resolution will be possible once the magnitude of the heat flux across the core-mantle boundary becomes more clear.

Bibliography

Becker, T. W., and O'Connell, R. J., 2001. Predicting plate velocities with mantle circulation models. *Geochemistry Geophysics Geosystems*, **2**, doi:10.1029/2001GC000171.

Billen, M. I., 2008. Modeling the dynamics of subducting slabs. *Annual Review of Earth and Planetary Sciences*, **36**, 325–356.

Bird, P., Liu, Z., and Rucker, W. K., 2008. Stresses that drive the plates from below: Definitions, computational path, model optimization, and error analysis. *Journal of Geophysical Research*, **113**, B11406, doi:10.1029/2007JB005460.

Chapple, W. M., and Tullis, T. E., 1977. Evaluation of the forces that drive the plates. *Journal of Geophysical Research*, **82**, 1967–1984.

Conrad, C. P., and Lithgow-Bertelloni, C., 2004. The temporal evolution of plate driving forces: Importance of "slab suction" versus "slab pull" during the Cenozoic. *Journal of Geophysical Research*, **109**, B10407, doi:10.1029/2004JB002991.

Davies, G. F., 1978. The roles of boundary friction, basal shear stress and deep mantle convection in plate tectonics. *Geophysical Research Letters*, **5**, 161–164.

Davies, G. F., 1988. Role of the lithosphere in mantle convection. *Journal of Geophysical Research*, **93**, 10,451–10,466.

Elsasser, W. M., 1969. Convection and stress propagation in the upper mantle. In Runcorn, S. K. (ed.), *The Applications of Modern Physics to the Earth and Planetary Interiors*. New York: Interscience, pp. 223–246.

Elsasser, W. M., 1971. Sea-floor spreading as thermal convection. *Journal of Geophysical Research*, **76**, 1101–1112.

Evans, B., and Kohlstedt, D. L., 1995. Rheology of rocks, in rock physics and phase relations: A handbook of physical constants, American geophysical union (Washington, DC). *AGU Reference Shelf*, **3**, 148–165.

Forsyth, D., and Uyeda, S., 1975. On the relative importance of the driving forces of plate motion. *Geophysical Journal of the Royal Astronomical Society*, **43**, 163–200.

Forte, A. M., and Peltier, W. R., 1987. Plate tectonics and aspherical Earth structure: The importance of poloidal–toroidal coupling. *Journal of Geophysical Research*, **92**, 3645–3679.

Forte, A. M., and Peltier, W. R., 1994. The kinematics and dynamics of poloidal–toroidal coupling in mantle flow: The importance of surface plates and lateral viscosity variations. *Advances in Geophysics*, **36**, 1–119.

Forte, A. M., Moucha, R., Rowley, D. B., Quéré, S., Mitrovica, J. X., Simmons, N. A., and Grand, S. P., 2009. Recent tectonic plate decelerations driven by mantle convection. *Geophysical Research Letters*, **36**, L23301, doi:10.1029/2009GL040224.

Funiciello, F., Faccenna, C., Heuret, A., Lallemand, S., Di Giuseppe, E., and Becker, T. W., 2008. Trench migration, net rotation and slab-mantle coupling. *Earth and Planetary Science Letters*, **271**, 233–240.

Gable, C. W., O'Connell, R. J., and Travis, B. J., 1991. Convection in three dimensions with surface plates. *Journal of Geophysical Research*, **96**, 8391–8405.

Hager, B. H., and O'Connell, R. J., 1981. A simple global model of plate dynamics and mantle convection. *Journal of Geophysical Research*, **86**, 4843–4867.

Hanks, T. C., 1977. Earthquake stress drops, ambient tectonic stresses and stresses that drive plate motions. *Pure and Applied Geophysics*, **115**, 441–458.

Harper, J. R., 1975. On the driving forces of plate tectonics. *Geophysical Journal of the Royal Astronomical Society*, **40**, 465–474.
Jarvis, G. T., and Peltier, W. R., 1982. Mantle convection as a boundary layer phenomenon. *Geophysical Journal of the Royal Astronomical Society*, **68**, 385–424.
Karato, S-i., and Wu, P., 1993. Rheology of the upper mantle: A synthesis. *Science*, **260**, 771–778.
Karpychev, M., and Fleitout, L., 1996. Simple considerations on forces driving plate motion and on the plate-tectonic contribution to the long-wavelength geoid. *Journal of Geophysical Research*, **127**, 268–282.
Lithgow-Bertelloni, C., and Richards, M. A., 1998. The dynamics of Cenozoic and Mesozoic plate motions. *Reviews of Geophysics*, **36**, 27–78.
Lowman, J. P., Gait, A. D., Gable, C. W., and Kukreja, H., 2008. Plumes anchored by a high viscosity lowermantle in a 3D mantle convection model featuring dynamically evolving plates. *Geophysical Research Letters*, **35**, L19309, doi:10.1029/2008GL035342.
Malvern, L. E., 1969. *Introduction to the Mechanics of a Continuous Medium*. Englewood Cliffs: Prentice-Hall.
McKenzie, D. P., 1969. Speculations on the consequences and causes of plate motions. *Geophysical Journal of the Royal Astronomical Society*, **18**, 1–32.
Poirier, J. P., 1985. *Creep of Crystals*. Cambridge: Cambridge University Press.
Ricard, Y., and Vigny, C., 1989. Mantle dynamics with induced plate tectonics. *Journal of Geophysical Research*, **94**, 17543–17559.
Richardson, R. M., Solomon, S. C., and Sleep, N. H., 1976. Intraplate stress as an indicator of plate tectonic driving forces. *Journal of Geophysical Research*, **81**, 1847–1856.
Richter, F., 1973. Dynamical models for sea floor spreading. *Reviews of Geophysics and Space Physics*, **11**, 223–287.
Richter, F., 1977. On the driving mechanism of plate tectonics. *Tectonophysics*, **38**, 61–88.
Schellart, W. P., 2008. Kinematics and flow patterns in deep mantle and upper mantle subduction models: Influence of the mantle depth and slab to mantle viscosity ratio. Geochemistry. Geophysics Geosystems **9**, Q03014, doi:10.1029/2007GC001656.
Solomon, S. C., and Sleep, N. H., 1974. Some simple physical models for absolute plate motions. *Journal of Geophysical Research*, **79**, 2557–2567.
Tackley, P. J., 2000. Mantle convection and plate tectonics: Toward an integrated physical and chemical theory. *Science*, **288**, 2002–2007.
Turcotte, D. L., and Oxburgh, E. R., 1967. Finite amplitude convective cells and continental drift. *Journal of Fluid Mechanics*, **28**, 29–42.
Vigny, C., Ricard, Y., and Froidevaux, C., 1991. The driving mechanism of plate tectonics. *Tectonophysics*, **187**, 345–360.

Cross-references

Continental Drift
Earth's Structure, Global
Earth's Structure, Upper Mantle
Energy Budget of the Earth
Lithosphere, Oceanic
Lithosphere, Oceanic: Thermal Structure
Mantle Convection
Mantle Plumes
Seismic Structure at Mid-Ocean Ridges
Subduction Zones

PLATE MOTIONS IN TIME: INFERENCES ON DRIVING AND RESISTING FORCES

Giampiero Iaffaldano[1], Hans-Peter Bunge[2]
[1]Research School of Earth Sciences, The Australian National University, Acton, ACT, Australia
[2]Geophysics Section, Department of Earth and Environmental Sciences, Ludwig Maximilians University of Munich, Munich, Germany

Introduction

Plate tectonics is a remarkable theory. Its central tenet, which is widely known, states that the surface of the Earth moves in a piecewise coherent fashion via a number of rigid plates (Wilson, 1965; Morgan, 1968), with most of the deformation focused along plate boundaries (Stein, 1987). The underlying dynamic mechanism responsible for plate motions has been long identified in lateral buoyancy variations within the Earth mantle (Forte, 2010). However, 40 years after the acceptance of plate tectonics, the details of how these forces may vary in time to modify plate movements remain limited, and we still fail to answer basic questions of fundamental importance: *Why do plates change their motions as shown by the geologic and geodetic records? – What are the spatial and temporal patterns of forces responsible for such changes?* The main difficulty stems from our poor knowledge of the *force balance* in plate tectonics. It is widely agreed that plate motions are driven by convection in the Earth's mantle (Hager and O'Connel, 1981; Davies and Richards, 1992). But the influence of other driving and resisting mechanisms, especially along plate margins, remains unclear. Forces concentrated along plate margins are known as plate boundary forces (Forsyth and Uyeda, 1975) and it is likely that they play an important role in modulating plate motions (Humphreys and Coblentz, 2007).

A key component of plate boundary forces is the gravitational collapse associated with large mountain belts extending along plate margins. Climate-induced variations in topography, related to uplift and erosion in orogens, have been long recognized as possible controls in tectonics (Koeppen and Wegener, 1924), and may be therefore capable to initiate variations in plate motion on regional and global scales. Conversely, geodynamicists can use inferences on topography and plate motion variations as powerful probes into the underlying force balance of plate tectonics.

Temporal variations in plate motions provide significant constraints on the budget of forces acting upon plates. This is because by virtue of Newton's first law of motion any change in plate motion is necessarily driven by variations in one or more driving or resisting forces. With the advent of space geodetic techniques (Dixon, 1991), increasingly accurate estimates of global plate motions and their temporal variations are now available (Sella et al., 2002). While geodesy measures the current rates of plate motion, one can also glean long-term constraints

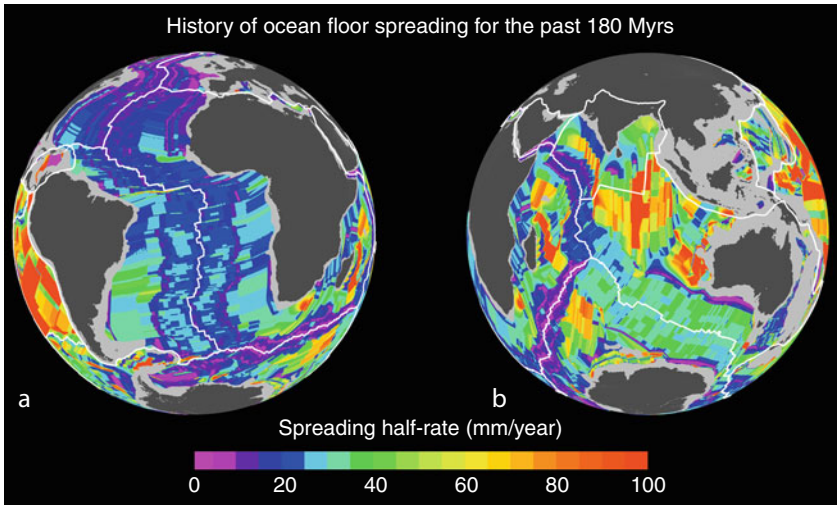


Plate Motions in Time: Inferences on Driving and Resisting Forces, Figure 1 Observed oceanic spreading half-rates for the past 180 Myrs after a recent global compilation by Müller et al. (2008). Plate boundaries are in white, continents in dark gray. Abrupt changes in spreading rates reveal short-term variations in global plate motions, particularly visible in the South Atlantic (**a**) as well as in the Indian Ocean (**b**). Such rapid variations are unlikely to originate from global changes in mantle driving forces, which occur on a longer time scale on the order of 50–100 Myrs as indicated by mantle circulation models. Instead, they are related to short-term variations in plate boundary forces caused, for example, by rapid growth of surface topography at convergent margins (see text). These observations point to the first-order importance of plate boundary forces in controlling global plate motions.

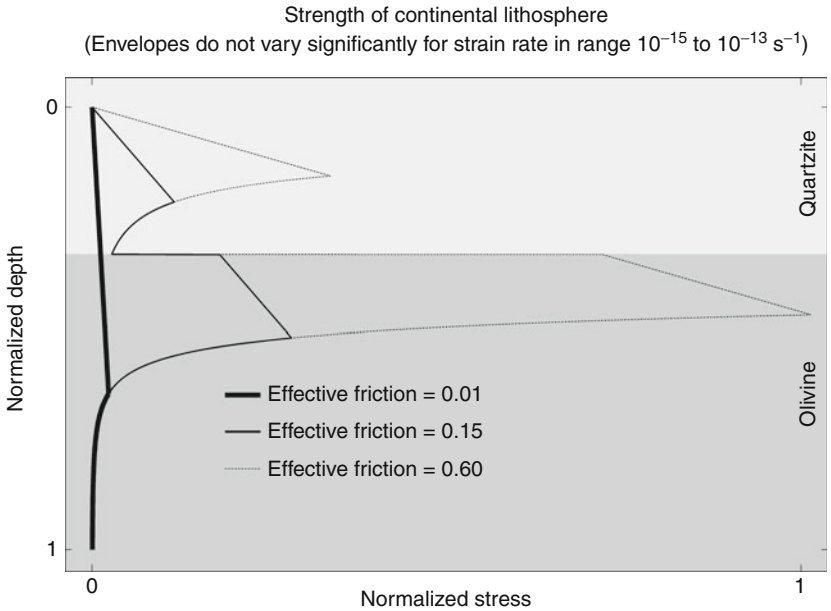


Plate Motions in Time: Inferences on Driving and Resisting Forces, Figure 2 Normalized strength/depth profiles for a simplified two-layer continental lithosphere, plotted for three different fault friction coefficients and a strain rate in the range of 10**–15 to 10**–13 1/s, typical of plate tectonics. Laboratory experiments performed mainly on quartzite and olivine (see text) indicate that lithosphere strength increases linearly with overburden pressure in the upper brittle part, and decreases exponentially with increasing temperature in the lower ductile part. Note that the brittle part contributes a significant portion (70–80%) of the total lithospheric strength, independently of the particular choice of friction coefficient and strain rate values. Experimental results suggest a friction coefficient around 0.6 (*dashed envelope*) for rocks under lithostatic pressure conditions, known as Byerlees law. However, various independent evidences for weak faults suggest much lower values between 0.01 and 0.15 (*solid envelopes*), so that plate boundaries experience considerably lower stresses even for high strain rates.

on plate motion from paleomagnetic observations derived from the magnetic isochron record of the ocean floor. In fact, recent paleomagnetic plate motion models approach temporal resolutions on the order of 1–2 Myrs and reveal a number of rapid *velocity-changes* (few cm/year) occurring over time-periods of a few Myrs or less (Müller et al., 2008). Some of these are particularly evident in the Southern Atlantic (Figure 1a) and in the Indian Ocean (Figure 1b). Their short duration makes it unlikely to attribute the changes to variations in the internal distribution of mantle buoyancies. In fact, these evolve globally on longer time scales on the order of 50–100 Myrs (Bunge et al., 1998), although locally changes may occur over some 10 Myrs (Forte et al., 2009). It is reasonable to link these rapid changes to variations in plate boundary forces.

Plate tectonics and the rheology of the lithosphere

The regime of plate tectonics requires high strain rates and low resistive stresses to coexist simultaneously along plate margins. But it is challenging to model this regime on a computer, because it is difficult to simulate shear failure along plate boundaries. To overcome this challenge some models of the lithosphere incorporate highly non-Newtonian, viscous creep, strain-rate weakening rheologies together with viscoplastic yielding. Moresi and Solomatov (1998) explored the effects of temperature-dependent viscosity in combination with a plastic yield stress: The former causes the cold upper boundary layer (lithosphere) to be strong, while the latter allows the boundary layer to fail locally in regions of high stress. The success, measured through a so-called plateness, is evident when extreme strain-softening rheologies, known as pseudo-stick-slip (Bercovici, 1995), are employed (Bercovici, 2003). Unfortunately, the rheological parameters required for pseudo-stick-slip agree poorly with laboratory experiments of ductile deformation performed on olivine (Kirby, 1983), particularly at the pressure and temperature conditions typical of the upper mantle (Karato and Jung, 2003). Figure 2 shows normalized strength envelopes of a simplified two-phase lithosphere. Depth-dependent strength is parameterized via empirical laws established through laboratory experiments performed on quartzite, abundant in the upper 20 km of continental lithosphere, and olivine, which dominates at greater depths. The laboratory results indicate that strength increases linearly with overburden pressure in the upper, brittle part of the lithosphere; it then decreases exponentially with increasing temperature in the lower, ductile part (Kohlstedt et al., 1995). High strength in the upper lithosphere reflects the resistance of rocks to failure at low temperature, or to sliding past each other when already faulted. Experimental results indicate a simple linear relationship to parameterize this behavior (Byerlee, 1978), with shear stress proportional to the normal pressure through a friction coefficient typically on the order of 0.6 (dashed envelope in Figure 2). There is, however, mounting evidence for significantly lower values (in the range 0.01–0.15, see solid envelopes in Figure 2) along faults and plate boundaries (Hickman, 1991; Bird, 1998; Suppe, 2007).

Computer models of the faulted lithosphere coupled with global mantle circulation models

Lithospheric faults exhibit low friction, as we have noted. In other words, they are mechanically weak and experience low stresses independently of the strain rate. It is therefore attractive to represent faults directly within the computational grid of numerical models of the lithosphere. This can be done, for instance, through the use of so-called contact-elements, in an approach known as neo-tectonic modeling. Neo-tectonic models have reached considerable sophistication. They solve the equations of mass and momentum conservation, and compute the instantaneous force balance and associated plate velocities. They implement empirical, depth-dependent rheologies of the

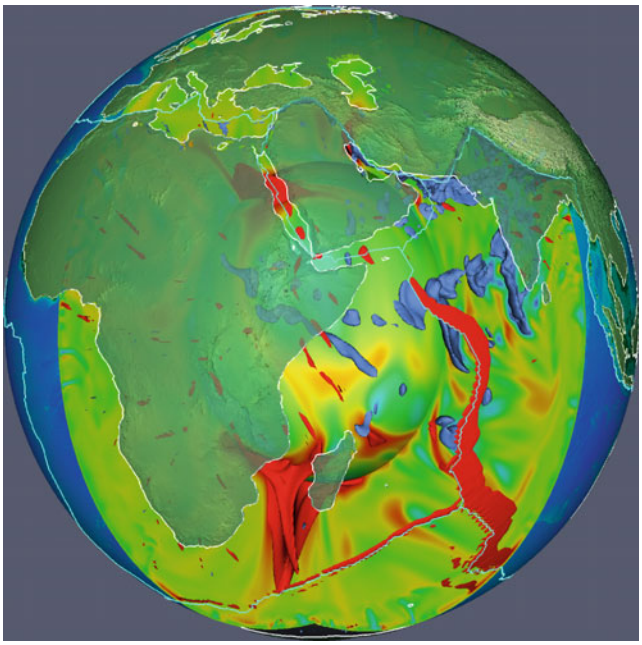


Plate Motions in Time: Inferences on Driving and Resisting Forces, Figure 3 Temperature distribution in the Earth's mantle from a recent, high-resolution 3-D global circulation model. *Blue* represents cold, denser material whereas *red* is hot and buoyant mantle. View is on Africa, coastline is in *white*, with continental topography in transparent *green color* scale. Present-day plate boundaries are outlined in *blue*. More than 100 million grid points discretize the Earth's mantle, equivalent to an average grid spacing of 20 km or less. Circulation models include radial variations in mantle viscosity (factor 40 increase from the upper to the lower mantle), internal heat generation from radioactivity, bottom heating from the core, and a history of subduction spanning the past 120 Myrs. A cold downwelling is visible beneath Tibet where the ancient Tethys Ocean subducted under Eurasia; a hot and buoyant upwelling is visible as well beneath the spreading triple-junction of the Antarctica, Africa, and Australia plates. The circulation model provides a realistic, first-order estimate of internal buoyancy forces driving global plate motions.

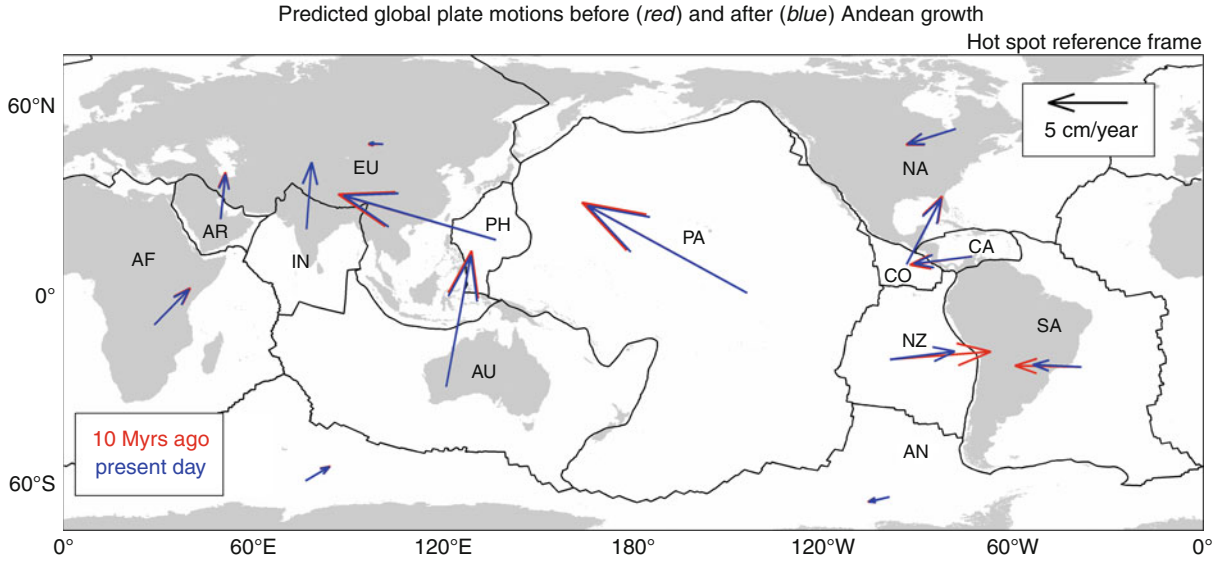


Plate Motions in Time: Inferences on Driving and Resisting Forces, Figure 4 Predicted plate motions in the Hot spot reference frame from coupled global mantle convection/lithosphere dynamics simulations, corresponding to assumed Andean paleotopography 10 Myrs ago (*red*) and present-day topography (*blue*). Plate boundaries are in *black*, continents in *gray*. AF Africa, AN Antarctica, AR Arabia, AU Australia, CA Caribbean, CO Cocos, EU Eurasia, IN India, NA North America, NZ Nazca, PA Pacific, PH Philippine, SA South America. IN and AU are treated as separate in the computational grid, based on recent evidences of plate separation (see text and Figures 6, 7). Note that a lower paleotopography assumed for the Andes 10 Myrs ago results in a predicted NZ/SA convergence of 10.1 cm/year at long 71.5°W, lat 25°S, whereas present-day topography results in a predicted convergence of 6.9 cm/year at the same position (see text). The rates compare remarkably well with observations inferred from a variety of data, which indicate a 30% reduction of NZ/SA plate convergence from 10.3 cm/year to 6.7 cm/year over the past 10 Myrs. The modeling results suggest that the reduction of NZ/SA convergence is caused by resisting plate-margin forces associated with the topographic load of Andes (see text). Similar numerical models moreover confirm that frictional variations along the boundary, arising from variations in trench sediment infill, are insufficient to explain the record of plate motion.

lithosphere and account for ductile deformation in the lower crust as well as brittle deformation along faults. In some cases, they take advantage of the so-called thin-sheet approximation to reduce the computational complexity from 3-D to 2-D (Kong and Bird, 1995) and to achieve greater computational efficiency.

Independent of advances in neo-tectonic models, there has been much progress in modeling the global circulation of the Earth's mantle (Tackley et al., 1994; Bunge et al., 1997; Zhong et al., 2000). Mantle circulation models (MCMs) account for the dynamic effects from a weak asthenosphere on the horizontal length-scales of the flow (Bunge et al., 1996), include internal heat generation from radioactivity, as well as a significant amount of heat from the core (Bunge, 2005; Quere and Forte, 2006; van der Hilst et al., 2007). Combined with constraints on the history of subduction, these models place first-order estimates on the internal mantle buoyancies driving plate motions. Figure 3 shows the temperature distribution in the mantle from one recent model with more than 100 million grid points (Oeser et al., 2006), equivalent to a grid point spacing of 20 km and less throughout the model mantle. MCMs provide first-order estimates of mantle buoyancy forces, but do not account for the complex processes in the lithosphere such the brittle failure. Similarly, neo-tectonic models include stresses originating within the lithosphere, and realistic plate boundary forces. But they rely on assumptions of the mantle buoyancy field to complete the force balance. The logical step is merging the two model classes to simulate the coupled global mantle convection/plate tectonics system. This makes it possible to account simultaneously for plate boundary forces and mantle-related components of lithospheric force balance, and to predict variations in global plate velocities that one may test explicitly against the geologic record. Several authors have recently undertaken such approach (Iaffaldano et al., 2006; Ghosh et al., 2008). Results are encouraging, and in the following, we review recent predictions of climate-related topography variations and their influence on plate velocities in the Southern Pacific, Southern Atlantic, as well as in the Indian Ocean that compare well with observations.

Recent plate motion changes

Kinematics of the Southern Pacific and Atlantic, and its relations to climate variations and topographic growth in the Andes

Paleomagnetic (Gordon and Jurdy, 1986; DeMets et al., 1994) and geodetic (Norabuena et al., 1999) data indicate a significant reduction (as high as 30%) of the Nazca/South America plate convergence over the

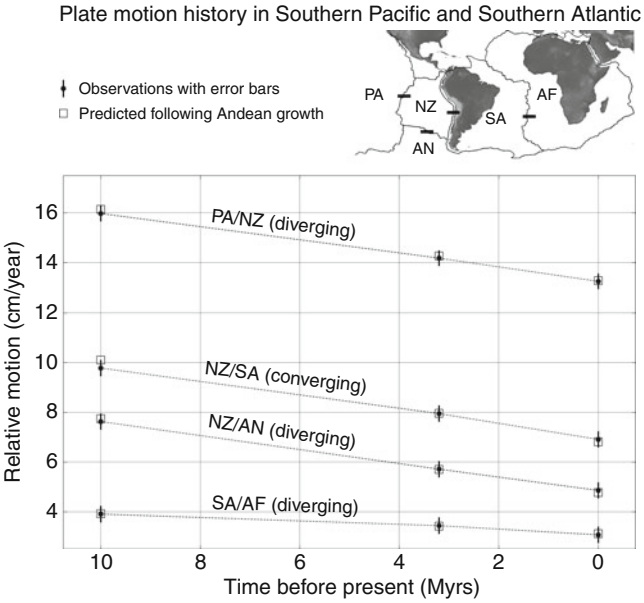


Plate Motions in Time: Inferences on Driving and Resisting Forces, Figure 5 Predicted and observed relative plate motions in the South Atlantic and South Pacific over the past 10 Myrs for a set of adjacent plate pairs: PA/NZ, NZ/SA, NZ/AN, and SA/AF (abbreviations as in Figure 6). *Black bold segments* in the small inset indicate positions along plate boundaries (*thin black*) at which relative motions have been computed. Observed plate motions (with *error bars*) inferred from paleomagnetic and geodetic data are represented by *black dots*, while *empty squares* indicate relative motions predicted from our simulations of the global coupled mantle/lithosphere system. The models explicitly account for the growth of the Andes over the past 10 Myrs, and demonstrate that the relative plate motion record can be entirely explained with the history of Andean orogeny. Our simulations thus point to the importance of far-field effects in plate tectonics, and imply that resisting plate margin forces due to Andean growth account for about 18% of global plate motion changes over the past 10 Myrs (see Figure 8).

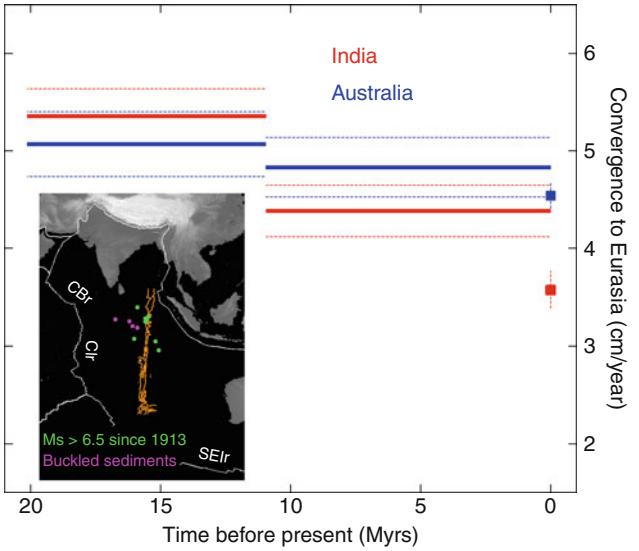


Plate Motions in Time: Inferences on Driving and Resisting Forces, Figure 6 Observed convergence of India (*red*) and Australia (*blue*) relative to fixed Eurasia over the past 20 Myrs. Convergence rates are computed through rigid-rotation Euler poles at long 86°E, lat 27°N of the India/Eurasia margin. Present-day values (*squares*) are derived from geodetic techniques while the paleomagnetic record (*solid lines*) is computed by averaging finite rotations of magnetic anomalies identified along the Carlsberg and South-East Indian ridges (labeled respectively, as CBr and SEIr in inset. Clr is Central Indian ridge). Note that convergence rates are very similar between 20 and 11 Myrs ago, when India and Australia appear to behave as one single plate with presumably little internal deformation. The convergence relative to Eurasia differs more distinctly over the last 11 Myrs, when India and Australia slowed down by almost 2 cm/year and 0.5 cm/year, respectively. Inset shows locations of identified unconformities of sediments (*magenta dots*) as well as great (Ms $>$ 6.5) earthquakes (*green dots*) indicating left-lateral strike-slip motion in the northern portion of the Ninety East Ridge (*orange contours*). Those evidences suggest diffuse deformation in the Indian Ocean particularly pronounced during late Miocene, and have been interpreted as separation between the India and Australia. Plate boundaries are in *white*, continental topography in *gray color* scale.

past 10 Myrs. The timing of the slowdown is significant in that it is coeval with major growth of the Andes (Allmendinger et al., 1997) inferred from a variety of independent data (Gregory-Wodzicki, 2000). Iaffaldano et al. (2006) tested the effect of topography on plate convergence by computing plate velocities before and after Andean topography growth, assuming mantle shear tractions from a MCM, as well as a low fault friction coefficient of 0.03 for the tectonic model. They performed two separate simulations of global plate motions, in order to estimate the increase of horizontal deviatoric force in the frictional region of the Nazca/South America interface arising from the gravitational collapse of the Andean belt, and its effect on the convergence rate. One simulation was based on present-day topography from the ETOPO 5 data set, as high as 5 km in the central Andes, to compute global plate velocities. A second simulation accounted for a lower topographic relief of continental South America, based on a reconstruction of Andean paleo-elevation 10 Myrs ago (Gregory-Wodzicki, 2000). Comparison of the two computed velocity fields predicts a 30% convergence reduction between the Nazca and South America plates (Figure 4), and agrees well with the record of present and past plate motions. More importantly, the contribution of the momentum from the gravitational collapse of the large plateau has been estimated to be of some 10**13 N/m along the central Andes. It should be pointed out that other mechanisms, such as temporal variations of friction along plate boundaries (Iaffaldano et al., 2008) and lateral buoyancy variations (Forte et al., 2009), may contribute to modify the convergence regime along the Nazca/South America margin to a minor extent. From Figure 4, it is evident that the total convergence reduction is unevenly partitioned between the Nazca and South America plates. This can be understood by recalling that mantle shear tractions exerted on the lithosphere-base scale to first order with

the basal surface area of the plate times its velocity. Because the Nazca plate is smaller than South America, a higher velocity reduction is required in the momentum balance for mantle shear tractions to equilibrate topography-generated plate-boundary forces along the margin. Iaffaldano and Bunge (2009) took a step forward and used global models to predict the history of relative motion for plates adjacent to Nazca and South America (Pacific, Africa, Antarctica), which are shown in Figure 5. Agreement between models and observations is remarkable, implying that the resisting forces along the Nazca/South America plate boundary are responsible for driving plate motion changes also in the Southern Atlantic and Southern Pacific regions.

Plate motion changes in the Indian Ocean

While instantaneous calculations of the plate tectonic momentum balance cannot be taken to model the temporal evolution of plate boundaries, they do allow us to test the effects of variations in plate geometry on global plate motions, and in particular the creation of new plate boundaries. From the principle of inertia, it follows that any such event would invariably trigger plate motion changes due to repartitions in the budget of basal drag and plate boundary forces. A recent such episode is thought to have occurred in the Indian Ocean, where a variety of evidence has been interpreted as the generation of a diffuse boundary between the India and Australia plates, dated between 8 and 20 Myrs ago (Wiens et al., 1985; Gordon et al., 1998). Ocean-floor deformation at about 8 Myrs is documented from buckling of marine sediments (Weissel et al., 1980). Ongoing deformation in the Indian Ocean is also supported by pronounced (Ms > 6.5) and localized seismicity (Stein and Okal, 1978) along the northern portion of the Ninety East Ridge (see inset in Figure 6), suggestive of left-lateral strike-slip motion. Figure 6 shows the observed convergence history of India and Australia relative to Eurasia since early Miocene based on geodetic (Sella et al., 2002) as well as paleomagnetic (Cande and Stock, 2004; DeMets et al., 1994; Gordon and Jurdy, 1986; Merkouriev and DeMets, 2006) data collected along the Carlsberg and South East Indian ridges (labeled respectively as CBr and SEIr in the inset). Within errorbars, convergence rates are almost indistinguishable between 20 and 11 Myrs ago, suggesting that India and Australia behaved as one single plate with presumably little deformation occurring in between. Over the past 11 Myrs however their convergence to Eurasia differs distinctly. While India slowed down by almost 2 cm/year, convergence of Australia to Eurasia remained almost steady, with only some 0.5 cm/year of reduction. Timing of the India/Eurasia plate-motion change coincides reasonably well with the occurrence of diffuse deformation in the Indian Ocean. More relevant is the fact that Tibet had attained most of its current elevation (Tapponnier et al., 2001) prior to the slowdown of the Indian plate and prior also to the presumed formation of the India/Australia plate boundary, implying that resistive plate

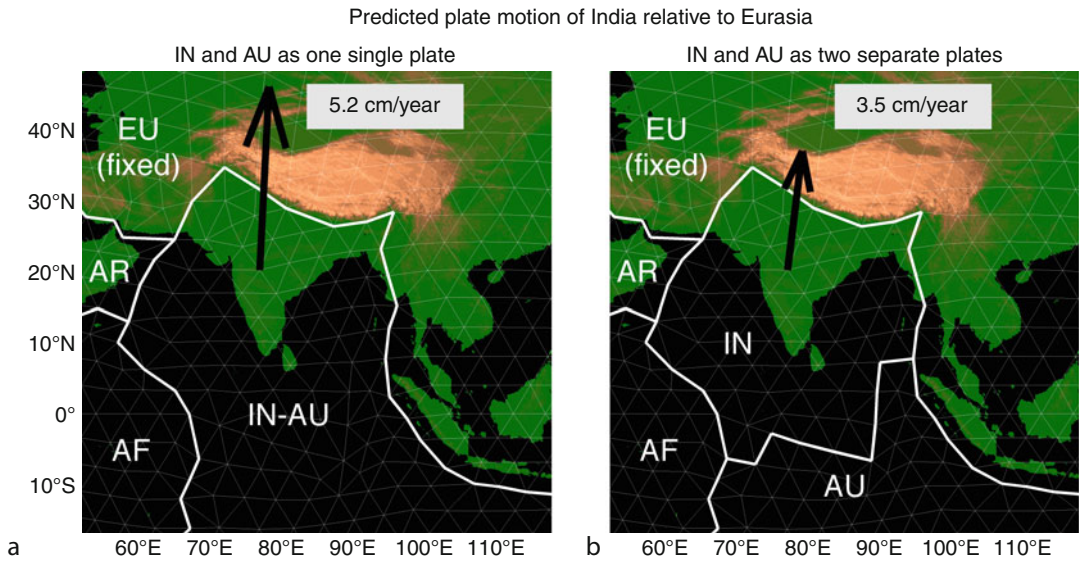


Plate Motions in Time: Inferences on Driving and Resisting Forces, Figure 7 Predicted India (IN) plate motion relative to Eurasia (EU) from two distinct simulations with India and Australia (AU) acting, respectively, (**a**) as one single and (**b**) as two separate plates, where plate boundaries in our computational mesh are shown in *bold white* and finite elements in thin white. Plate motions are computed at long 86°E, lat 27°N. Abbreviations of plate names as in Figure 6. Note that a single India/Australia plate results in a predicted convergence of 5.2 cm/year relative to EU, incompatible with the geodetic estimate (see Figure 12). Two separate plates result in a convergence of 3.5 cm/year of IN relative to EU, similar to the present-day observation. In the latter scenario, resisting forces from the gravitational load of Tibet act only against the smaller India plate, and are thus more effective in slowing the convergent motion.

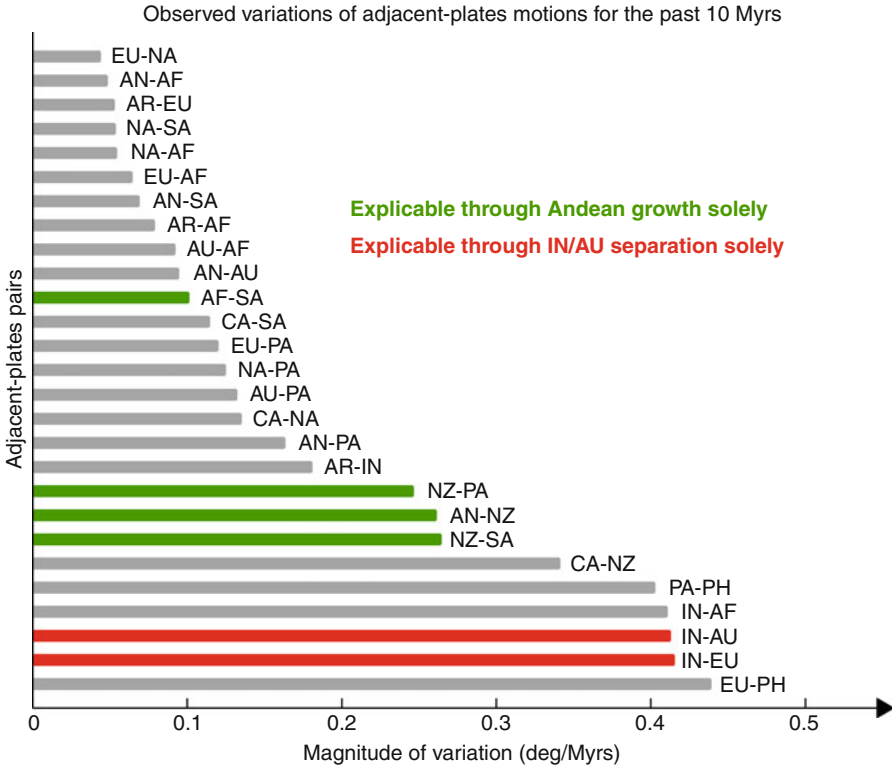


Plate Motions in Time: Inferences on Driving and Resisting Forces, Figure 8 Observed variations of adjacent-plates motions over the past 10 Myrs. Abbreviations of plate names as in Figure 6. For each couple of adjacent plates, variations are computed as magnitude of difference between relative rotation poles at 10 Myrs, derived from paleomagnetic reconstruction, and at present day, obtained through geodetic techniques (GPS). The Cocos oceanic plate is not considered, since a geodetic estimate for its rotation pole is not available. Variations of AF/SA, AN/NZ, NZ/SA, and NZ/PA adjacent-plates systems (*green bars*) can be entirely explained through the effect of Andean growth (see Figures 6 and 7). They amount to about 18% of the global relative motions changes over the past 10 Myrs. Variations of IN/EU as well as IN/AU relative motions (*red bars*) can be entirely explained through the effect of separation between India and Australia (see Figure 7). They amount to about 17% of the global relative motions changes over the past 10 Myrs. Thus, our models of mantle/lithosphere dynamics explicitly predict about 35% of the global plate motion changes observed over the past 10 Myrs from two well-identified tectonic variations.

boundary forces arising from the gravitational load of Tibet were already in place to act against convergence. Iaffaldano and Bunge (2009) tested explicitly whether plate-boundary forces from high Tibet are sufficient to explain the observed reduction of India/Eurasia plate convergence, once the former is separated from Australia by an additional plate boundary (Figure 7). Specifically, they performed two distinct simulations of global plate motions, one with India and Australia cast as one single plate and the other with two plates built into the computational finite-element grid. A single India/Australia plate results in a predicted convergence of 5.2 cm/year at long 86°E, lat 27°N (Figure 7a), whereas India being separated from Australia implies a convergence of 3.5 cm/year at the same position (Figure 7b). The latter prediction compares remarkably well with the geodetic estimate (see Figure 6). Finally, it is worth mentioning that simulations also predict an increased convergence between India and Australia, concentrated in the Indian Ocean, compatible with the aforementioned geologic and geodetic record.

Conclusions

Recent results indicate that joint modeling of the mantle/lithosphere system begins to achieve a level of maturity that allows explicit testing of a range of hypotheses on the force balance in plate tectonics, and identifying key controlling parameters. While buoyancy forces from MCMs contribute significantly to the dynamics of plate motion, it is clear that plate boundary forces are of sufficient magnitude relative to these driving forces to affect plate motions and plate deformation, and to initiate rapid plate motion changes. One key controlling parameter in regulating plate velocity is the elevation of large mountain belts, because their topographic load consumes a considerable amount of the driving forces available in plate tectonics, as much as 10**13 N/m. Along the Nazca/South America plate boundary these forces are sufficient to reduce the convergence rate over the past 10 Myrs by some 30%. This reduction is, however, not an isolated episode of a rapid plate motion slow down. Instead many such variations are documented from the global compilation of Müller et al. (2008), which points to the

importance of topography and erosion in the global tectonic system (Cloetingh et al., 2007). The fact that models accurately predict the spreading history of the Pacific/Nazca, Nazca/South America, Nazca/Antarctica, and South America/Africa plate boundaries is of equal interest. The result is not entirely surprising, and arises from the kinematic constraints of plate tectonics on the sphere. This suggests that far-field effects cannot be neglected in the geologic record at least in some cases. The strong influence of mountain belts on the plate tectonic force balance could have important implications. In an influential paper, Raymo and Ruddiman (1992) advanced the notion that Cenozoic climate change may have been caused by the uplift of Tibet. In other words, the rise of large mountain plateaus may act as a tectonic force on climate (Strecker et al., 2007). Low erosion rates have been implicated as a prerequisite for the creation of large mountain plateaus (Sobel et al., 2003; Clift and Vannucchi, 2004). This implication suggests conversely that climate can act – through large topography – as a force in plate tectonics. Overall, a significant portion of recent changes in global plate motions can be attributed to topography-related forcing along plate boundaries rather than to mantle buoyancy. These findings are summarized in Figure 8, where the relative plate motion changes observed globally over the past 10 Myrs are plotted. Green and red bars show variations in plate motion that are related, respectively, to the growth of the high Andes or to the presumed recent separation between India and Australia, and amount to about 35% of the total change over the Earth surface. This remarkable first-order result clearly demonstrates the ability of plate boundary forces to affect the global plate velocity field. The level of maturity achieved by neo-tectonic simulations coupled with 3-D MCMs thus allows geodynamicists to make explicit predictions of the plate tectonic force balance that can be tested against the geologic record of present and past plate motions.

Bibliography

Allmendinger, R. W., Jordan, T. E., Kay, S. M., and Isacks, B. L., 1997. The evolution of the Altiplano-Puna Plateau of the Central Andes. *Annual Review of Earth and Planetary Sciences*, **25**, 139–74.

Bercovici, D., 1995. A source-sink model of the generation of plate tectonics from non-Newtonian mantle flow. *Journal of Geophysical Research*, **100**, 2013–2030.

Bercovici, D., 2003. The generation of plate tectonics from mantle convection. *Earth and Planetary Science Letters*, **205**, 107–121.

Bird, P., 1998. Testing hypotheses on plate-driving mechanisms with global lithosphere models including topography, thermal structure, and faults. *Journal of Geophysical Research*, **103**, 10115–10129.

Bunge, H.-P., 2005. Low plume excess temperature and high core heat flux inferred from non-adiabatic geotherms in internally heated mantle circulation models. *Physics of the Earth and Planetary Interiors*, **153**, 3–10.

Bunge, H.-P., Richards, M. A., and Baumgardner, J. R., 1996. The effect of depth dependent viscosity on the planform of mantle convection. *Nature*, **379**, 436–438.

Bunge, H.-P., Richards, M. A., and Baumgardner, J. R., 1997. A sensitivity study of 3-D spherical mantle convection at 10**8 Rayleigh number: Effects of depth dependent viscosity, heating mode and an endothermic phase change. *Journal of Geophysical Research*, **102**, 11991–12007.

Bunge, H.-P., Richards, M. A., Lithgow-Bertelloni, C., Baumgardner, J. R., Grand, S. P., and Romanowicz, B. A., 1998. Time scales and heterogeneous structure in geodynamic earth models. *Science*, **280**, 91–95.

Byerlee, J. D., 1978. Friction of rocks. *Pure and Applied Geophysics*, **116**, 1189–1198.

Cande, S. C., and Stock, J. M., 2004. Pacific-Antarctic-Australia motion and the formation of the Macquarie plate. *Geophysical Journal International*, **157**, 399–414.

Clift, P., and Vannucchi, P., 2004. Controls on tectonic accretion versus erosion in subduction zones: implications for the origin and recycling of the continental crust. *Reviews of Geophysics*, **42**, RG2001.

Cloetingh, S. A. P. L., Ziegler, P. A., Bogaard, P. J. F., Andriessen, P. A. M., Artemieva, I. M., Bada, G., van Balen, R. T., Ben-Avraham, Z., Brun, J.-P., Bunge, H.-P., Burov, E. B., Carbonell, R., Facenna, C., Gallart, J., Green, A. G., Heidbach, O., Jones, A. G., Matenco, L., Mosar, J., Oncken, O., Pascal, C., Peters, G., Sli-aupa, S., Soesoo, A., Spakman, W. R. S., Thybo, H., Torsvik, T. H., de Vicente, G., Wenzel, F., Wortel, M. J. R., and the TOPO-EUROPE Working Group, 2007. TOPO-EUROPE: the geoscience of coupled deep earth – surface processes. *Global and Planetary Change*, **58**, 1–118.

Davies, G. F., and Richards, M. A., 1992. Mantle convection. *Journal of Geology*, **100**, 151–206.

DeMets, C., Gordon, R. G., Argus, D. F., and Stein, S., 1994. Effect of recent revisions to the geomagnetic reversal time scale on estimates of current plate motions. *Geophysical Research Letters*, **21**, 2191–2194.

Dixon, T. H., 1991. An introduction to the Global Positioning System and some geological applications. *Reviews of Geophysics*, **29**, 249–276.

Forsyth, D. W., and Uyeda, S., 1975. Relative importance of driving forces of plate motion. *Geophysical Journal of the Royal Astronomical Society*, **43**, 163–200.

Forte, A. M., 2010. Plate driving forces. *Encyclopedia of solid earth geophysics*. Springer 2011.

Forte, A. M., Moucha, R., Rowley, D. B., Quere, S., Mitrovica, J. X., Simmons, N. A., and Grand, S. P., 2009. Recent tectonic plate decelerations driven by mantle convection. *Geophysical Research Letters*, **36**, L23301.

Ghosh, A., Holt, W. E., Wen, L., Haines, A. J., and Flesch, L. M., 2008. Joint modeling of lithosphere and mantle dynamics elucidating lithosphere-mantle coupling. *Geophysical Research Letters*, **35**, L16309.

Gordon, R. G., and Jurdy, D. M., 1986. Cenozoic global plate motions. *Journal of Geophysical Research*, **91**, 12389–12406.

Gordon, R. G., DeMets, C., and Royer, J.-Y., 1998. Evidence for long-term diffuse deformation of the lithosphere of the equatorial Indian Ocean. *Nature*, **395**, 370–374.

Gregory-Wodzicki, K. M., 2000. Uplift history of the central and northern Andes: A review. *Geological Society of America Bulletin*, **112**, 1091–1105.

Hager, B. F., and O'Connel, R. J., 1981. A simple global-model of plate dynamics and mantle convection. *Journal of Geophysical Research*, **86**, 4843–4867.

Hickman, S. H., 1991. Stress in the lithosphere and the strength of active faults. *Reviews of Geophysics*, **29**, 759–775.

Humphreys, E. D., and Coblentz, D. D., 2007. North American dynamics and Western US tectonics. *Reviews of Geophysics*, **45**(2), RG3001.

Iaffaldano, G., and Bunge, H.-P., 2008. Strong plate coupling along the nazca/south america convergent margin. *Geology*, **36**, 443–446.

Iaffaldano, G., and Bunge, H.-P., 2009. Relating rapid plate-motion variations to plate-boundary forces in global coupled models of the mantle/lithosphere system: effects of topography and friction. *Tectonophysics*, **474**, 393–404.
Iaffaldano, G., Bunge, H.-P., and Dixon, T. H., 2006. Feedback between mountain belt growth and plate convergence. *Geology*, **34**, 893–896.
Karato, S., and Jung, H., 2003. Effects of pressure on high-temperature dislocation creep in olivine. *Philosophical Magazine*, **83**, 401–414.
Kirby, S. H., 1983. Rheology of the lithosphere. *Reviews of Geophysics*, **21**, 1458–1487.
Koeppen, W., and Wegener, A., 1924. *Die Klimate der geologischen Vorzeit*. Berlin: Borntraeger.
Kohlstedt, D. L., Evans, B., and Mackwell, S. J., 1995. Strength of the lithosphere: constraints imposed by laboratory experiments. *Journal of Geophysical Research*, **100**, 17587–17602.
Kong, X., and Bird, P., 1995. SHELLS: a thin-shell program for modeling neotectonics of regional or global lithosphere with faults. *Journal of Geophysical Research*, **100**, 22129–22132.
Merkouriev, S., and DeMets, C., 2006. Constraints on India plate motion since 20 Ma from dense Russian magnetic data: implications for Indian plate dynamics. *Geochemistry Geophysics Geosystems*, **7**, Q02002.
Moresi, L., and Solomatov, V., 1998. Mantle convection with a brittle lithosphere: thoughts on the global tectonic styles of the Earth and Venus. *Geophysical Journal International*, **133**, 669–682.
Morgan, W. J., 1968. Rises, trenches, great faults and crustal blocks. *Journal of Geophysical Research*, **73**, 1959.
Müller, R. D., Sdrolias, M., Gaina, C., and Roest, W. R., 2008. Age, spreading and spreading asymmetry of the world's ocean crust. *Geochemistry Geophysics Geosystems*, **9**, Q04006.
Norabuena, E. O., Dixon, T. H., Stein, S., and Harrison, C. G. A., 1999. Decelerating Nazca-South America and Nazca-Pacific plate motions. *Geophysical Research Letters*, **26**, 3405–3408.
Oeser, J., Bunge, H.-P., and Mohr, M., 2006. Cluster design in the earth sciences: TETHYS, high performance computing and communications – second international conference, HPCC 2006, Munich, Germany. *Lecture Notes in Computer Science*, **4208**, 31–40.
Quere, S., and Forte, A. M., 2006. Influence of past and present-day plate motions on spherical models of mantle convection: implications for mantle plumes and hotspots. *Geophysical Journal International*, **165**, 1041–1057.
Raymo, M. E., and Ruddiman, W. F., 1992. Tectonic forcing of late Cenozoic climate. *Nature*, **359**, 117–122.
Sella, G. F., Dixon, T. H., and Mao, A., 2002. REVEL: a model for recent plate velocities from space geodesy. *Journal of Geophysical Research*, **107**, 2081–2111.
Sobel, E. R., Hilley, G. E., and Strecker, M. R., 2003. Formation of internally drained contractional basins by aridity-limited bedrock incision. *Journal of Geophysical Research*, **108**, 2344.
Stein, R. S., 1987. Contemporary plate motion and crustal deformation. *Reviews of Geophysics*, **25**(5), 855–863.
Stein, S., and Okal, E., 1978. Seismicity and tectonics of the Ninetyeast Ridge area: evidence for internal deformation of the Indian plate. *Journal of Geophysical Research*, **83**, 2233–2246.
Strecker, M. R., Slonso, R. N., Bookhagen, B., Carrapa, B., Hilley, G. E., Sobel, E. R., and Trauth, M. H., 2007. Tectonics and climate of the Southern Central Andes. *Annual Review of Earth and Planetary Sciences*, **35**, 747–787.
Suppe, J., 2007. Absolute fault and crustal strength from wedge tapers. *Geology*, **35**(12), 1127–1130.
Tackley, P. J., Stevenson, D. J., Glatzmaier, G. A., and Schubert, G., 1994. Effects of multiple phase transitions in a three-dimensional spherical model of convection in Earth's mantle. *Journal of Geophysical Research*, **99**, 15877–15902.
Tapponnier, P., Zhiquin, X., Roger, F., Meyer, B., Arnaud, N., Wittlinger, G., and Jingsui, Y., 2001. Oblique stepwise rise and growth of the Tibet plateau. *Science*, **294**, 1671–1677.
van der Hilst, R. D., de Hoop, M. V., Wang, P., Shim, S.-H., Ma, P., and Tenorio, L., 2007. Seismostratigraphy and thermal structure of Earth's core-mantle boundary region. *Science*, **315**, 1813–1817.
Weissel, J. K., Anderson, R. N., and Geller, C. A., 1980. Deformation of the Indo-Australian plate. *Nature*, **287**, 284–291.
Wiens, D. A., DeMets, C., Gordon, R. G., Stein, S., Argus, D., Engeln, J. F., Lundgren, P., Quible, D., Stein, C., Weinstein, S., and Woods, D. F., 1985. A diffuse plate boundary model for Indian Ocean tectonics. *Geophysical Research Letters*, **12**, 429–432.
Wilson, J. T., 1965. A new class of faults and their bearing on continental drift. *Nature*, **207**, 343–347.
Zhong, S., Zuber, M. T., Moresi, L., and Gurnis, M., 2000. Role of temperature-dependent viscosity and surface plates in spherical shell models of mantle convection. *Journal of Geophysical Research*, **105**, 11063–11082.

Cross-references

Continental Drift
Geodesy, Ground Positioning and Leveling
Geodesy, Networks and Reference Systems
GPS, Tectonic Geodesy
Lithosphere, Oceanic
Plate Driving Forces
Plates and Paleoreconstructions

PLATE TECTONICS, PRECAMBRIAN

Y. J. Bhaskar Rao, E. V. S. S. K. Babu
National Geophysical Research Institute, Council of Scientific and Industrial Research, Hyderabad, India

Introduction

The plate tectonics paradigm, developed in the late 1960s, provides an excellent framework to explain the tectonics of the Earth's crust since the Cenozoic period. In general, the tectonic style reflects how the planets lose their internal heat and plate tectonics represent one of the at least three styles in which planets cool (Sleep, 2007). Intriguingly, Earth is the only planet that presently exhibits plate tectonics (Stevenson, 2003), an unusual mode of cooling. It is now widely appreciated that the excess density of the oceanic lithosphere that is sinking deeply in subduction zones drives the modern-style plate tectonics ("subduction tectonics," Stern, 2005). However, *why and when on Earth plate tectonics began* has been highly controversial and remains one of the most challenging unresolved problems in our understanding of the evolution of the planet (for a recent review on the subject see Condie and Pease, 2008 and articles therein; Ernst, 2009). While many authors propose that plate tectonics, in some form, operated since early Archaean times (Kroner, 1981; Ernst, 1983; Sleep, 1992; Parman et al., 2001; Smithies et al., 2003, 2005; Sleep, 2005; Condie, 2005; Cawood et al.,

2006; Condie and Kroner, 2008; Dilek and Polat, 2008), others favor a much later beginning in the Earth's history (Davies, 1992; Hamilton, 1998, 2003; Stern, 2005, 2007; Hamilton, 2007; Dewey, 2007; Brown, 2008; Ernst, 2009).

Demonstration that plate tectonics operated at any given time requires evidence for lithospheric subduction and independent plate motions. However, since the oldest oceans on present Earth are not older than Jurassic, one is forced to consider the geological record on continents that dates back to ~4.4 Gyr (billion years). This section summarizes the debate on the onset of modern-style plate tectonics on Earth, but before addressing this, an overview of the current perspectives on modern-style plate tectonics in terms of thermal evolution of Earth and other silicate planets is considered pertinent. Subsequently, the footprints of plate tectonics in the deep geological past, the Precambrian (covering an early part of the Earth's history, including, Hadean- prior to 3.8 Gyr, Archaean-3.8–2.5 Gyr and Proterozoic-2.5–0.54 Gyr) are traced through a consideration of empirical observations and multidisciplinary datasets that are believed to serve as proxies for modern-style plate tectonics in the ancient geologic record.

Modern-style plate tectonics: some broad perspectives

In its initial conceptualization, plate tectonics offered a purely kinematic description of how lithospheric plates move and interact with one another on the nearly spherical surface of the Earth. There was little clarity on the forces that cause plate motions. Notwithstanding the traditional view that the asthenospheric convection powers contemporary plate motions, modern geodynamicists find an increasing consensus for predominance of "top-down tectonics" (Anderson, 2001), at least during some stages in the Earth's history, and suggest that plates drive themselves as they organize the overall mantle convection (Davies, 1999; Hynes, 2005).

The thermal boundary layer – the lithosphere – is buoyant when new oceanic crust is created at the Mid Oceanic Ridges (MOR). As the lithosphere ages and moves away from the MOR, it cools (conductively), thickens, and increases in density leading to a buoyancy-crossover in the lithosphere–asthenosphere system. On the present Earth, such a crossover occurs in 20–40 Myr (million years). This "ridge-push" coupled with "gravitational descent" of the lithosphere in the subduction zones contributes to much of the plate motion, so much so that the modern plate tectonic style can be referred to as "subduction tectonics" (Stern, 2005). Based on seismic tomography, it is evident that the lithospheric slabs on Earth reach great depths, typically 1100–1300 km, and possibly down to the core-mantle boundary (Vander Hilst et al., 1997). While the "slab pull" of the sinking lithosphere and the associated "slab suction" of the entrained ambient mantle at subduction zones move the plates (Conrad and Lithgrow-Bertelloni, 2004), the viscous resistance of the asthenosphere impedes plate motion. Resistance to the sinking lithosphere is also from dissipation forces that are manifest as deformation (mainly faulting and bending) in the subducting slab and phenomena such as regional mantle down swelling. The nature and extent of equilibrium between the driving and resisting forces is indicated by the limited range of the observed plate velocities, which possibly did not exceed ~17 cm/y, for instance, during the Mesozoic drift of the Indian lithospheric plate, a case of extremely rapid plate motion.

As stated earlier, Plate tectonics seems to be an unusual way for silicate planets to cool, and perhaps represents a specific stage in the ~4.57 Gyr cooling history of the Earth (Ernst, 2009 and references cited therein). This implies a temporal change in the cooling/tectonic style of the planet. Ernst (2007) suggested that the Earth's plate tectonic history is divisible into four stages. In the early three stages, plate motions were controlled by asthenospheric convection dragging buoyant lithosphere down inwards. Contrary to this, during the fourth stage alone, the subducting lithosphere causes plate motions. The latter stage is referred to as the "Proterozoic plate tectonic regime," while the early stages comprise "Proto-plate tectonics" (Stern, 2007) broadly analogous to delamination tectonics and/or the shallow subduction mode of tectonics (Foley et al., 2003).

Evolution of earth's tectonic regimes

Numerical modeling, geochronological and isotopic results on meteorites, and the oldest known terrestrial rocks and minerals, especially ex-situ magmatic zircons in ancient sediments, offer valuable constraints on the physical conditions of the planet during its early history. A general consensus is that the condensation of the accreting hot Earth was completed by ~4.57 Gyr (Ernst, 2009 and references cited therein). Further rapid increase in its internal heat is ascribed to core-formation, radioactive decay of long-lived elements (U, Th, ^{40}K) and short-lived isotopes (e.g., ^{26}Al), bolide impacts and very early impact of a Mars-sized body and such other factors (Wetherill, 1976; Kleine et al., 2004; Cameron and Benz, 1991; Halliday, 2004). As illustrated in Figure 1 (Stage I), this phase was succeeded by the formation of a near-surface magma mush ocean several hundred kilometers thick (Abe, 1997; Boyet and Carlson, 2005; Davies, 1999), prior to 4.4 Gyr. The Earth cooled rapidly thereafter leading to the formation of a thin silicate-proto crust by ~4.4 Gyr, consistent with the well-documented occurrence of 4.2–4.4 Gyr igneous zircons from the Archaean sediments of Western Australia (Wilde et al., 2001; Watson and Harrison, 2005). It is generally perceived that such a Hadean crust may have been recycled back into the mantle by viscous drag and mantle overturn (Boyet and Carlson, 2005). Significant cooling of the Earth's crust (below the boiling point of water) by ~4.0 Gyr is inferred from preservation of the ~3.9 Gyr Isua volcano-sedimentary assemblages, western Greenland, which suggest the possibility that water-oceans marked the planet's outer rind by the Late Hadean times. Presence of microdiamond ± graphite inclusions in the detrital zircons

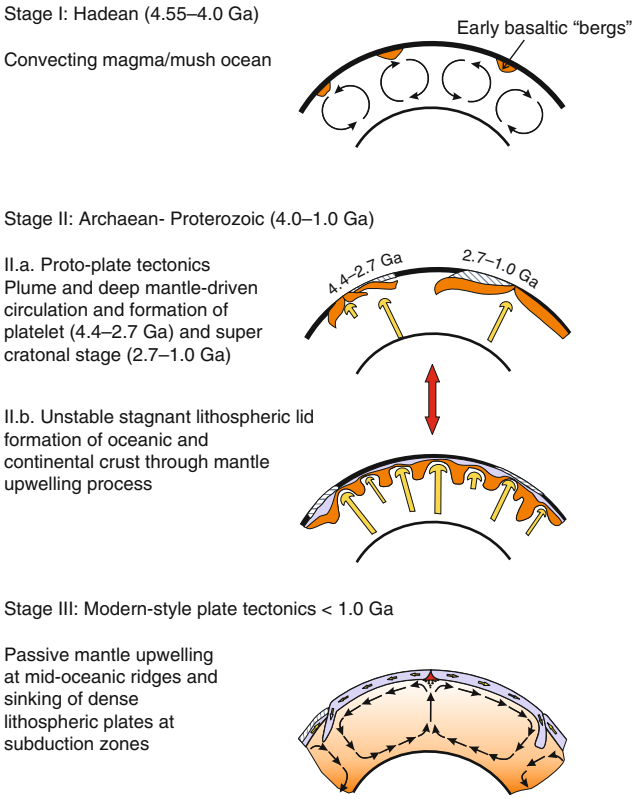


Plate Tectonics, Precambrian, Figure 1 Cartoon illustrating the evolution of global tectonic styles from the Hadean to present involving three main stages. See text for more details (Modified after Ernst, 2007; Stern, 2008 and references cited therein).

(Menneken et al., 2007) suggests thick Archaean lithosphere, at least locally. However, if these inclusions were initially graphite, but transformed to diamonds subsequently through subduction processes, the thick lithosphere can be inferred much later, during the Mid Archaean rather than Hadean/Early Archaean (Williams, 2007). Indeed, Mid Archaean diamonds have also been sourced in younger kimberlites from several cratons, suggesting thick lithosphere and possible plate motion by the Mid Archaean (Shirey et al., 2004).

Thermal models of planetary evolution suggest three principal modes of non-radiative planetary heat transfer (Sleep, 2000, 2007): (1) magma mush oceans, (2) plume and plate tectonics and (3) stagnant-lid convection (illustrated in Figure 1 in corresponding stages). The magma mush ocean mode is, by far, the most energetic mode and probably characterizes the Hadean tectonic style. The stagnant-lid convection, the least energetic mode, characterizes the present Venus and Mars and is possibly present on Earth episodically. Much of the modern, Wilson-style plate tectonics reflect convective overturn in the suboceanic lithosphere–asthenosphere system. On the other hand, the craton-capped lithosphere scenario mimics the stagnant-lid mode where the build-up of sublithospheric temperature leads to episodic break up and drift of continents (Gurnis, 1988; Lowman and Jarvis, 1996; Silver and Behn, 2008). In this process, some low-density lithosphere is returned to the mantle by viscous drag. However, the high-brittle yield strength combined with chemical and mineralogical buoyancy and high viscosity of the cratonic roots promote long-term stabilization of cratonic lithosphere compared to the oceanic lithosphere. While the latter is continuously recycled back into the mantle, the cratonic lithosphere accumulates over time and forms a time-integrated aggregate. Mantle convection is driven by both bottom-up and top-down thermal and gravitational instabilities: hot-buoyant plumes of mantle overturn and cold-dense sinking lithospheric slabs (Davies, 1999 and Anderson, 2001). It is further perceived that in the Hadean Earth, the widespread vigorous bottom-up overturn and plume ascent predominated (Condie, 1994, 2005; Sleep, 2007). High-temperature regimes and low-mantle viscosities may have led to rapidly moving plates and avalanches of lithospheric descent into the lower mantle interspersed with intervals of relative quiescence (O'Neill et al., 2007). In contrast, by the Phanerozoic time, thermal heterogeneities in the deep-mantle lessened, and progressively thicker and cooler oceanic lithospheric plates became negatively buoyant leading to the foundering and driving mantle convection from the surface downward. Craton-capped lithospheric plates being chemically and mineralogically buoyant would stabilize at the Earth's surface. Stagnant-lid tectonics is a dominant mode of other silicate bodies today, and it is likely that Earth also experienced some episodes of this mode. However, in the Hadean, stagnant-lid tectonics may have been very unstable (Hamilton, 2007).

Through a correlation between mantle potential temperatures (T_P) – a measure of mantle temperature adiabatically decompressed to pressures corresponding to the planetary surface, and the Urey ratio (heat production/heat loss) for progressively cooling silicate bodies, Sleep (2000) showed that whereas the stagnant-lid style encompasses a wide range of thermal conditions from thermally "active" to "dead" silicate bodies, the modern plate tectonics can only occur when appropriate mantle thermal conditions exist. Plate tectonics are thus an effective mode of planetary heat loss, but conditions favorable for this style may only exist for a relatively brief interval in planetary evolution. Ernst (2007) showed that a Urey ratio of ~0.3 extrapolated back in time yields a "thermal catastrophe" at ~1.0–1.5 Gyr (Korenaga, 2006).

The density of the present oceanic lithosphere is controlled largely by the thickness of its mantle component. Although a small part, the oceanic crustal thickness is controlled by T_p that determines the amount of mantle melting for a given amount of mantle upwelling (McKenzie and Bickle, 1988). The mantle potential temperature for the Archaean Earth was probably 300–500°C higher than today (Nisbet et al., 1993). Thus, decompression melting of the hotter Archaean mantle must have generated thicker oceanic crust and a more buoyant lithosphere. Presently, the oceanic lithosphere reaches neutral buoyancy in about

20 million years (Davies, 1992), while in a hotter Archaean mantle, it would take longer (approximately 80 Myr) to reach neutral buoyancy and much longer lithosphere–asthenosphere crossover times. In the Arcahean, the negatively buoyant lithosphere would have been removed much more rapidly than it could form and subduction may have been possible intermittently, considering the crust was denser and hence gravitational instability would also favor subduction. This scenario is in contrast to the protracted time through which subduction is sustained in the Phanerozoic.

The transition from stagnant-lid to plate-tectonic modes requires lithospheric rupture, subsidence and asthenosphere flow over the lithosphere. Mechanisms for efficient cooling of relatively hotter Archaean lithospheric plates may include three possible scenarios: (1) greater length of the Archaean ocean ridge system (Hargraves, 1981), about 27 times to the present length, (2) inversion of basalt to eclogite in buoyantly subducted Archaean oceanic crust, and (3) negatively buoyant oceanic mantle lithosphere because of latent heat loss associated with melt extraction at MORs.

Criteria for recognizing plate tectonics in the precambrian geological record

Recognizing the onset of modern-style plate tectonics on planet Earth involves tracking the manifestation of the processes in the geological record. For instance, evidences for the creation of new oceanic lithosphere, the existence of "ocean-plate stratigraphy," subduction processes where lithosphere is returned to the mantle, transform and transcurrent faulting at any given time. The problem of preservation of the diagnostic criteria figures prominently, as the evidence in all likelihood, can be fragmentary and incomplete owning to obliteration by tectonic and surface processes. From the theoretical considerations presented in the foregoing sections, the modern-style plate tectonics are not only unusual in a planetary perspective, but also in the Earth's evolutionary history. Thus, the Huttonian approach "The present is the key to the past" may not be entirely helpful to understand the Earth's earliest tectonic history (Rollinson, 2007; Stern, 2008). In any case, best approaches should consider multiple working hypothesis and methodological uniformitarianism emphasizing upon constancy of natural laws rather than substantive uniformitarianism invoking uniform geological processes through time (Stern, 2008). Some of the most robust criteria for recognizing the operation of the plate tectonics in the past include: ophiolites, blueschists and Ultra High-Pressure (UHP) metamorphic belts, eclogites, passive margins, transform faults, paleomagnetic evidence, igneous rocks with subduction-related geochemistry and isotopic evidences for recycling. In general, the Archaean crust is typified by granite-greenstone terrains, with characteristic rock types such as komatiitic ultramafics and basalts and TTG gneisses. Archaean terrains generally preclude mafic alkaline igneous rocks, undisputedly recognized ophiolites, widespread multicyclic sediments, UHP terrains and blueschist-eclogite metamorphic assemblages. Several of these form distinctive lithologies ascribed to Wilson-style plate-tectonic processes prevalent in the Phanerozoic times. Lithological associations such as stratigraphically intact platformal sedimentary formations do not appear unequivocally until the latest Archaean time. Such temporal distribution of rock formations are considered to reflect changes in tectonic style (Stern, 2007, 2008 and references cited therein).

Evidence for existence of paired metamorphic belts and subduction-related batholiths also form useful criteria. In general, different authors apply the criteria in different ways. For example, while some consider when a particular assemblage or property first appeared, others emphasize on when it became common. Thus, the inferences on the onset of plate tectonics on the Earth could be quite subjective, even by the same criterion. A broad description and significance of some of the important criteria are given below.

Ophiolites: Ophiolites are slices of the oceanic lithosphere tectonically emplaced onto the continental crust and could be reliable indicators of past plate tectonics. A complete ophiolite sequence (ocean-plate stratigraphy) would include pelagic sediments, pillow basalt, sheeted dykes, gabbros, and tectonised ultramafites (lherzolites and harzburgites). However, generally, fragmentary and dismembered sequences are encountered in Precambrian terrains. Well- preserved ophiolites are known from many Neoproterozoic terrains (up to ca. 1.0 Gyr old; Stern, 2008 and Ernst, 2007 and 2009). Some of the better-preserved Neoproterozoic ophiolites include, Purtuneniq ophiolites, Canada (Scott et al., 1992) and Jormua ophiolites, Finland (Peltonen et al., 1996), although ophiolites as old as ca. 3.8 Gyr (the Isua suite, Greenland; Furnes et al., 2007) have been proposed, they remain controversial (also see Bickle et al., 1994).

Blueschists: Blueschists are metamorphosed mafic rocks bearing predominantly a sodic amphibole that stabilizes at high-P, low-T typical of Pacific-type of orogenic belts. Best studied examples are the Franciscan and Sanbagwa terrains (Ernst, 2003). The oldest known blueschists are Neoproterozoic (800–700 Myr; Maruyama et al., 1996) and undisputed pre-Neoproterozoic blueschists are virtually unknown.

Eclogites: Eclogites (s.s) are metabasalts and metagabbros that equilibrated at $P > 1.2$ GPa (> 45 km) and T (> 600–$650°C$) with a characteristic pyropic-garnet and omphacitic-clinopyroxene-bearing assemblage. Different types of eclogites have been described, but Colman's group C (associated with glaucophane schists) and the "low-T eclogites" (Carswell, 1990) could be reliable indicators of subduction process.

UHP terrains: UHP terrains form when continental crust initially subducted to greater than 100 km returns to the surface through subsequent tectonic processes.

Typical assemblages include coesite and diamond, indicative of 700–900°C and 3–4 GPa, or more (Liou et al., 2004). The oldest known UHP terrain is in Mali, ca. 620 Myr (Jahn et al., 2001) and Kazakhsthan, 530 Myr (Maruyama and Liou, 1998).

Paired metamorphic belts: The occurrence of paired metamorphic belts was originally conceptualized by Miyashiro (1972). They serve as one of the diagnostic criteria for modern-style plate tectonics and typically form on the overriding plate with the high-P low-T (blueschist facies) terrane forming next to the trench, whereas the low-P high-T terrane (greenschist facies) develops near to the arc. Paired metamorphic belts represent thermal structure of a mature subduction zone (Kelemen et al., 2003 and van Keken et al., 2002). As these belts are usually very wide, tens of hundreds of kilometers, they could be robust features that survive several episodes of orogeny and erosion (Stern, 2008). As only ~30% of convergent plate boundaries are accretionary, a large proportion, remain erosional (Clift and Vannucchi, 2004). In a paired metamorphic belt, accretionary convergent margins are preserved as broad zones, whereas the erosional boundaries could be much narrower (Stern, 2008 and references cited therein).

Batholiths: Mere presence of magmatic "island arc" features suggests modern-style plate tectonics with deep subduction (100–250 km). In such a setting from older terrains, volcanic rocks are usually removed by erosion, leaving behind well-preserved mid-crustal batholiths (Stern, 2002, Hamilton and Myers, 1967). Felsic plutonic lithologies represented by TTG suites (Tonalite-Trondjhemite-Granodiorite) dominate Precambrian plutons coalesced into batholiths. Subduction-related batholiths also serve as polarity indicators when combined with other evidences such as grade of metamorphism. etc.

Composition of igneous rocks: Considerable geochemical evidence in terms of depleted mantle source signatures for mid-crustal magmatic complexes suggests relevance of plate tectonic processes by the end of Hadean and in subsequent periods (Shirey et al., 2008). Fluid mobile elements such as K, Sr, Pb, HFSE ratios such as Nb/Th, Th/U, rare earth-element chemistry and other trace element compositions are often used to establish the 'arc-like' signatures in igneous suites (Kerrich and Polat, 2006 and references cited therein), which support the general thesis of plate tectonics since the Early Archaean.

Palaeomagnetic measurements: Palaeomagnetic data indicate differential movement of continents in pre-Neoproterozoic times (Pesonen et al., 2003; Cawood et al., 2006 and references cited therein).

Isotopic compositions: So far, the strongest evidence for crust-mantle differentiation is from the isotopic record of key rock units. There is enough evidence for the recycling of early crust and sediments into the mantle since Early Archaean (Shirey et al., 2008; Wilde et al., 2001 and Valley et al., 2006 and references cited therein). Isotopic studies on diamonds, for example ^{33}S from the inclusions preserved in the Archaean diamonds, provide the best evidence that the surface materials of the early Earth were recycled into the mantle (Farquhar et al., 2002). However, such recycling could also be due to proto-plate-tectonic activity or delamination associated with stagnant lid (Stern, 2008).

Summary and conclusions

The onset of modern-style plate tectonics on Earth remains controversial. A critical evaluation of the Precambrian geological record in terms of many of the criteria considered in the foregoing reveals that plate tectonics were operational in some form in some places on the planet possibly during the Early-Meso Archaean, but became widespread since ~2.7 Gyr. The problem however, concerns three of the indicators, namely, ophiolites, UHP metamorphism, and Blue schists, which clearly suggest much later onsets of more specifically, modern-style plate tectonics around 1.0 Gyr (Stern, 2005; Ernst, 2009; Condie and Kroner, 2008 and references cited therein). A closer consideration of these three criteria in terms of the inferred physical conditions prevalent in the Hadean and Archaean assumes importance. If the Archaean oceanic crust were significantly thicker than the modern oceanic crust (Sleep and Windley, 1982 and Foley et al., 2003), the deeper layers of the oceanic crust may not get accreted to the continents in the subduction zones. Thus, the rock associations such as, layered gabbro, ultramafic cumulates could be missing from the Archaean ophiolites, which may have been recycled into the mantle (Condie and Benn, 2006). From a different perspective, obduction of thin oceanic crust also results in delamination of the middle-to lower-crustal sections and the mantle lithosphere during successive collisional tectonic events. Thus, the preservation of only the upper pillow lavas of the oceanic crust (oceanic flake tectonics of Hoffman and Ranalli, 1988) may be expected. The widespread occurrence of mafic plain type greenstones in the Archaean may be remnants of oceanic crust than oceanic plateaus (Tomlinson and Condie, 2001). The apparent absence of blueschists and UHP rocks before ca. 1.0 Gyr (especially in the Archaean rock record) could also be ascribed to the relatively hotter Archaean Earth, wherein the subduction geotherms did not pass into blueschist stability field (Peacock, 2003; Ernst, 2009; Condie and Kroner, 2008 and references cited therein). Indeed the thermo-barometric studies of Archaean metamorphic rocks suggest steeper subduction geotherms (Moyen et al., 2006). Further, it is possible that the rate of uplift of UHP rocks may have been very slow in the Archaean, such that these assemblages recrystallized and hence no longer recognized.

Although the most robust evidences (temporal distribution of ophiolites, UHP metamorphism, and blueschist terrains) lead us to the thesis that modern-style of plate tectonics was undisputedly operational in the Neoproterozoic time, there are no evidences to suggest that manifestation in the geologic records is not strictly

synchronous. As summarized by Condie and Kroner (2008), it is quite unlikely that the plate tectonics on the Earth began as single global event at a distinct time in the early Archaean rather, it is probable that it began locally and progressively became more widespread from early to the late Archaean. By the late Archaean (2.7 Gyr), steep subduction similar to modern times was widespread on the planet. O'Neill et al. (2007) suggested that subduction may have been episodic during the Precambrian, possibly explaining the peak in crustal production at 2.7 Gyr.

Bibliography

Abe, Y., 1997. Thermal and chemical evolution of the terrestrial magma ocean. *Physics of the Earth and Planetary Interiors*, **100**, 27–39.

Anderson, D. L., 1987. Thermally induced phase changes, lateral heterogeneity of the mantle, continental roots, and deep slab anomalies. *Journal of Geophysical Research*, **90**, 13965–13980.

Anderson, D. L., 2001. Top-down tectonics? *Science*, **293**, 2016–2018.

Bickle, M. J., Nisbet, E. G., and Martin, A., 1994. Archean greenstone belts are not oceanic crust. *Journal of Geology*, **102**, 121–138.

Bowring, S. A., and Williams, I. S., 1999. Priscoan (4.00–4.03Ga) Orthogneisses from north-Western Canada. *Contributions to Mineralogy and Petrology*, **134**, 3–16.

Boyet, M., and Carlson, R. W., 2005. 142^{Nd} evidence for early (>4.53) global differentiation of the silicate Earth. *Science*, **309**, 576–589.

Brown, M., 2008. Characteristic thermal regimes of plate tectonics and their metamorphic imprint through Earth history. In Condie, K. C., and Pease, V. (eds.), *When Did Plate Tectonics Begin on Planet Earth?* Geological Society of America Special Paper 440, Colorado, USA, pp. 97–128.

Cameron, A. G. W., and Benz, W., 1991. Origin of the moon and the single impact hypothesis IV. *Icarus*, **92**, 204–216.

Carswell, D. A., 1990. Eclogites and the eclogite facies: definitions and classification. In Carswell, D. A. (ed.), *Eclogite Facies Rocks*. Glasgow: Blackie, pp. 1–13.

Cawood, P. A., Kroner, A., and Pisarevsky, S., 2006. Precambrian plate tectonics: criteria and evidence. *GSA Today*, **16**, 4–11.

Clift, P., and Vannucchi, P., 2004. Controls on tectonic accretion versus erosion in subduction zones: implications for the origins and recycling of the continental crust. *Reviews of Geophysics*, **42**, doi:10.1029/2003RG000127.

Coleman, R. G., Lee, D. E., Beatty, L. B., and Brannock, W. E., 1965. Eclogites and eclogites: their differences and similarities. *Geological Society of America Bulletin*, **76**(5), 438–508.

Condie, K. C., 1994. Greenstones through time. In Condie, K. C. (ed.), *Archean Crustal Evolution*. Amsterdam: Elsevier, pp. 85–120.

Condie, K. C., 2005. *Earth is an Evolving Planetary System*. Amsterdam: Academic, p. 447.

Condie, K. C., and Benn, K., 2006. Archean geodynamics: similar to or different from modern geodynamics? *American Geophysical Union, Monograph*, **164**, 47–59.

Condie, K. C., and Kroner, A., 2008. When did plate tectonics begin? Evidence from the geologic record. In Condie, K. C., and Pease, V. (eds.), *When Did Plate Tectonics Begin on Planet Earth?* Geological Society of America Special Paper 440, pp. 281–294.

Condie, K. C., and Pease, V. (eds.), 2008. *When Did Plate Tectonics Begin on Planet Earth?* Geological Society of America Special Paper 440, Boulder, Colorado, USA.

Conrad, C. D., and Lithgrow-Bertelloni, C., 2004. The temporal evolution of plate driving forces: Importance of 'slab suction' versus 'slab pull' during the Cenozoic. *Journal of Geophysical Research*, **109**, B10407.

Davies, G. F., 1992. On the emergence of plate tectonics. *Geology*, **20**, 963–966.

Davies, G. F., 1999. *Dynamic Earth: Plate, Plumes, and Mantle Convection*. New york: Cambridge University Press.

Dewey, J. F., 2007. The secular evolution of plate tectonics and the continental crust: an outline. In Hatcher Jr., R. D., Carlson, M. P., McBride, J. H., and Martinez Catalan, J. R. (eds.), *4-D Framework of Continental Crust*, Geological Society of America Memoir, V. 200, pp. 1–7.

Dilek, Y., and Polat, A., 2008. Suprasubduction zone ophiolites and archean tectonics. *Geology*, **36**, 431–432.

Ernst, W. G., 1983. The early earth and archean rock record. In Schopf, J. W. (ed.), *The Earth's Earliest Biosphere: Its Origin and Evolution*. Princeton: Princeton University, pp. 41–52.

Ernst, W. G., 2003. High-pressure and ultrahigh-pressure metamorphic belts-Subduction, recrystallization, exhumation, and significance for ophiolite studies. In Dilek, Y., and Newcombe, S. (eds.), *Ophiolite Concept and Evolution of Geological Thought*. Geological Society of America Special Paper, 373, pp. 365–384.

Ernst, W. G., 2007. Speculations on evolution of the terrestrial lithosphere-asthenosphere system-Plumes and plates. *Gondwana Research*, **11**, 38–49.

Ernst, W. G., 2009. Archean plate tectonics, rise of Proterozoic supercontinentality and onset of regional, episodic stagnant-lid behaviour. *Gondwana Research*, **15**, 243–253.

Farquhar, J., Wing, B. A., McKeegan, K. D., Harris, J. W., Cartigny, P., and Thiemens, M. H., 2002. Mass-independent sulfur of inclusions in diamond and sulfur recycling on the early Earth. *Science*, **298**, 2369–2372.

Foley, S. F., Buhre, S., and Jacob, D. E., 2003. Evolution of the Archean crust by shallow subduction and recycling. *Nature*, **421**, 249–252.

Furnes, H., de Witt, M., Staudigel, H., Rosing, M., and Muehlenbachs, K., 2007. A vestige of earth's oldes ophiolite. *Science*, **315**, 1704–1707.

Gurnis, M., 1988. Large-scale mantle convection and the aggregation and dispersal of supercontinents. *Nature*, **332**, 695–699.

Halliday, A. N., 2004. Mixing, volatile loss and compositional change during impact-driven accretion of the Earth. *Nature*, **427**, 505–509.

Hamilton, W. B., 1998. Archaean magmatism and deformation were not products of plate tectonics. *Precambrian Research*, **91**, 143–179.

Hamilton, W. B., 2003. An alternative Earth. *GSA Today*, **13**(11), 4–12.

Hamilton, W. B., 2007. Earth's first two billion years-The era of internally mobile crust. In Hatcher, R. D., Jr., Carlson, M. P., McBride, J. H., and Martinez, J. R. (eds.), *4-D Framework of Continental Crust*. Geological Society of America Memoir, pp. 233–296.

Hamilton, W. B., and Myers, W. B., 1967. The nature of batholiths, U.S. *Geological Survey Report*, **554-C**, C1–C30.

Hargraves, R. B., 1981. Precambrian tectonic style: a liberal uniformitarian interpretation. In Kroner, A. (ed.), *Precambrian Plate Tectonics*. Amsterdam: Elsevier, pp. 21–56.

Hoffman, P. F., and Ranalli, G., 1988. Archean oceanic flake tectonics. *Geophysical Research Letters*, **15**, 1077–1080.

Hynes, A., 2005. Buoyancy of the oceanic lithosphere and subduction initiation. *International Geology Review*, **47**, 938–951.

Jahn, B. M., Caby, R., and Monie, P., 2001. The oldest UHP eclogites of the world: age of UHP metamorphism, nature of protoliths and tectonic implications. *Chemical Geology*, **178**, 143–158.

Kelemen, P. B., Rilling, J. L., Parmentier, E. M., Mehl, L., and Hacker, B. R., 2003. Thermal structure due to solid-state flow in the mantle wedge beneath arcs. In Eiler, J., and Hirschmann, M. (eds.), *Inside the Subduction Factory*. American Geophysical Union: Washington, DC, pp. 293–311.

Kerrich, R., and Polat, A., 2006. Archaean greenstone-tonalite duality: thermochemical mantle convection models or plate tectonics in the early Earth global geodynamics? *Tectonophysics*, **415**, 141–165.

Kleine, T., Mezger, K., Palme, H., and Munker, C., 2004. The W isotope evolution of the bulk silicate Earth – constraints on the timing and mechanisms of core formation and accretion. *Earth and Planetary Science Letters*, **228**, 109–123.

Korenaga, J., 2006. Archaean geodynamics and the thermal evolution of Earth. In: Benn, K., Mareschal, J.-C., and Condie, K. C. (eds.), *Archaean Geodynamic Processes*. American Geophysical Union Geophysical Monograph. 164. pp. 7–32.

Kroner, A., 1981. Precambrian plate tectonics. In Kroner, A. (ed.), *Precambrian plate tectonics*. Amsterdam: Elsevier, pp. 57–90.

Liou, J. G., Tsujimori, T., Zhang, R. Y., Katayama, I., and Maruyama, S., 2004. Global UHP metamorphism and continental subduction/collision: the Himalayan model. *International Geology Review*, **46**, 1–27.

Lowman, J. P., and Jarvis, G. T., 1996. Continental collisions in wide aspect ratio and high Rayleigh number two-dimensional mantle convection models. *Journal of Geophysical Research*, **101**, 25485–25497.

Maruyama, S., and Liou, J. G., 1998. Ultrahigh-pressure metamorphism and its significance on the Proterozoic-Phanerozoic boundary. *The Island Arc*, **7**, 6–35.

Maruyama, S., Liou, J. G., and Terabayashi, M., 1996. Blueschists and eclogites of the world and their exhumation. *International Geology Review*, **38**, 485–594.

McKenzie, D., and Bickle, M. J., 1988. The volume and composition of melt generated by extension of the lithosphere. *Journal of Petrology*, **29**, 625–679.

Menneken, M., Nemchin, A. A., Geisler, T., Pidgeon, R. T., and Wilde, S. A., 2007. Hadean diamonds in zircon from Jack Hills, Western Australia. *Nature*, **448**, 917–921.

Miyashiro, A., 1972. Metamorphism and related magmatism in plate tectonics. *American Journal of Science*, **272**, 629–656.

Moyen, J. F., Stevens, G., and Kisters, A., 2006. Record of mid-Archean subduction from metamorphism in the Barberton terrain, South Africa. *Nature*, **442**, 559–562.

Nisbet, E. G., Cheadle, M. J., Arndt, N. T., and Bickle, M. J., 1993. Constraining the potential temperature of the Archean mantle: a review of the evidence from komatiites. *Lithos*, **30**, 291–307.

O'Neill, C., Lenardic, A., Moresi, L., Torsvik, T. H., and Lee, C. T. A., 2007. *Episodic Precambrian Subduction: EOS Transactions*. American Geophysical Union, **88**(23), Abstract U44A-04.

Parman, S. W., Grove, T. L., and Dann, J. C., 2001. The production of Barberton komatiites in an Archean subduction zone. *Geophysical Research Letters*, **28**, 2513–2516.

Peacock, S. M., 2003. Thermal structure and metamorphic evolution of subducting slabs. In Eiler, J., and Hirschman, M. (eds.), *Subduction Factory*. American Geophysical Union Monograph, Vol. 138, pp. 7–22.

Peltonen, P., Kontinen, A., and Huhma, H., 1996. Petrology and geochemistry of metabasalts form the 1.95 Ga Jormua ophiolite, northeastern Finland. *Journal of Petrology*, **37**, 1359–1383.

Pesonen, L. J., Elming, S.-A., Metanen, S., Pisarevsky, S. A., D'Agrella-Filhio, M. S., Meert, J., Schmidt, P. W., Abrahmsen, N., and Bulund, G., 2003. Paleomagnetic configuration of continents during the Proterozoic. *Tectonophysics*, **375**, 289–324.

Rollinson, H., 2007. *Early Earth Systems: A Geochemical Approach*. Malden, MA: Blackwell, p. 285.

Scott, D. J., Helmstaedt, H., and Bickle, M. J., 1992. Purtuniq Ophiolite, Cape Smith Belt, northern Quebec, Canada: a reconstructed section of early Proterozoic oceanic crust. *Geology*, **20**, 173–176.

Shirey, S. B., Richardson, S. H., and Harris, J. W., 2004. Integrated models of diamond formation and craton evolution. *Lithos*, **77**, 923–944.

Shirey, S. B., Kamber, B. S., Whitehouse, M. J., Mueller, P. A., and Basu, A. R., 2008. A review of the isotopic and trace element evidence for mantle and crustal processes in the Hadean and Archaean: implications for the onset of plate tectonic subduction. In Condie, K. C., and Pease, V. (eds.), *When Did Plate Tectonics Begin on Planet Earth?* Geological Society of America Special Paper 440: Boulder, Colorado, USA, pp. 1–29.

Silver, P. G., and Behn, M. D., 2008. Intermittent plate tectonics. *Science*, **319**, 85–88.

Sleep, N. H., 1992. Archean plate tectonics: what can be learned from continental geology? *Canadian Journal of Earth Sciences*, **29**, 2066–2071.

Sleep, N. H., 2000. Evolution of the mode of convection within terrestrial planets. *Journal of Geophysical Research*, **105**, 17563–17578.

Sleep, N. H., 2005. Evolution of the continental lithosphere. *Annual Review of Earth and Planetary Sciences*, **33**, 369–393.

Sleep, N. H., 2007. Weak thermal convection within tilted plume conduits. *Geochemistry, Geophysics, Geosystems*, **8**, 11.

Sleep, N. H., and Windley, B. F., 1982. Archean plate tectonics: constraints and inferences. *Journal of Geology*, **90**, 363–379.

Smithies, R. H., Champion, D. C., and Cassidy, K. F., 2003. Formation of earth's early archean continental crust. *Precambrian Research*, **127**, 89–101.

Smithies, R. H., Champion, D. C., Van Kranendonk, M. J., Howard, H. M., and Hickman, A. H., 2005. Modern-style subduction processes in the Meso-archean: geochemical evidence from the 3.12 Ga Whundo intra-oceanic arc. *Earth and Planetary Science Letters*, **231**, 221–237.

Stern, R. J., 2002. Subduction zones. *Reviews of Geophysics*, **40**, 1012.

Stern, R. J., 2005. Evidence from ophiolites, blueschists, and ultrahigh-pressure metamorphic terranes that the modern episode of subduction tectonics began in Neoproterozoic time. *Geology*, **33**, 557–560.

Stern, R. J., 2007. When did plate tectonics begin on Earth? Theoretical and empirical constraints. *Chinese Science Bulletin*, **52**(5), 578–591.

Stern, R. J., 2008. Modern-style plate tectonics began in Neoproterozoic time: an alternate interpretation of Earth dynamics and history. In Condie, K. C., and Pease, V. (eds.), *When Did Plate Tectonics Begin on Planet Earth?* Geological Society of America Special Paper 440, pp. 265–280.

Stevenson, D. J., 2003. Styles of mantle convection and their influence on planetary evolution. *Comptes Rendus Geosciences*, **335**, 99–111.

Tomlinson, K. Y., and Condie, K. C., 2001. Archean mantle plumes: evidence from greenstone belt geochemistry, In: Ernst, R. E., and Buchan, K. L. (eds.), *Mantle Plumes: Their Identification through Time*. Geological Society of America Memoir 352, pp. 341–357.

Valley, J., Cavosie, A. J., Fu, B., Peck, W. H., and Wilde, S. A., 2006. Comment on "Heterogeneous Hadean Hafnium: Evidence

of Continental Crust at 4.4 to 4.5 Ga". *Science*, **312**, 1139a, doi:10.1126/science.1125301.
Van Keken, P. E., Kiefer, B., and Peacock, S. M., 2002. High-resolution models of subduction zones: implications for mineral dehydration reactions and the transport of water into the deep mantle. *Geochemistry, Geophysics, Geosystems*, **3**(10), 1056.
Vander Hilst, R. D., Widyantoro, S., and Engdahl, E. R., 1997. Evidence for deep mantle circulation from global tomography. *Nature*, **228**, 657–659.
Watson, E. B., and Harrison, T. M., 2005. Zircon thermometer reveals minimum melting conditions on earliest Earth. *Science*, **208**, 841–844.
Wetherill, G. W., 1976. The role of large bodies in the formation of the Earth and Moon. *Proceedings of the 7th Lunar Science Conference*, Houston, TX, pp. 3245–3257.
Wilde, S. A., Valley, J. W., Peck, W. H., and Graham, C. M., 2001. Evidence from detrital zircons for the existence of continental crust and oceans on the Earth 4.4 Ga ago. *Nature*, **409**, 175–178.
Williams, I. S., 2007. Old diamonds and the upper crust. *Nature*, **448**, 880–881.

Cross-references

Plate Driving Forces
Plate Motions in Time: Inferences on Driving and Resisting Forces

PLATES AND PALEORECONSTRUCTIONS

Alan G. Smith
Department of Earth Sciences, University of Cambridge, Cambridge, UK

Definition

The *lithosphere* is the outer rigid part of the Earth, forming a shell whose thickness may range up to about 200 km.
A *tectonic plate* is a part of the Earth's lithosphere that is bounded by active plate margins.
A present-day *plate margin* is a seismically active zone that cuts the lithosphere.
Conservative plate boundaries are plate boundaries that separate two plates that are sliding past one another along a *transform fault* (which must cut the lithosphere).
Divergent (*extensional*) *plate boundaries* mark where the two plates are separating from one another, generally forming *passive* (*or Atlantic*) *continental margins* (though extension can also occur behind island arcs).
Convergent plate boundaries are located between a converging oceanic and continental plate forming an *active* (*Pacific*) *continental margin*, or between a converging oceanic plate and an island arc.
A *global paleoreconstruction* shows a reassembly of the major continents (and oceans) relative to one another at some time in the past. If the reassembly is made using ocean-floor spreading magnetic anomalies and fracture zones, and derives the geographic poles from paleomagnetism, it is a *global paleomap*. If only the continents are depicted, it is a *global paleocontinental reconstruction* (*or map*). If each continent is projected separately onto a global map frame (as in pre-Mesozoic maps) is it is a *global composite reconstruction.*
A *paleo-plate reconstruction* is simply a paleoreconstruction that shows the plate boundaries of the time concerned, that is, ridges, trenches, and major transform faults.

Earthquakes and plates

Earthquakes are caused by the catastrophic release of accumulated elastic strain. The strain accumulates as a result of differential movement between parts of the lithosphere. A map of earthquakes over the past 40 years or so shows that differential movement is generally confined to narrow seismic zones, particularly in the oceans (Figure 1). Elsewhere there is little or no seismicity, showing that most of the lithosphere is not undergoing differential motion and is therefore rigid. The present-day surface extent of a plate is marked by zones of active seismicity that show the plate boundaries.

Plate compositions, sizes, and shapes

Plates are made up of *Lithosphere, Continental*, with a thickness that probably averages about 200 km (Priestley and McKenzie, 2006), and *Lithosphere, Oceanic*, whose thickness varies from a small value at a mid-ocean ridge, where it is created, to about 100 km for old oceanic lithosphere (Stein and Stein, 1992; Crosby et al., 2006). Most of the plates in the Pacific Ocean are almost entirely, or entirely, made of oceanic lithosphere, but other plates are made up of varying proportions of continental and oceanic lithosphere (Figure 1). There are no wholly continental plates. The shapes of most plates are quite irregular, reflecting their past histories of growth and destruction.

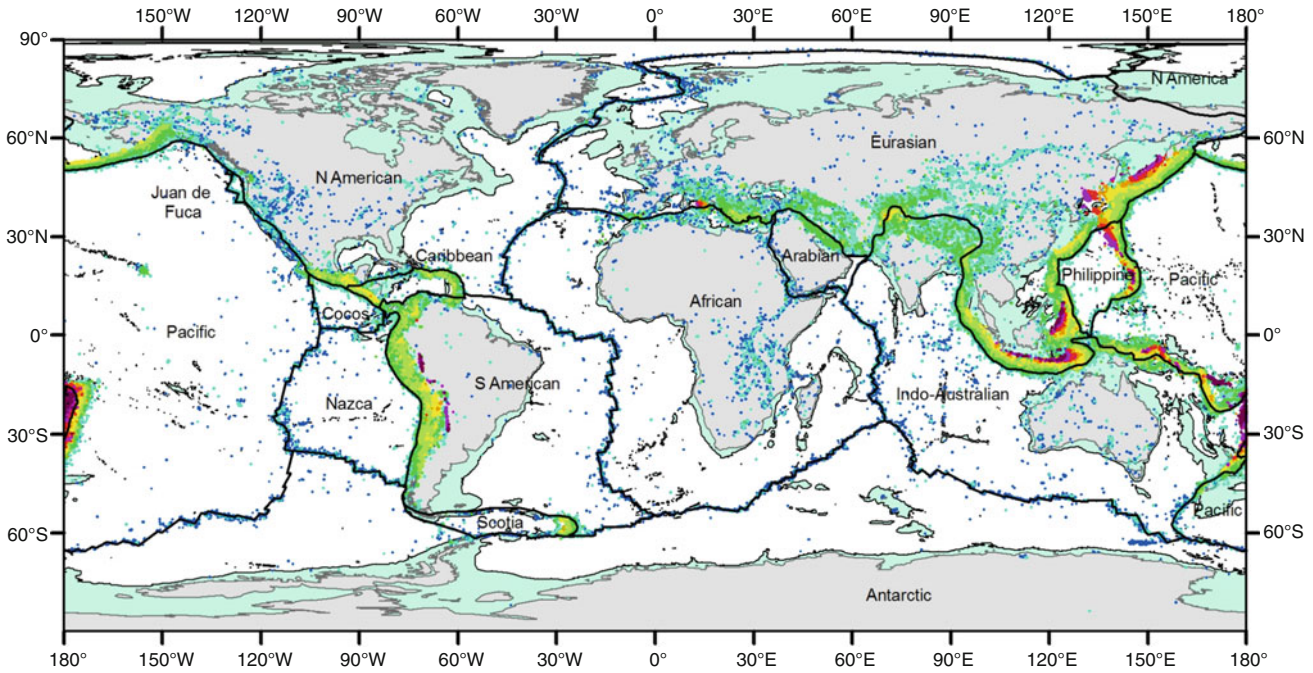
Plate tectonics

Plate tectonics describe the motions of the lithospheric plates and the effects they cause. It is now the major branch of terrestrial tectonics, which is the study of the origins of the large-scale structures of the Earth, particularly the lithosphere. The name tectonics is derived from the Greek word tektonikos (τεκτονικός) "pertaining to building."

Plate motions and Euler's theorem

Plates are rigid bodies that can move relative to one another in three ways: one, they can slide past one another; two, they can move away from one another; three, they can move toward one another.

According to a theorem attributed to Euler, the relative motion of two rigid bodies on a sphere can be described as a rotation about a line passing through the Earth's center that cuts the Earth's surface at a point known as the rotation pole. The pole will have a latitude, λ, and a longitude, φ, with the rotation that takes place in



Plates and Paleoreconstructions, Figure 1 Present-day earthquakes and simplified map of the present-day plates. The Philippine plate is also known as the Philippine Sea plate. The plate boundary between Eurasia and North America in northeast Asia is poorly defined. Cylindrical equidistant world map showing land areas as *light gray*. Areas between 2,000 m water depth and the coastline are *blue* (colored version) or *light gray* (monochrome version). Deeper oceanic areas are uncolored. Over 400,000 earthquakes are shown. In the colored map, the color code is: 0–10 km (79478) *blue*; 10–35 km (168263) *light blue*; 35–100 km (79890) *green*; 100–200 km (47056) *yellow-green*; 200–300 km (11144) *yellow*; 300–400 km (3874) *orange-brown*; 400–500 km (4674) *red*; 500–600 km (8427) *magenta*; 600–700 km (2477) *deep magenta*; >700 km (4) *black*. The monochrome version shows earthquakes ranging from *light gray* (shallow) to *black* (deepest). The rainbow-colored areas in the Australasia, together with smaller areas along the western coast of the Americas, indicate zones of seismic activity of increasing depth. The continuous *black* line is the plate network showing 14 plates, all labeled. It is unclear whether the Indo–Australian plate should be considered to be two plates with a boundary in the NW Indian Ocean. The plate margins in the oceans are very narrow and conform to ideal plate margins; those in the continents are much broader. In particular, the wide belt of active deformation in the Alpine–Himalayan belt shows that the deformation is not restricted to the narrow plate margin on the map between the African, Arabian, and Indo–Australian plates on the one hand, and the southern Eurasian plate on the other. Rigid plate tectonic descriptions of such areas may be inappropriate. (Copyright A. G. Smith). The earthquakes are those of magnitude 4 or greater listed in http://earthquake.usgs.gov/earthquakes/eqarchives/epic/

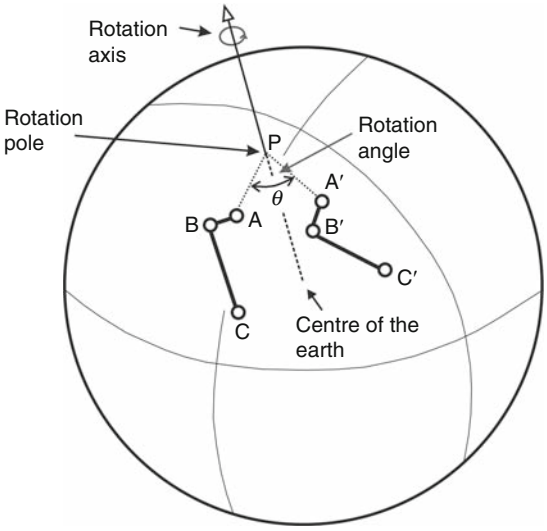
a given time given by θ (Figure 2). To a good approximation the Earth is a sphere and tectonic plates are rigid. Thus the motion of one plate relative to another in a given time interval is given by their Euler rotation (λ, φ, θ). λ varies from 90° N, the N pole, conventionally given the value of +90, to 90° S, the S pole, or −90. Longitudes are positive measured E from the Greenwich meridian, conventionally taken as 0°, or as 360°. If measured W from the Greenwich meridian, the longitudes are negative. Maps showing the whole world run from 0° E to 360° E, or 0 to −360. They could also run from 180° W, that is, −180, to 180° E, or +180. The sign of the rotation θ is found by fixing one plate as a reference and determining whether the rotation moves the second plate in an anticlockwise direction when looking down on it – a positive rotation – or moves clockwise – a negative rotation. Changing the reference plate changes the sign of the rotation.

Sliding motions

When plates slide past one another they neither create nor destroy lithosphere. They conserve the lithosphere and are described as conservative plate boundaries. On a sphere such boundaries are circular arcs forming what are known as transform faults as in the San Andreas fault. The Euler pole is the point at the center of the arc.

Divergent motions

When a continent breaks up and the pieces move away from one another the motion is divergent. At first the margins of the continents-to-be are stretched, forming what will become passive continental margins, as in the East African Rift. Eventually the stretching reaches the stage where new ocean-floor has to form, as in the Red Sea or the Gulf of Aden. The Euler rotations for each stage are given by the rotations needed to bring the different



Plates and Paleoreconstructions, Figure 2 The line ABC on a rigid body on the surface of a sphere (e.g., Earth) is moved to A′B′C′ by a rotation through the angle θ, about the rotation axis. The rotation axis passes through the center of the Earth and cuts the surface at point P with geographic coordinates of latitude λ and longitude φ (not shown), leading to an Euler rotation of (λ, φ, θ). (Copyright A. G. Smith).

markers (passive margins; anomalies of the same age) together.

Convergent motions

The local motions between present-day converging plates can be found from the slip vectors of earthquakes. If the convergent margin is long enough it is also possible to use these motions to estimate the position of the Euler pole. However, the local motions and Euler poles for ancient plate margins are difficult to estimate. In favorable circumstances it is possible to calculate them from a plate circuit. For example, the Cenozoic motion along a convergent plate boundary between India and the rest of Eurasia can be calculated by successively summing the motions from Eurasia to North America, North America to Africa, and Africa to the Indo–Australian plate (Figure 1). All of these motions are known from the ocean-floor spreading patterns in the Atlantic and Indian Oceans. Their sum gives the required result. The absence of any Paleozoic and Precambrian ocean-floor means that the motions and Euler poles at all pre-Mesozoic convergent plate boundaries are highly uncertain.

Reference frames

Three classes of reference frames are relevant to global reconstructions: any plate, the paleomagnetic reference frame, and that given by "hot-spots."

Plates as reference frames

Any plate, large or small, that can be linked to all the other plates can be used as reference frame to make a global reassembly: there is nothing fundamental in the choice of a plate. These global reassemblies are not geographic maps because they do not show the geographic latitude and longitude of the time concerned. To turn a reassembly into a map one needs to know the position of the geographic pole at the time concerned, which is given by paleomagnetism.

Paleomagnetism as a reference frame

Over a period of several tens of thousands of years or more, the Earth's magnetic field averages to what is known as a geocentric axisymmetric dipole field, or GADF. Virtually all paleomagnetic measurements use this field model to interpret the results. The field is geocentric because it behaves as if it were at the Earth's center; it is axisymmetric because the field is symmetric about the Earth's spin (or rotation) axis; and it behaves like a bar magnet and is therefore a dipole. Such a field has a very simple relationship between the angle, I, known as the inclination that the lines of magnetic force make with the horizontal, and the latitude, λ, of the point of observation:

$$\tan I = 2 \tan \lambda, \qquad (1)$$

where tan = tangent of the angle (I or λ).

As a plate moves, the latitude of a point on it changes and the orientation of the plate varies. Both these effects are recorded in the remanent magnetization – or *Paleomagnetism, Principles* – preserved in rocks. The actual position of the north magnetic pole at the time must lie along the line through the observation point that coincides with the direction of the declination. It is at a distance of ($90 - \lambda$) degrees from the observation point along that line. Thus, in principle, paleomagnetic data can be used to reposition that plate at the time that the old magnetism was superimposed on the rocks.

The determination of the age and direction of magnetization in rocks becomes progressively more difficult as one goes back in time, but these problems will not be discussed here. However, any Mesozoic or Cenozoic global reassembly can be turned into a map by finding where the mean north magnetic pole lies on the reassembly and making that point the north geographic pole of the map (Smith et al., 1994). A map of Pangea just before break-up is shown in Figure 3.

Although paleomagnetic measurements give the past latitude and orientation of a plate, they do not give its past longitude. Thus it is generally not possible to make a paleoreconstruction from paleomagnetic measurements alone. However, paleo-longitude differences between any two points on different continents will be correctly given if the relative position of the two points has been determined via a plate tectonic circuit whose individual motions are known.

Hot-spots as a reference frame

The origin of linear volcanic chains like the Hawaiian Islands has been attributed to the slow movement of

Plates and Paleoreconstructions, Figure 3 Mollweide projection of Pangea reassembled at the end of the Triassic period, ~200 Ma. The *green* areas are the present-day coastlines; the *blue* areas all lie above 2,000 m water depth, showing the approximate present-day extent of the continental crust. The latitude-longitude grid is drawn at 30° and shows in particular the estimated position of the late Triassic equator. Pangea was assembled by the collision of the southern continents (Gondwana) with the northern continents (Laurasia). A large oceanic area, known as the Tethys, separated southern Eurasia from northeast Gondwana. The distribution of continental fragments lying within the Tethys is uncertain, particularly that of China and adjacent areas (the large fragment lying east of Eurasia). Some continental slivers and/or island arcs off western North America are shown schematically. Several other slivers probably lay adjacent to other continental margins of Pangea, but no attempt has been made to depict them. (Copyright A. G. Smith).

a tectonic plate over a fixed deep plume of hot mantle, giving rise to "hot-spots" at the surface (Wilson, 1963; Morgan, 1981). A purely descriptive term for them is that they are examples of "large igneous provinces" created in the case of hot-spots by intense, generally basaltic, magnetism. From the point of view of making reconstructions they show an interesting property in that they move relative to one another at rates that are an order of magnitude slower than those of plates. Hot-spots have been used as a reference frame that is believed to provide longitudes as well as latitudes for a global reassembly, forming a so-called absolute reference frame (Duncan, 1981; Müller et al., 1993). The paleomagnetic reference and hot-spots reference frames agree within the limits of error for much of Cenozoic time, then gradually diverge for reasons that are not clear (Livermore et al., 1988).

Paleoreconstructions

Paleoreconstructions attempt to show where the oceans and continents were in the geological past. That this is possible is due to several effects: (1) conservation of continental crust during plate motions; (2) the record of Mesozoic and Cenozoic continental separation in the intervening ocean-floor; and (3) the record of changes in continental latitude and orientation in *Paleomagnetism, Principles*.

Conservation of continental shapes

The extension on passive continental margins is probably about 100 km for a margin that is 200 km wide at the present-day, such as parts of the eastern North American margin (e.g., Sawyer, 1985). In other areas, the extension may be much higher: Powell et al. (1988) estimated a total of 360 km extension across the Australia–Antarctica margin before the formation of ocean-floor (180 km) equivalent to 1.6° of latitude for each margin, assuming symmetric extension.

To a good approximation, continental shapes on a global scale are conserved during plate motions. In detail this is clearly not the case. For example, when continents separate the passive margins have been stretched. When two continents converge, the continental lithosphere of one continent overrides that of the other, which is then partially subducted. But because continental lithosphere is less dense than the underlying mantle, it cannot be subducted to any great depth. Thus the continental crust in the collisional zone becomes thickened and the area of continental crust decreases. In the case of the collision between India and Asia, the Tibetan plateau, and the Himalayas may represent a loss of continental area in a strip that may be more than 500 km wide (Coward et al., 1988; Shackleton, 1988). As one goes back in time, the subducted Indian crust should be pulled out from under Eurasia and added to the Indian continent. Similar adjustments, but on a generally smaller scale, are necessary to the shapes of all continents caught up in collision zones, but are rarely made.

The continental shapes used for reconstructions generally show the edges of continents at about 2,000 m water depth. This shape is unfamiliar where there is a wide shallow sea between the 2 km submarine bathymetry contour and the coastline. To aid in recognition, the present-day coastline is commonly shown on continents, but the coastline of the time concerned may be quite different, as well as being very difficult to pin down.

Record of continental separation in the ocean-floor

The ocean-floor generated during the break-up of Pangea and the creation *of* the Atlantic, India, Antarctic, and

Arctic Oceans is still mostly preserved. Successively older oceanic transform faults, fracture zones, and ocean-floor magnetic anomalies and the continental edges themselves can be fitted together to give the Euler rotations needed to reposition all the major continents to a high degree of precision for most of Mesozoic and all of Cenozoic time, or roughly back to 200 Ma. The net results are paleocontinental reconstructions, that is, a display of the relative positions of the continents in the past (Smith et al., 1973).

Paleozoic and Precambrian composites

For pre-Mesozoic time the only ocean-floor that remains is tectonically highly deformed. Thus Paleozoic and Precambrian global "maps" are made by geographically repositioning individual continents using paleomagnetism. The uncertainties in paleomagnetic poles are considerable, particularly for Precambrian time. Although the order in which the continents are arranged on the globe is known, at least for Paleozoic time, the longitude separations of adjacent continents are not. These separations are subjectively estimated from evidence such as the distribution of fossils, tectonic effects, and the like. The results have been called composites, rather than maps, to emphasize the fact that they include qualitative data (Smith et al., 1973). The Precambrian has few fossils that provide an accurate indication of age and only scattered age determinations. Their absence, together with the absence of ocean-floor data and the uncertainties in pole positions, means that there are no generally agreed composites for most of Precambrian time. The Precambrian is still a veritable terra incognita.

Paleo-plate reconstructions

As noted above, a *global paleoreconstruction* is a reassembly of the major continents (and oceans) relative to one another at some time in the past. To turn this into a *paleo-plate reconstruction*, one needs to add the plate margins of the time concerned. For Jurassic and younger periods the former positions of ocean ridges are known from the ocean-floor record in the Atlantic, Indian, and Southern Oceans, but much of the Pacific ocean-floor has been subducted and for these areas former ridge positions have to be estimated. For Triassic and older periods the positions of former ocean ridges is largely a matter of informed guesswork (e.g., Stampfli and Borel, 2002, 2004).

Orogenic belts, that is, those areas made up of deformed, metamorphosed, and igneous rocks, mark the former positions of convergent margins. However, the margins are commonly difficult to locate precisely and many orogenic belts include more than one former convergent margin, together with the locations of former oceans that have since been subducted. Thus although one can join former ridges to orogenic belts to make a sketch plate boundary map for the time concerned, this map is necessarily imprecise.

Precambrian *global paleoreconstructions* are generally schematic (e.g., Collins and Pisarevsky, 2005). Because paleo-plate reconstructions build on an initial global reassembly, most Paleozoic and all Precambrian paleo-plate reconstructions are essentially cartoons that show the kinds of plate boundaries that may have existed at a given time but whose locations may have considerable errors.

Conclusions

The ocean-floor magnetic anomalies, transform faults, and fracture zones, together with the paleomagnetic frame allow very precise global Cenozoic and Mesozoic maps to be made. The inferred relative longitudes of Paleozoic and Precambrian continents depend on qualitative data, giving rise to poorly defined "composites." Fossils enable precise dating of Paleozoic features that help in the construction of composites, but such precision is not obtainable from Precambrian fossils, giving rise to considerable uncertainties in all Precambrian reconstructions. How to convert these global paleoreconstructions into verifiable paleo-plate reconstructions is a largely unsolved problem for pre-Mesozoic time.

Bibliography

Collins, A. S., and Pisarevsky, S. A., 2005. Amalgamating eastern Gondwana: the evolution of the Circum-Indian Orogens. *Earth Science Reviews*, **71**, 229–270.

Cox, A., and Hart, R. B., 1986. *Plate Tectonics: How It Works*. Oxford: Blackwell.

Crosby, A. G., McKenzie, D., and Sclater, J. G., 2006. The relationship between depth, age and gravity in the oceans. *Geophysical Journal International*, **166**, 553–573.

Duncan, R. A., 1981. Hotspots in the Southern Ocean – An absolute frame of reference for motions of the Gondwana continents. *Tectonophysics*, **74**, 29–42.

Isacks, B., Oliver, J., and Sykes, L. R., 1968. Seismology and the new global tectonics. *Journal of Geophysical Research*, **73**, 5855–5899.

Livermore, R. A., Vine, F. J., and Smith, A. G., 1983. Plate motions and the geomagnetic field. 1: quaternary and late tertiary. *Geophysical Journal of the Royal Astronomical Society*, **73**, 153–171.

McKenzie, D. P., and Parker, R. L., 1967. The North Pacific: an example of tectonics on a sphere. *Nature*, **216**, 1276–1280.

Morgan, W. J., 1968. Rises, trenches, great faults and crustal blocks. *Journal of Geophysical Research*, **73**, 1959–1982.

Morgan, W. J., 1981. Hotspot tracks and the opening of the Atlantic and Indian Oceans. In Emiliani, C. (ed.), *The Oceanic Lithosphere*. New York: Wiley, pp. 443–487.

Müller, R. D., Royer, J. Y., and Lawver, L. A., 1993. Revised plate motions relative to the hotspots from combined Atlantic and Indian Ocean hotspot tracks. *Geology*, **21**, 275–278.

Powell, C. M., Roots, S. R., and Veevers, S. J., 1988. Pre-breakup continental extension in East Gondwanaland and the early opening of the eastern Indian Ocean. *Tectonophysics*, **155**, 261–283.

Priestley, K., and McKenzie, D., 2006. The thermal structure of the lithosphere from shear wave velocities. *Earth and Planetary Science Letters*, **244**, 285–301.

Sawyer, D. S., 1985. Total tectonic subsidence: a parameter for distinguishing crust type at the U.S. continental margin. *Journal of Geophysical Research*, **90**, 7751–7769.
Shackleton, R. M., 1988. Tectonic evolution of the Himalayas and Tibet. In Shackleton, R. M., Dewey, J. F., and Windley, B. F. (eds.), *Philosophical Transactions of the Royal Society of London, Series A: Mathematical and Physical Sciences*. London: The Royal Society of London.
Smith, A. G., Briden, J. C., and Drewry, G. E., 1973. Phanerozoic world maps. In Hughes, N. F. (ed.), *Organisms and Continents through Time, Volume 12: Special Paper*. London: Palaeontological Association, pp. 1–42.
Smith, A. G., Smith, D. G., and Funnell, B. M., 1994. *Atlas of Mesozoic and Cenozoic Coastlines*. Cambridge: Cambridge University Press.
Stampfli, G. M., and Borel, G. D., 2002. A plate tectonic model for the Paleozoic and Mesozoic constrained by dynamic plate boundaries and restored synthetic oceanic isochrons. *Earth and Planetary Science Letters*, **196**, 17–33.
Stampfli, G. M., and Borel, G. D., 2004. The TRANSMED transects in space and time: constraints on the Paleotectonic evolution of the Mediterranean domain. In Cavazza, W., Roure, F. M., Spakman, W., Stampfli, G. M., and Ziegler, P. A. (eds.), *The TRANSMED Atlas – The Mediterranean Region from Crust to Mantle*. Berlin: Springer, pp. 53–80.
Vine, F. J., and Matthews, D. H., 1963. Magnetic anomalies over oceanic ridges. *Nature*, **199**, 947–949.
Wilson, J. T., 1963. A possible origin of the Hawaiian Islands. *Canadian Journal of Physics*, **41**, 863–870.

Cross-references

Lithosphere, Continental
Lithosphere, Oceanic
Magnetic Anomalies, Interpretation
Magnetic Methods, Surface
Paleomagnetism, Principles
Subduction Zones

POROELASTICITY

Ran Bachrach
Geophysics and Planetary Sciences Department, Tel Aviv University, Tel Aviv, Israel

Definition

Poroelasticity is a joint formulation for the behavior of a solid–fluid coupled porous system. Poroelasticity describes the behavior of porous continuum and pore fluid with respect to applied displacements and stresses. Poroelasticity jointly treats the solid frame and pore fluid to calculate changes in pore pressure, solid displacements, and fluid displacements due to stresses and displacements associated with external or internal processes.

The theory of poroelasticity is based on the work of Terzaghi, 1941; Biot, 1941, 1956a, b, 1962; and Gassmann, 1951. Poroelasticity is often called Biot theory as the work of M. A. Biot is considered to establish the complete foundation for the theory.

Basic concepts

Consider a connected network of pores fully saturated with fluid. The following are basic concepts that form the basis for the theory of poroelasticity.

Drained and undrained deformation

For a fully saturated porous material, we consider two fundamental states (or thought experiments): *Drained* and *undrained* (sometimes these are called *jacketed* and *unjacketed* conditions).

Drained conditions are achieved when pore pressure in a representative elementary volume (REV) is in equilibrium with the surrounding throughout the deformation of the saturated porous medium. A special case (described as a lab experiment) is when pore pressure and fluid can freely defuse in and out of the sample during the deformation process. Undrained (or jacketed) conditions refer to the condition in which the boundary of the porous solid (REV) is impermeable and pore fluid cannot exit the rock during the deformation process.

Physical measurements on both states (laboratory measurements) will determine the poroelastic coefficients of the rock. Often "dry conditions" are associated with the drained experiment. In poroelasticity the term "dry" elastic modulus refers to the drained experiment and not the elastic modulus when the material is dry.

Linear stress–strain formulation for poroelastic media

Consider the displacement vector in the solid $\mathbf{u} = [u_x, u_y, u_z,]$ and the average fluid displacement vector in the pore space $\mathbf{U} = [U_x, U_y, U_z,]$.

The volume of fluid displaced through a unit area normal to the x, y, z, direction is $\phi\mathbf{U}$ and the fluid displacement field is $\mathbf{w}$. The isotropic poroelastic stress–strain relations that relate stress, σ, and pore pressure, P, to displacement and strain, e, can be written in an abbreviated notation as:

$$\sigma_{ij} = 2\mu e_{ij} + \delta_{ij}(\lambda_c\theta - \alpha M\zeta)$$
$$P = -\alpha M\theta + M\zeta$$
$$\theta = \nabla\cdot\mathbf{u}, \quad \zeta = \varphi\nabla\cdot(\mathbf{u}-\mathbf{U})$$

where λ_c is the lame's coefficient for the closed (undrained) system and μ is the shear moduli. Note that the shear strain and stress are not affected by pore pressure and fluid presence. M, α are the two poroelastic constants associated with the poroelastic stress–strain formulation. M is known as Biot modulus, or P wave modulus, and relates dry response to saturated response as follows:

$$\lambda_C = \lambda_{\text{dry}} + \alpha^2 M,$$
$$\frac{1}{M} = \frac{\alpha-\varphi}{K_0} + \frac{\varphi}{K_f}.$$

where K_f is the pore fluid bulk modulus and φ is the porosity. α is the Biot parameter and is defined as:

$$\alpha = 1 - K_{\text{dry}}/K_0$$

where K_{dry} is the bulk modulus of the rock measured under drained conditions and K_0 is the bulk modulus of the frame mineral.

Note that when the system is jacketed, that is, $\zeta = 0$ so there is no relative displacement between the fluid and matrix, the pore pressure and the total stress are related and the bulk modulus is defined in terms of "closed" elastic properties.

When the matrix is open, the assumption is that there is a constant pore pressure in the system (pore pressure is in equilibrium with the surroundings) and thus fluid can flow freely in and out of the system. One special case of such conditions is when the pore pressure is equal to zero. The term "dry" frame conditions often refers to the states of zero pore pressure.

The poroelastic linear stress–strain relation can be written in terms of pore pressure as the coupling variable rather than the relative fluid displacement (or divergence). A popular form of the linear isotropic poroelastic stress–strain relations is given by:

$$\sigma_{ij} + \alpha P\delta_{ij} = 2\mu e_{ij} + \lambda_{dry}\theta\delta_{ij}$$
$$\zeta = (1/M)P + \alpha\theta$$

Note that the pressure direction here is assumed to be negative, for example, the pore pressure resists a positive confining stress that pushes the grains together.

The above equation can be written as an equivalent linear elasticity relation by substituting the total stress with Biot effective stress $\sigma^*_{ij} = \sigma_{ij} + \alpha P\delta_{ij}$ as:

$$\sigma_{ij}* = 2\mu e_{ij} + \lambda_{dry}\theta\delta_{ij}$$
$$\zeta = (1/M)P + \alpha\theta$$

In terms of the differential pressure (also known as Terzaghi's effective stress or just the "effective" stress) defined as $\sigma'_{ij} = \sigma_{ij} + P\delta_{ij}$ the linear stress–strain relations are given by:

$$\sigma_{ij}' - (1-\alpha)P\delta_{ij} = 2\mu e_{ij} + \lambda_{dry}\theta\delta_{ij}$$
$$\zeta = (1/M)P + \alpha\theta$$

When $\alpha = 1$ (soft materials where $K_{dry} << K_0$) the Biot effective stress is equal to the differential stress (often called Tarahzagi's effective stress). Simple derivation of Biot's effective stress law is given by Nur and Byerlee, 1971.

Gassmann's equation and fluid substitution

Gassman's equation (Gassmann, 1951) relates the elastic constant of saturated material in closed or undrained state to the drained or open conditions. In its isotropic form, Gassmann's equation can be written as:

$$\frac{K_{sat}}{K_0 - K_{sat}} = \frac{K_{dry}}{K_0 - K_{dry}} + \frac{K_{fl}}{\varphi(K_0 - K_{fl})}$$
$$\mu_{sat} = \mu_{dry}$$

where K_{sat} refers to the saturated bulk modulus of the closed system. The shear modulus for the open system (dry) and the closed system are the same $\mu_{sat} = \mu_{dry}$ as fluid does not contribute to the shear stiffness of the system. Gassman's equation can be derived directly from Biot's equations or from basic elasticity considerations by accounting for mass conservation of pore fluid in a closed system. Gassmann's equation assumes that pore fluid is in equilibrium throughout the sample and that the pore fluid does not interact with the mineral frame. For more information see Mavko et al. (1998). Anisotropic extension to Gassmann's equation is derived by Brown and Korringa (1975) in terms of the elastic compliance tensor:

$$S^{dry}_{ijkl} - S^{sat}_{ijkl} = \frac{\left(S^{dry}_{ij\alpha\alpha} - S^{0}_{ij\alpha\alpha}\right)\left(S^{dry}_{kl\alpha\alpha} - S^{0}_{kl\alpha\alpha}\right)}{\left(S^{dry}_{\alpha\alpha\beta\beta} - S^{0}_{\alpha\alpha\beta\beta}\right) + (\beta_{fl} - \beta_0)}$$

where $S^{dry}_{ijkl}, S^{sat}_{ijkl}$ are the saturated and dry 4th order compliance tensors, β_{fl} is the fluid compressibility (or inverse bulk modulus), and β_0 is the compressibility of the mineral frame.

Fluid substitution

One of the many practical applications of Gassmann's equation is the ability to predict the elastic modulus of a known rock saturated with different fluids. One example is the prediction of seismic velocity of a gas-saturated sediment given the seismic velocity of the same sediment saturated with brine. This fluid substitution scheme is readily derived from Gassmann's equation:

$$\frac{K_{sat1}}{K_0 - K_{sat1}} - \frac{K_{fl1}}{\varphi(K_0 - K_{fl1})} = \frac{K_{sat2}}{K_0 - K_{sat2}} - \frac{K_{fl2}}{\varphi(K_0 - K_{fl2})}$$
$$\mu_{sat2} = \mu_{sat1}$$

Skempton coefficient

If the system is closed (jacketed), the pore pressure and the external stress are directly related as:

$$\theta = \frac{(\bar{\sigma} - \alpha P)}{K_{dry}} \rightarrow P = \frac{1}{\alpha}\left(\bar{\sigma} - K_{dry}\theta\right), \quad \bar{\sigma} = \sigma_{kk}/3$$

$$\theta = e_{kk} = \frac{\bar{\sigma}}{K_{sat}} \rightarrow P = \frac{1}{\alpha}\left(K_{sat} - K_{dry}\right)\theta \rightarrow P = B\bar{\sigma},$$
$$B = \frac{1}{K_{sat}}\left(\frac{K_{sat} - K_{dry}}{\alpha}\right)$$

B is known as the Skempton coefficient and is one of the most measured laboratory quantity.

Three-dimensional consolidation

Poroelasticity provides the analytical framework for three-dimensional consolidation of soils and sediments as follows:

Darcy's law relates fluid motion to pore pressure gradient and gradient of the gravitational potential G and can be written as

$$\partial \mathbf{w}/\partial t = \frac{-\boldsymbol{\kappa}}{\eta}\nabla P - \frac{-\boldsymbol{\kappa}}{\eta}\nabla G$$

where $\boldsymbol{\kappa}$ is the permeability tensor and η is the fluid viscosity.

The total stress field must satisfy the equilibrium equations (neglecting the inertia terms)

$$\sum \frac{\partial \sigma_{ij}}{\partial x_j} + \rho b_i = 0,$$

where b_i is the body force acting on the media (Gravity) and will be neglected in the following expressions. Combining Darcy's law with the equilibrium equation, we obtain the general consolidation equations that couple fluid and solid displacements of the vectors $\boldsymbol{u}$ and $\boldsymbol{w}$ as:

$$2\sum_{ij}\frac{\partial}{\partial x_j}(\mu e_{ij}) + \frac{\partial}{\partial x_j}(\lambda_C\theta - \alpha M\zeta) = 0$$

$$\frac{\partial \mathbf{w}}{\partial t} = \frac{\boldsymbol{\kappa}}{\eta}\nabla(\alpha M\theta - M\zeta)$$

for the special case where the poroelastic coefficients are constant, the equation of motion is written (after application of the divergence operator and some algebra) as:

$$(2\mu + \lambda_C)\nabla^2\theta - \alpha M\nabla^2\zeta = 0$$

$$\frac{\partial \zeta}{\partial t} = \frac{\boldsymbol{\kappa}}{\eta}M_C\nabla^2\zeta$$

$$M_C = M(2\mu + \lambda)/(2\mu + \lambda_C)$$

The latter equation can be written in terms of pore pressure as the familiar diffusion equation with an added volumetric solid deformation term:

$$\frac{\partial P}{\partial t} - \frac{\boldsymbol{\kappa}}{\eta}M\nabla^2 P = -\alpha M\frac{\partial \theta}{\partial t}$$

This equation is used extensively in the study of soil consolidation, compacting reservoirs, and stress–fluid coupled processes in the earth.

Dynamic poroelasticity and wave propagation in saturated porous medium

The wave equation in saturated porous media is derived by adding the inertia terms to the equation of motion. For isotropic media with isotropic permeability, it can be written in terms of fluid and solid displacement vectors:

$$2\sum_{ij}\frac{\partial}{\partial x_j}(\mu e_{ij}) + \frac{\partial}{\partial x_i}(\lambda_C\theta - \alpha M\zeta) = \frac{\partial^2}{\partial t^2}\left(\rho u_i + \rho_f w_i\right)$$

$$\frac{\partial}{\partial x_i}(\alpha M\theta - M\zeta) = \frac{\partial^2}{\partial t^2}\left(\rho_f u_i + m w_i\right) + \frac{\eta}{k}\frac{\partial w_i}{\partial t}$$

where $m = \left(\varphi\rho_f + \rho_a\right)/\varphi^2$ is the mass coupling parameters, and $\rho_a = \varphi\rho_f(a - 1)$ is the apparent density that relates the dynamic tortuosity to fluid density and rock porosity (Berryman, 1980).

The equations of motion can be written in terms of rotation-free and dilatation-free equations similar to the equation for P and S wave in linear elasticity. However, unlike linear elasticity, in isotropic linear poroelastic media, there are two P waves that travel with two different velocities and a single shear wave velocity. The two P wave velocities are known as the fast and slow Biot wave velocities and have been observed in the laboratory (Plona, 1980). When the fluid viscosity is zero, the two P waves are not attenuated but when the fluid viscosity is finite the slow Biot wave is highly attenuated and quickly decays into pore pressure diffusion. In his 1956 papers, Biot discusses many details associated with wave propagation in the low frequency range, which is the appropriate range for seismic investigation, and the high frequency range, which covers the ultrasonic case.

Bibliography

Berryman, J. G., 1980. Conformation of Biot's theory. *Applied Physics Letters*, **37**, 382–384.

Biot, M. A., 1941. General theory of three dimensional consolidation. *Journal of Applied Physics*, **12**, 155–164.

Biot, M. A., 1956a. Theory of propagation of elastic waves in fluid saturated porous solid 1. Low frequency range. *The Journal of the Acoustical Society of America*, **28**, 168–178.

Biot, M. A., 1956b. Theory of propagation of elastic waves in fluid saturated porous solid 2. Higher frequency range. *The Journal of the Acoustical Society of America*, **28**, 179–191.

Biot, M. A., 1962. Mechanics of deformation and acoustic propagation in porous media. *Journal of Applied Physics*, **33**, 1482–1498.

Brown, R., and Korringa, J., 1975. On the dependence of elastic properties of porous rock on the compressibility of pore fluid, *Geophysics*, **40**, 608–616.

Gassmann, F., 1951. Uber die Elastizitat poroser Medien. *Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürich*, **96**, 1–23.

Mavko, G., Mukerji, T., and Dvorkin, J., 1998. *The rock physics handbook*, Cambridge university press: New York.

Nur, A., and Byerlee, J. D., 1971. An exact effective stress law for elastic deformation of rock with fluids. *Journal of Geophysical Research*, **76**, 6414–6419.

Plona, T. J., 1980. Observation of a second bulk compressional wave in a porous medium at ultrasonic frequencies. *Applied Physics Letters*, **36**, 259–261.

Terzaghi, K., 1941. *Theoretical Soil Mechanics*. New York: Wiley.

PROPAGATION OF ELASTIC WAVES: FUNDAMENTALS

Francisco J. Sánchez-Sesma[1], Ursula Iturrarán-Viveros[2]
[1]Instituto de Ingeniería, Universidad Nacional Autónoma de México, Ciudad Universitaria, México DF, Mexico
[2]Departmento de Matemáticas, Facultad de Ciencias, Universidad Nacional Autónoma de México, Ciudad Universitaria, México DF, Mexico

Definition

Dynamic elasticity allows predicting and explaining the propagation of perturbations in solids. The contributions of Newton, Hooke, Cauchy, Poisson, and Navier laid the basis of dynamic elasticity by the beginnings of the nineteenth century. It is possible to solve the equation of motion in terms of displacements using potentials which are solutions of the wave equations for longitudinal (or compression) and transverse (or shear) waves. The waves that propagate in fluids like water and air are called acoustic waves. Plane waves are the simplest solutions of wave equation as discovered by D'Alembert. In conjunction with Fourier transforms, plane waves are central in the solutions of a variety of problems like waves in layered media and surface guided waves.

Introduction

Waves are part of our existence and manifest themselves at many scales and with an enormous variety of effects. Perhaps the most familiar waves could be those generated in a pond after a pebble is thrown. Other waves that we cannot see are the sound waves that propagate in the air. They allow both hearing and speech. Elastic waves are of the same type of acoustic waves but can be also of transverse nature as they can propagate in solids. In what follows we establish the fundamental equations and delineate solution strategies. Recent findings are pointed out.

Newton's second law

During the second half of eighteenth century, Isaac Newton developed groundbreaking ideas of mechanics and optics; we can credit him the invention of calculus although this has been a matter of bitter debate with Leibnitz and his followers. Regarding the motion of bodies, his descriptions enabled the industrial revolution. Consider a particle with mass m subjected to loads. Second Newton's law establishes that:

$$\sum \mathbf{F} = m\ \mathbf{a}, \tag{1}$$

which means that the sum of forces equals the product of mass and acceleration. In this equation boldface letters represent vectors. For instance, the acceleration vector $\mathbf{a} = (a_1, a_2, a_3)^T$ has components a_j for the three directions x_j (with j = 1, 2, and 3). The upper index T means the transpose operation.

As velocity is the change of displacement with time, acceleration is the change of velocity with time. Therefore, acceleration is the second derivative of displacement **d** with respect to time and is written as:

$$\mathbf{a} = \frac{\mathrm{d}^2\mathbf{d}}{\mathrm{d}t^2}. \tag{2}$$

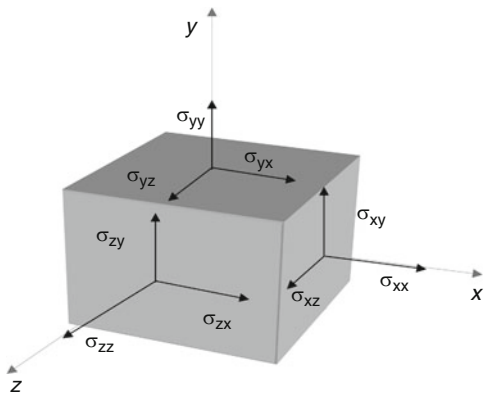
Newtonian mechanics explained the motion of both celestial bodies and everyday objects. By means of the methods of calculus, developed significantly by Newton himself, these ideas where applied to deformable solids. For a successful application a relationship between forces and displacements, a constitutive law, is needed.

Cauchy's stress and strain

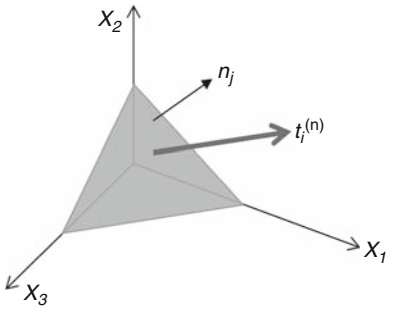
Assuming Cartesian coordinates and studying equilibrium within a continuum we can state Newton's second law per unit volume in modern form using both index notation and partial derivatives with respect to particular variables. This is possible thanks to the concept of stress tensor developed by Cauchy. Therefore we can write

$$\frac{\partial \sigma_{ji}}{\partial x_j} + f_i = \rho \frac{\partial^2 u_i}{\partial t^2}, \tag{3}$$

where $u_i(\mathbf{x},t)$ = displacement vector of point **x** (represented by x_j, with j = 1, 2, and 3, respectively, or by x, y and z) and time t, $\sigma_{ji}(\mathbf{x},t)$ = stress tensor components (force per unit area) on the faces of a particle oriented along the axis j (see Figure 1), f_i = body force pr unit volume, ρ = mass density, and the right hand term is precisely the mass per unit volume multiplied by acceleration. Moreover, it is possible to show that in order to guarantee equilibrium on moments the stress tensor must be symmetric (i.e., $\sigma_{ji} = \sigma_{ij}$). The equilibrium at a given boundary subjected to tractions or forces per unit area leads to Cauchy equation $\sigma_{ij} n_j = t_i^{(\mathbf{n})}$ which relates the stress tensor at a point with traction associated to a given normal vector as depicted in Figure 2.



Propagation of Elastic Waves: Fundamentals, Figure 1 Stress components on the faces of a particle oriented along the axes *x*, *y*, and *z*.



Propagation of Elastic Waves: Fundamentals, Figure 2 Traction vector on a plane with normal n_j.

The motion at point $\mathbf{x} + d\mathbf{x}$ allows for a first order description if it is within a small neighborhood of $\mathbf{x}$ and can be expressed in terms of values at $\mathbf{x}$ by means of:

$$\begin{aligned} u_i(\mathbf{x}+d\mathbf{x},t) &= u_i(\mathbf{x},t)+\frac{\partial u_i}{\partial x_k}dx_k \\ &= u_i(\mathbf{x},t)+\frac{1}{2}\left(\frac{\partial u_i}{\partial x_k}+\frac{\partial u_k}{\partial x_i}\right)dx_k \\ &\quad +\frac{1}{2}\left(\frac{\partial u_i}{\partial x_k}-\frac{\partial u_k}{\partial x_i}\right)dx_k. \end{aligned} \tag{4}$$

The first and third terms of last equality exhibit the motion at $\mathbf{x} + d\mathbf{x}$ as the composition of a rigid body translation and rotation, respectively, which do not produce any strains. The middle term really implies strains and is the product of the Cauchy infinitesimal strain tensor,

$$e_{ik}=\frac{1}{2}\left(\frac{\partial u_i}{\partial x_k}+\frac{\partial u_k}{\partial x_i}\right), \tag{5}$$

and the spatial differentials dx_k. This tensor is a natural candidate for a constitutive law.

Hooke's law

Solids undergo deformations when subjected to loads. If deformations are small a linear relationship can be assumed. This is what Robert Hooke did when studying springs. To secure his finding he even wrote an anagram: *ceiiinosssttuv* to hide the Latin sentence: *ut tensio sic vis* which literally means that the displacement (*vis*) is proportional to the force (*tensio*).

Nowadays, we express Hooke's law simply by writing $F = k \times \delta$ where $F =$ force, $k =$ spring constant or stiffness, and $\delta =$ displacement. Hooke and Newton were contemporary, but not very friendly to each other. Hooke claimed Newton stole from him the notion of gravitational force as proportional to inverse square of distance; Newton ignored him. Actually, such an idea was implicit in Kepler's laws and many suspected it, but only Newton gave a precise quantitative account.

For linear elastic materials the more general constitutive law is

$$\sigma_{ij}=c_{ijpq}\,e_{pq}. \tag{6}$$

This is Hooke's law expressing a linear relationship between stresses σ_{ij} and strain e_{pq} at a point. Very likely, Hooke never imagined such an extension. Technically speaking these are second order tensors. Both are due to Cauchy.

Elastic properties appear in the tensor c_{ijpq} which is of fourth order. At a given point, it represents 81 entries but after recognizing the symmetry of the stress and strain tensors we have 36 independent constants. Energy considerations allow having only 21 constants for a general anisotropic solid. If the medium is isotropic we have

$$c_{ijpq}=\lambda\delta_{ij}\delta_{pq}+\mu(\delta_{ip}\delta_{jq}+\delta_{iq}\delta_{jp}), \tag{7}$$

where $\delta_{ij} =$ Kronecker delta (=1 if $i = j$; =0 if $i \neq j$; note that $\delta_{jj} = 3$) and the parameters $\lambda, \mu =$ Lamé constants. Equation 7 shows that only two independent constants are needed to describe a linear isotropic solid.

Navier's equation

For isotropic materials Hooke's law can be written as

$$\sigma_{ij}=\lambda e_{kk}\delta_{ij}+2\mu e_{ij}. \tag{8}$$

Substituting Equation 8 in Equation 3 and considering the definition of strain in Equation 5, we can write the governing equation for motion within an elastic medium:

$$\mu\frac{\partial^2 u_i}{\partial x_j\partial x_j}+(\lambda+\mu)\frac{\partial^2 u_j}{\partial x_j\partial x_j}+f_i=\rho\frac{\partial^2 u_i}{\partial t^2}. \tag{9}$$

This is Navier's equation. It is possible to show that under very general circumstances the homogeneous solution (for $f_i = 0$) can be expressed by means of:

$$u_i=\frac{\partial\phi}{\partial x_i}+\varepsilon_{ijk}\frac{\partial\psi_k}{\partial x_j} \quad \text{with} \quad \frac{\partial\psi_k}{\partial x_k}=0. \tag{10}$$

Here $\varphi = \varphi(\mathbf{x}, t)$ and $\psi_k = \psi_k(\mathbf{x}, t)$ are functions called potentials (see Achenbach, 1973). The first is a scalar while the second is a vector. These functions must be solutions of the homogeneous wave equations:

$$\frac{\partial^2\phi}{\partial x_j\partial x_j}=\frac{1}{\alpha^2}\frac{\partial^2\phi}{\partial t^2} \quad \text{and} \quad \frac{\partial^2\psi_k}{\partial x_j\partial x_j}=\frac{1}{\beta^2}\frac{\partial^2\psi_k}{\partial t^2}, \tag{11}$$

respectively. This is supported by Lamé's theorem (for details see Aki and Richards, 1980). Here $\alpha=\sqrt{(\lambda+2\mu)/\rho}$ and $\beta=\sqrt{\mu/\rho}$ are the velocities of compressional (or P) and shear (or S) waves, respectively. The existence of two types of waves was suspected earlier thanks to experiments that show light as transversely polarized waves. Poisson employed Newtonian intermolecular elastic forces and found theoretically the existence of P and S waves. His considerations lead him to conclude that $\alpha=\sqrt{3}\beta$. This happens when $\lambda = \mu$ (or

$v = 0.25$) and therefore, a medium with this property is called a Poisson solid.

D'Alembert solution of wave equation

For the linearly elastic, isotropic, homogeneous elastic medium, Navier's equation can be solved through the superposition of solutions of wave equation called potentials (see Aki and Richards, 1980). Therefore, our capability to solve the canonical equation of dynamic elasticity rests upon the ability to solve wave equation. This is subject of significant mathematical interest and many methods exist to look for solutions. Here we follow instead a quick, intuitive approach to solve the wave equation. In our opinion, the simplest solution is the one proposed by D'Alembert, whom was a disciple of the eighteenth century French Illustration. Extending his idea we can write

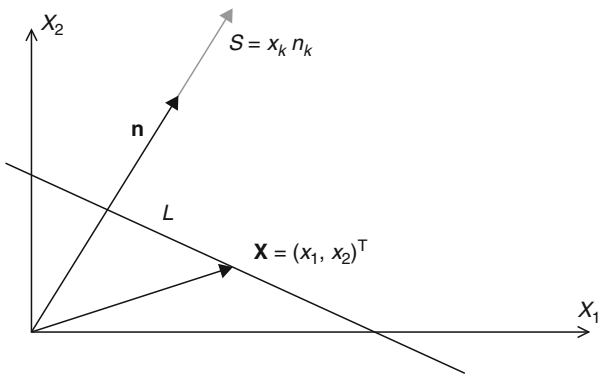
$$\phi(\mathbf{x},t) = f\left(t - \frac{x_k n_k}{\alpha}\right) + g\left(t + \frac{x_k n_k}{\alpha}\right), \tag{12}$$

which is a solution of a wave equation for almost any function f and g, propagating along the unit vectors n_k and $-n_k$, respectively. Equation 12 displays the beauty and power of D'Alembert solution. One powerful strategy is present in many methods and consists in sending explicitly plane waves in many directions.

This is what comes out after an exam of the arguments in Equation 12. Consider that $x_k n_k = s$ is the equation of a plane in 3D space with a distance s from the origin, measured along n_k. In vector notation we can write $\mathbf{x} \cdot \mathbf{n} = s$. This is the definition of a new coordinate s and is illustrated using the Cartesian system in 2D depicted in Figure 3. The line L corresponds to a constant value for s.

Fourier transforms and plane waves

In order to describe and generalize the power of the plane wave concept, consider the representation of a function



Propagation of Elastic Waves: Fundamentals,
Figure 3 Coordinate s along the vector **n**. Line L corresponds to a constant value for s.

$$f(t) = \frac{1}{2\pi}\int_{\infty}^{\infty} F(\omega)\exp(+i\omega t)\mathrm{d}\omega, \tag{13}$$

where $F(\omega)$ = Fourier transform of $f(t)$ given by

$$F(\omega) = \int_{-\infty}^{\infty} f(t)\exp(-i\omega t)\mathrm{d}t. \tag{14}$$

Equation 13 exhibits an arbitrary function in terms of continuous superposition of sines and cosines. This fundamental property can be practically exploited under the assumptions that signals are discrete and periodic. We have then a powerful tool: the Discrete Fourier Transform. Moreover, the Fast Fourier Transform (FFT) algorithm allows for efficient computation. With this powerful tool we are effectively holding infinity in our hands. As the variable t in Equation 13 is a dummy variable we can substitute it by the D'Alembert argument (t-s/c). Therefore, it is possible to write

$$\begin{aligned} f(t-s/c) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)\exp[+i\omega(t-s/c)]\mathrm{d}\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)\exp[-i\omega s/c]\exp[+i\omega t]\mathrm{d}\omega. \end{aligned} \tag{15}$$

We first see D'Alembert argument at the complex exponential but then the contribution of the plane wave is given by the transfer function that multiplies the Fourier transform of signal. Let us consider the transfer function

$$\exp[-\mathrm{i}\omega s/c] = \exp[-\mathrm{i}(\omega/c)\mathbf{n}\cdot\mathbf{x}] = \exp[-\mathrm{i}\mathbf{k}\cdot\mathbf{x}], \tag{16}$$

which represents a harmonic plane wave propagating and identify that the wave vector $\mathbf{k} = (\omega/c)\mathbf{n}$ that gives the direction of propagation of the plane wave, where **n** is a unit vector ($n_1^2 + n_2^2 + n_3^2 = 1$).

It is possible to extend the idea that the unit vector **n** may have complex values then we have inhomogeneous plane waves. As an example of the significant representation power of generalized plane waves we can express a harmonic spherical wave as a superposition of both homogeneous and inhomogeneous plane waves using the Weyl integral (see Aki and Richards, 1980):

$$\frac{1}{R}\exp\left(-i\omega\frac{R}{c}\right) = \frac{1}{2\pi}\iint \frac{\exp(-ik_x x - ik_y y - \gamma|z|)}{\gamma}\mathrm{d}k_x\mathrm{d}k_y. \tag{17}$$

Here $\gamma = \sqrt{\omega^2/c^2 - k_x^2 - k_y^2}$ with the sign chosen so that Re $\gamma > 0$ and $R = \sqrt{x^2+y^2+z^2}$. Plane waves are

then building blocks to construct solutions for the half-space problems and layered media. They play a central role in seismology. Certainly, it is the success of a simple idea.

Green's function for homogeneous space

The Green's function is the displacement field produced within an elastic medium by a concentrated, impulsive unit load at a given point. For an infinite, homogeneous, isotropic, elastic medium Stokes found the analytical solution at the middle of the nineteenth century. More than 50 years later, at the eve of the twentieth century, Lamb (1904) found the solution for a vertical point load at the surface of a half-space. The solution for the horizontal load came again more than 50 years later (Chao, 1960).

If the body force is a unit concentrated impulsive it can be represented with Dirac's delta functions for space and time. Moreover, to specify the force direction we add Kronecker delta δ_{ij} (=1 if $i = j$, =0 for $i \neq j$). Therefore, we can define $f_i \equiv \delta_{ij}\delta(|\mathbf{x} - \xi|)\delta(t)$ and $u_i(\mathbf{x}, t) \equiv G_{ij}(\mathbf{x}, \xi; t)$ = Green's function for the infinite elastic domain. It is the displacement in direction i at $\mathbf{x}$ when a unit impulse is applied in direction j at point ξ.

The Green's function should satisfy Navier's equation, which can be written as

$$\mu\frac{\partial^2 G_{ij}}{\partial x_k \partial x_k} + (\lambda+\mu)\frac{\partial^2 G_{kj}}{\partial x_i \partial x_k} - \rho\frac{\partial^2 G_{ij}}{\partial t^2} = -\delta_{ij}\delta(|\mathbf{x} - \xi|)\delta(t). \tag{18}$$

On November 26th, 1849, G. G. Stokes read his paper on the Dynamical Theory of Diffraction. He intended to model a light source in the luminiferous Ether and assumed a force in an infinite domain. Based upon previous results of Poisson regarding Laplace equation, he obtained the exact solution for the displacement field due to a single force within the unbounded elastic medium. Without knowing it, he had conceived the first mathematical model of an earthquake. This solution (Stokes, 1849) can be expressed by means of

$$G_{ij}(\mathbf{x}, \xi; t) = \frac{1}{4\pi\mu r}\left\{f_1\gamma_i\gamma_j + f_2(\delta_{ij} - \gamma_i\gamma_j)\right\}, \tag{19}$$

where

$r = |\mathbf{x} - \xi| = \sqrt{(x_1 - \xi_1)^2 + (x_2 - \xi_2)^2 + (x_3 - \xi_3)^2}$

and $\gamma_j = (x_j - \xi_j)/r$ is the unit vector form ξ to $\mathbf{x}$. The functions f_1 and f_2 are given by

$$f_1(r, t) = \frac{\beta^2}{\alpha^2}\delta(t - r/\alpha) + 2\beta^2\int_{1/\alpha}^{1/\beta}\delta(t - r\kappa)\kappa d\kappa \tag{20}$$

and

$$f_2(r, t) = \delta(t - r/\beta) - \beta^2\int_{1/\alpha}^{1/\beta}\delta(t - \kappa r)\kappa d\kappa. \tag{21}$$

We can notice that Stokes' solution decays as $1/r$ and is modulated angularly by sines and cosines. It is remarkable that the longitudinal P-wave pulse has the same time dependence as the source followed by a similar transverse pulse with much larger amplitude. Between these main P and S arrivals we find a disturbance decaying more rapidly away from the source.

In the frequency domain (after applying Fourier transform) these functions can be written as follows:

$$f_1(r, \omega) = (\beta^2/\alpha^2)[1 - i2/qr - 2/q^2r^2]\exp(-iqr) + [i2/kr + 2/k^2r^2]\exp(-ikr) \tag{22}$$

and

$$f_2(r, \omega) = (\beta^2/\alpha^2)[i/qr + 1/q^2r^2]\exp(-iqr) + [1 - i/kr - 1/k^2r^2]\exp(-ikr). \tag{23}$$

Here $q = \omega/\alpha$ and $k = \omega/\beta$ are the wave numbers of P and S waves, respectively. If ω equals zero, the static case, we obtain the constants $f_1 = 1$ and $f_2 = (1 + \beta^2/\alpha^2)/2$ that correspond to the Kelvin solution for a unit static load in the full elastic space. The Stokes solution for the infinite space is frequently used in the integral BEM or IBEM formulations (see Bouchon and Sánchez-Sesma, 2007).

For a homogeneous or layered half-space, Fourier methods allow to formulate the problems in integral form. Obtaining the displacement and stress fields requires adequate discretization. A remarkable compendium of fundamental solutions is the one due to Kausel (2006).

The theory allows solving the homogeneous boundary conditions in terms of an eigenvalue problem. This lead to surface guided waves.

Green's function retrieval from correlations

If an elastic, inhomogeneous, anisotropic medium is subjected to isotropic, equipartitioned, illumination, the resulting field is said to be diffuse as its intensities are governed by a diffusion-like process. This may be easily argued if the role of multiple scattering is invoked. Under these circumstances, the Green's function can be retrieved from averaging cross-correlations of the recorded motions of such diffuse field (e.g., Weaver and Lobkis, 2004; Wapenaar, 2004; Sánchez-Sesma and Campillo, 2006; Gouédard et al., 2008).

The pioneering studies of Aki have contributed to the understanding of coda waves and seismic noise. Various scattering formulations have been developed in order to explain coda features (see Aki, 1957; Aki and Chouet, 1975; Sato and Fehler, 1998). When multiple scattering takes place, the intensities, which are related to energy densities, follow diffusion-like equations. These waves arrive at the receiver from different directions, sampling the medium along their paths.

When source and receiver are both at the same point, we have interesting consequences. From the observable side, the autocorrelation is proportional to the energy

density at a given point, and the imaginary part of the Green function at the source is finite because the singularity of the Green's function is restricted to the real part.

Thus, the directional energy densities at a point are proportional to the imaginary part of Green's function tensor components at the source itself, which is indeed the given point. The relationships among energy densities and its partitions have been recently studied by Perton et al. (2009) and Margerin et al. (2009). The Green´s function may be useful to imaging the subsurface structure at the site. In fact, the connection of the imaginary part of the Green´s function at the source with the optical theorem has been explored by Snieder et al. (2009).

For horizontally layered systems and 1D excitation, the relationship between reflection response and autocorrelation of surface motion was discovered by Claerbout (1968). An algorithm to identify reflection coefficients based upon Claerbout´s results was proposed by Scherbaum (1987). In fact, this is a single-station inversion method and, in principle, it allows imaging of the subsurface impedance structure. The idea was tested using small locally recorded earthquakes.

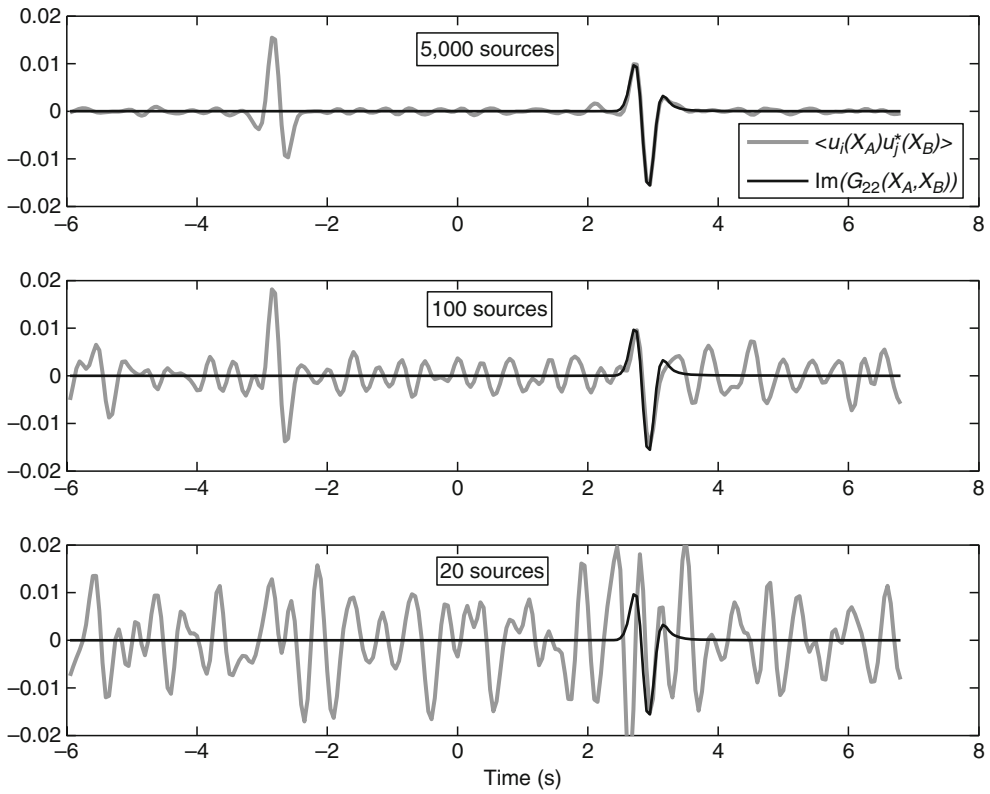
Seismic noise is produced by nearby, superficial low-strength sources that we will generically call microtremors. Noise sources are 3D in nature and have significant geometrical spreading. The relative sampling of the medium with respect to depth is considerably reduced due to this attenuation.

In what follows, we review the Green´s function retrieval from correlations of field fluctuations. For the layered medium with free surface overlaying a half-space we compute the imaginary part of the Green function when source and receiver coincide at the free surface.

It has been demonstrated (e.g., Weaver and Lobkis, 2004; Wapenaar, 2004; Sánchez-Sesma et al., 2008) that if a 3D diffuse harmonic displacement vector field $u_i(\mathbf{x}, \omega)$ is established within an elastic medium, the average cross-correlations of such motions at points $\mathbf{x_A}$ and $\mathbf{x_B}$ can be written as:

$$\left\langle u_i(\mathbf{x_A}, \omega) u_j^*(\mathbf{x_B}, \omega) \right\rangle = -2\pi E_S k^{-3} \times \mathrm{Im}\left[G_{ij}(\mathbf{x_A}, \mathbf{x_B}, \omega)\right]. \tag{24}$$

In this equation, the Green's function $G_{ij}(\mathbf{x}_\mathrm{A}, \mathbf{x}_\mathrm{B}, \omega) =$ displacement at $\mathbf{x}_\mathrm{A}$ in direction i produced by a unit harmonic point force acting at $\mathbf{x}_\mathrm{B}$ in direction j, $\omega =$ circular frequency, $k = \omega/\beta =$ shear wavenumber, $\beta =$ shear wave propagation velocity, and $E_S = \rho\omega^2 S^2 =$ average energy density of shear waves ($\rho =$ mass density,



Propagation of Elastic Waves: Fundamentals, Figure 4 Correlations are computed with 20, 100, and 5,000 sources, for the scalar case. We can see that as the number of sources considered increases the retrieval, the G_{22} function is better achieved.

S^2 = average spectral density). Note that the asterix implies the complex conjugate and the angular brackets mean azimuthal average. Equation 1 is the analytical consequence of a correlation-type elastic representation theorem and has been verified recently in canonical examples of a full space (Sánchez-Sesma and Campillo, 2006) and for an elastic inclusion embedded in a full space and for more general circumstances (Sánchez-Sesma et al., 2006, 2008). For the scalar case it is possible to compare the Green´s function retrieval from correlations. In Figure 4, we show how the G_{22} function is retrieved taking into account 20, 100, and 5,000 sources. The correlations are computed with these numbers of sources, and we can see that as the number of sources considered increases, the retrieval is better achieved.

Energy densities at given points and directions

In what follows we compute the theoretical energy density at a given point $\mathbf{x_A}$. In order to do so we rewrite Equation 24 assuming $\mathbf{x_A} = \mathbf{x_B}$:

$$E(\mathbf{x_A}) = \rho\omega^2 \langle u_m(\mathbf{x_A}) u_m^*(\mathbf{x_A}) \rangle = -2\pi\mu E_S k^{-1} \times \mathrm{Im}[G_{mm}(\mathbf{x_A}, \mathbf{x_A})], \quad (25)$$

where $\mu =$ shear modulus. The energy density of shear waves E_S is a measure of the strength of the diffuse illumination. We see that the total energy density at a point is proportional to the imaginary part of the trace of the Green tensor for coincident receiver and source.

Note that this is possible because the singularity of Green's function is restricted to the real part. The imaginary part is finite and regular and represents the rate of energy injected by the unit harmonic load at that point. This quantity "detects" the energy that goes back to the source-receiver and may be used to imaging. Equation 25 is valid even if the summation convention is ignored. In that case $E(\mathbf{x_A}) \equiv E_m(\mathbf{x_A})$and the energy density is associated to a particular direction (for discussions see Perton et al., 2009 and Snieder et al., 2009).

To illustrate these ideas from theoretical point of view let us consider again the Stokes' (1849) solution. It is possible to show that

$$\mathrm{Im}[G_{ij}(\mathbf{x}, \mathbf{x})] = -\frac{\omega}{12\pi\rho}\left(\frac{1}{\alpha^3} + \frac{2}{\beta^3}\right)\delta_{ij}. \quad (26)$$

This shows that the power injected into the infinite elastic medium by the unit harmonic load grows linearly with frequency, as expected, it is an isotropic tensor, and the energy labeled as being related to P and S waves is injected into the medium with the proportions given by the principle of equipartition for elastic waves $E_S/E_P = 2\alpha^3/\beta^3$. This is more than a fortuitous coincidence. It shows that an uncorrelated set of random forces uniformly distributed within a medium (simulating multiple scattering) can produce an equipartitioned elastic field, at least in some part of the elastic domain.

Summary

In this article dynamic elasticity fundamentals are discussed. This theory is useful to explain the propagation of elastic waves in solids. Newton's second law allowed writing the equilibrium equation within an elastic medium. Cauchy developed the concept of stress and strain while Hooke's idea to relate linearly stress and strain lead to Navier's equation which is in fact Newton's equation in terms of displacements. It can be solved using solutions of wave equations for P and S waves, discovered by Poisson. Both static and dynamic elasticity were developed in the nineteenth century. However, the first synthetic seismogram was computed at the beginning of the twentieth century. Plane waves are the simplest solutions of wave equation and are a key component to study elastic wave propagation in layered media and surface waves. The Green's functions in time or frequency domain are the solutions for an impulsive or harmonic point load, respectively. The canonical case of an infinite homogeneous space is due to Stokes (1849) and is given here.

It has been recently discovered that the Green's function can be retrieved from averaging correlations of a diffuse field. This applies for a very general medium. The relationship of energy and Green function allows establishing identities between average normalized measurements (correlations or autocorrelations) and intrinsic properties of the medium (imaginary part of Green function between two points or at the source itself) and opens the door to use noise measurements to construct virtual sources and do medium characterizations of unprecedented quality.

Acknowledgments

We thank E. Kausel and H. Sato for their keen and constructive remarks. This work was partially supported by DGAPA-UNAM Project IN121709, Mexico.

Bibliography

Achenbach, J. D., 1973. *Wave Propagation in Elastic Solids*. Amsterdam: North-Holland.

Aki, K., 1957. Space and time spectra of stationary stochastic waves with special reference to microtremors. *Bulletin. Earthquake Research Institute*, **35**, 415–456.

Aki, K., and Chouet, B., 1975. Origin of coda waves: Source, attenuation and scattering effects. *Journal of Geophysical Research*, **80**, 3322–3342.

Aki, K., and Richards, P. G., 1980. *Quantitative Seismology. Theory and Methods*. San Francisco: W. H. Freeman.

Bouchon, M., and Sánchez-Sesma, F J., 2007. Boundary integral equations and boundary elements methods in elastodynamics. In Wu, R.-S., Maupin, V., and Dmowska, T. (eds.), *Advances in Wave Propagation in Heterogeneous Earth, Advances in Geophysicis*. Elsevier-Academic, New York, Vol. 48. pp. 157–189.

Chao, C. C., 1960. Dynamical response of an elastic half-space to tangential surface loadings. *Journal of Applied Mechanics*, **27**, 559–567.

Claerbout, J. F., 1968. Synthesis of a layered medium from its acoustic transmission response. *Geophysics*, **33**, 264–269.

Gouédard, P., Stehly, L., Brenguier, F., Campillo, M., Colin de Verdiére, Y., Larose, E., Margerin, L., Roux, P., Sánchez-Sesma,

F. J., Shapiro, N. M., and Weaver, R. L., 2008. Cross-correlation of random fields: mathematical approach and applications. *Geophysical Prospecting*, **56**, 375–393.
Kausel, E., 2006. *Fundamental solutions in elastodynamics. A Compendium Cambridge*. New York: Cambridge University Press.
Lamb, H., 1904. On the propagation of tremors over the surface of an elastic solid. *Philosophical Transactions of the Royal Society (London), A*, **203**, 1–42.
Margerin, L., 2009. Generalized eigenfunctions of layered elastic media and application to diffuse fields. *The Journal of the Acoustical Society of America*, **125**, 164–174.
Margerin, L., Campillo, M., van Tiggelen, B. A., and Hennino, R., 2009. Energy partition of seismic coda waves in layered media: theory and application to Pinyon Flats Observatory. *Geophysical Journal International*, **177**, 571–585.
Perton, M., Sánchez-Sesma, F. J., Rodríguez-Castellanos, A., Campillo, M., and Weaver, R. L., 2009. Two perspectives on equipartition in diffuse elastic fields in three dimensions. *The Journal of the Acoustical Society of America*, **126**, 1125–1130, doi:10.1121/1.3177262.
Sánchez-Sesma, F. J., and Campillo, M., 2006. Retrieval of the green function from cross-correlation: the canonical elastic problem. *Bulletin. Seismological Society of America*, **96**, 1182–1191.
Sánchez-Sesma, F. J., Pérez-Ruiz, J. A., Campillo, M., and Luzón, F., 2006. The elastodynamic 2D Green function retrieval from cross-correlation: The canonical inclusion problem. *Geophysical Research Letters*, **33**, L13305, doi:10.1029/2006 GL026454.
Sánchez-Sesma, F. J., Pérez-Ruiz, J. A., Luzón, F., Campillo, M., and Rodríguez-Castellanos, A., 2008. Diffuse fields in dynamic elasticity. *Wave Motion*, **45**, 641–654.
Sato, H., and Fehler, M., 1998. *Wave Propagation and Scattering in the heterogeneous Earth*. New York: Springer.
Scherbaum, F., 1987. Seisimic imaging of the site response using microearthquake recordings. Part 1. Method, *Bulletin. Seismological Society of America*, **77**, 1905–1923.
Snieder, R., Sánchez-Sesma, F. J., and Wapenaar, K., 2009. Field fluctuations, imaging with backscattered waves, a generalized energy theorem, and the optical theorem. *SIAM Journal on Imaging Sciences*, **2**, 763–776.
Stokes, G. G., 1849. On the dynamical theory of diffraction. *Transactions of the Cambridge Philosophocal Society*, **9**, 1–62.
Wapenaar, K., 2004. Retrieving the elastodynamic Green's function of an arbitrary inhomogeneous medium by cross correlation. *Physical Review Letters*, 93:254301-1-4.
Weaver, R. L., and Lobkis, O. I., 2004. Diffuse fields in open systems and the emergence of the Green's function. *The Journal of the Acoustical Society of America*, **116**, 2731–2734.

Cross-references

Body Waves
Poroelasticity
Propagation of Elastic Waves: Fundamentals
Seismogram Interpretation
Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting
Statistical Seismology
Surface Waves

R

RADIOACTIVITY IN EARTH'S CORE

V. Rama Murthy
Department of Earth and Planetary Sciences, Institute of Meteoritics, University of New Mexico, Albuquerque, NM, USA

Definition

Lithophile. Affinity for silicates
Chalcophile. Affinity for sulfur
Siderophile. Affinity for metals
BSE. Bulk Silicate Earth refers to primitive silicate material in the Earth from which the core is separated. Corresponds to the total mantle + crust system of silicates now.
P, T. Pressure, temperature
TW. Terawatt. 1 TW = 10^{12} W
D_K. Concentration of K in Fe-alloy/Concentration of K in silicate melt.
GPa. GigaPascal =10^9 Pascals (1 GigaPascal is equivalent to 10 kilobars of pressure)
Geoneutrinos. Electron antineutrinos produced inside the Earth due to β^- decay of naturally occurring radioactive elements in the Earth.
Oxygen fugacity, fO_2. A measure of the oxidation state of a system irrespective of the presence or absence of a gas phase containing free oxygen.
IW buffer. Iron–Wustite buffer. A synthetic redox mineral reference buffer representing the oxygen fugacity of a system where Fe_{metal} is in equilibrium with FeO (wustite), according to the reaction $Fe_{metal} + ½O_2 =$ FeO (wustite). The IW buffer corresponds to log $fO_2 = -12.5$ at 1,200°C.

Introduction

The Earth is a thermal engine driven by its internal heat. The bulk of this heat is produced by the elements potassium (K), uranium (U), and thorium (Th) through the radioactive decay of their isotopes ^{40}K, ^{235}U, ^{238}U, and ^{232}Th. Table 1 lists the atomic percentages, radioactive decay constants, half-life, and heat production characteristics of these radioactive isotopes.

Because K, U, and Th are strongly lithophile, differentiation of the Earth into a metallic core and a silicate primitive mantle (BSE) will have partitioned them overwhelmingly into the BSE. A long-standing convention is that the radioactivity of the BSE represents the total radioactivity of the Earth and that the metallic core has no radioactivity (e.g., McDonough, 1999, 2003). However, radioactivity in the core has come into focus due to recent developments in three areas: (1) modern high P, T experiments on the metal-silicate partitioning behavior of K and U; (2) theoretical and experimental advances in our understanding of the electronic structure of potassium under conditions of high pressure and temperature; and (3) thermal models of the core that satisfy the size and age of the inner core and the energy to sustain geomagnetic field. Also, it is known that the heat due to radioactivity in BSE, about 19 TW (e.g., compiled in Lodders and Fegley, 1998), falls short of the surface heat flux estimated at 30–44 TW (Lee, 1970; Hofmeister and Criss, 2005; Pollack et al., 1993). This shortfall is suggestive of either a higher than BSE radioactive content of the mantle and/or radioactivity in the core (see Murthy, 2006). In the discussion below, new experimental developments are emphasized, citing where necessary the models and calculations that bear upon the radioactivity in the core.

Potassium radioactivity in the Earth's core

Potassium radioactivity in the Earth's core has been a topic of discussion for nearly four decades. Purely on geochemical grounds, Hall and Murthy (1971) and Lewis (1971) proposed that K enters the metallic core of the Earth as a consequence of the presence of sulfur in the core

Harsh K. Gupta (ed.), *Encyclopedia of Solid Earth Geophysics*, DOI 10.1007/978-90-481-8702-7,

Radioactivity in Earth's Core, Table 1 Important heat-producing radioactive isotopes in the Earth: atomic percentages, decay constants, half-lives, and heat production values (After Van Schmus, 1995)

Isotope	Atomic percentage	Decay constant λ (year^{-1})	Half-life (years)	Specific isotopic heat production (μWkg^{-1})
^{40}K	0.01167	5.54×10^{-10}	1.251×10^{9}	29.17
^{235}U	0.7200	9.85×10^{-10}	7.038×10^{8}	568.7
^{238}U	99.2743	1.551×10^{-10}	4.468×10^{9}	94.65
^{232}Th	100.00	4.95×10^{-10}	1.401×10^{10}	26.38

(Murthy and Hall, 1970). Since then, the implications of K in the Earth's core for the energetics of the core, convection and core cooling rate, growth rate of the inner core, the geomagnetic field, and thermal evolution of the Earth have been noted in several studies (e.g., Verhoogen, 1973; Goettel, 1976; Gubbins et al., 1979; Stevenson et al., 1983). Early experiments, however, concluded against the presence of K in the core (Oversby and Ringwood, 1972; Ganguly and Kennedy, 1977; Murrell and Burnett, 1986). These low pressure (1 bar to −1.5 GPa) and low temperature (1,030–1,350°C) studies yielded very low distribution coefficients, D_K for potassium. The laboratory D_K values applied to core formation led to the conclusion that K cannot be present in an S-bearing core.

Since these first-generation experiments, we have made a substantial progress in understanding the partitioning behavior of elements. For example, we know now that the partitioning of an element between metal and silicate depends on a number of variables such as pressure, temperature, oxygen fugacity (fO_2), and the composition of both the metal and silicate melts (Murthy, 1991; Hillgren et al., 1996; Jana and Walker, 1997; Righter et al., 1997; Gessmann and Rubie, 1998, 2000; Gessmann et al., 1999; Li and Agee, 2001; Bouhifd and Jephcoat, 2003; Chabot et al., 2005). In addition, core formation models suggest metal-silicate partitioning at a pressure range of 30–60 GPa and a temperature of 3,000–4,000°C (e.g., see Rubie et al., 2003 and Chabot et al., 2005 and the discussion and references cited therein). In the context of these developments, several new experiments have investigated the possibility of K-radioactivity in the core.

Chabot and Drake (1999) made the first systematic study of the effect of some of the variables on K partition into metal alloys at 1.5 GPa and 1,900°C. They showed that D_K was dependent on the S content of the metal alloy and on the composition of the silicate melt. Their experiments yielded low values of D_K ($\sim 6 \times 10^{-3}$) leading to their conclusion that the K content in an S-bearing metallic core will be trivially small (<1 ppm). The D_K values obtained by Chabot and Drake (1999) are lower by an order of magnitude and at variance with a number of later studies. This discrepancy may be due to their capsule choice (Gessman and Wood, 2002) or due to other experimental artifacts caused by the extreme lability of K under experimental conditions (Murrell and Burnett, 1986; Murthy et al., 2003).

Studies of K partitioning at high pressure and temperature (2.5–24 GPa and 1,500–1,900°C) by Gessman and Wood (2002) found a significant entry of K into Fe-S-O alloys. K partition into the metallic liquids depended strongly on S and O contents of metal alloy and increased with temperature. K had little tendency to partition into S-free metal alloys, as was noted in many previous studies. They show clearly that the partitioning behavior of K is strongly influenced by anion/cation ratio of the Fe-sulfide liquid as well as the depolymerization level of the silicate liquid noted previously by Chabot and Drake (1999). Evaluating the effect of these and other variables on D_K, Gessman and Wood (2002) suggest the possibility of 100 ppm of K content in a core containing about 10% S and possibly up to a maximum of ~250 ppm for conditions relevant to core separation from a deep magma ocean.

Murthy et al. (2003) found a significant partition of K into Fe–FeS liquids in experiments at 1–3 GPa at temperatures above the liquidus of both metal and silicate phases (1,200–1,800°C) and at fO_2 ~1.5 log units below IW buffer, corresponding to core separation conditions in a magma ocean (Walter and Thibault, 1995; Chabot and Agee, 2003; Chabot et al., 2005; Wade and Wood, 2005). D_K seemed positively correlated with temperature as noted in previous studies but not correlated with pressure in the pressure range of the experiments. Extrapolating to deep magma ocean conditions and a core with ~10% S, the inferred K content is 60–130 ppm, a little lower but comparable to that inferred by Gessman and Wood (2002).

Bouhifd et al. (2007) studied the partition of K between silicates and FeS-rich alloy and pure Fe metal at 5–15 GPa and a temperature of 1,900°C at an fO_2 between 1.5 and 3 log units below IW buffer. D_K was clearly positively correlated to S and O contents of the metal alloy and temperature as observed in previous experiments. In the pressure range of their experiments, no pressure dependence of D_K was observed both in sulfur-free and sulfur-bearing metal alloys. A parameterization of all existing D_K data showed a linear relationship between D_K and temperature, a relationship useful to predict D_K values at plausible temperature relevant to core formation if the S and O contents of the core are known. For reasonable assumptions of S and O contents, Bouhifd et al. (2007) found K content of ~250 ppm in the core. If the core is free of S and O, the K content falls as low as ~25 ppm.

Partitioning experiments between peridotite silicate liquids and Fe-Ni-S-C-O melts at pressures of 1.0–7.7 GPa and temperatures of 1,650–2,200°C by Corgne et al. (2007) did not find any relationship between D_K and temperature, as observed by Gessman and Wood (2002), Murthy et al. (2003), and Bouhifd et al. (2007). Nor did they find any dependence of D_K on the S content of metal alloy, as found in Chabot and Drake (1999), Gessman and Wood (2002), Murthy et al. (2003) and Bouhifd et al. (2007). A compositional control on D_K by the O content of the metal alloy was observed leading to their conclusion that a core containing S, C, and about 2 wt.% O could only have up to 25 ppm of K irrespective of the S and C levels. For a possible O content of ~5 wt. % (Badro et al., 2007), the K content of the core is ~80 ppm. However, the discrepancy between this work and the other recent studies is unresolved at present.

Quantum mechanical calculations predict a change in electronic structure of K at high pressures (from 4d-like orbitals to 3s-like orbitals), making K behave like a transition element (Bukowinski, 1976; Lee et al., 2004). This prediction is experimentally confirmed by the work of Lee and Jeanloz (2003) that showed the alloying of K with a high-pressure polymorph of Fe at pressures >26 GPa and temperature 2,500 K. The change in chemical bonding behavior of K under pressure is noted in several other experiments (Ito et al., 1993; Parker et al., 1996; Hirao et al., 2006; Bouhifd et al., 2007). These experiments and calculations show that K entry into a Fe-metal core is facilitated by high temperatures and high pressures, making it siderophile at extreme conditions.

Radioactivity in the core is suggested in a number of studies concerned with the age and size of the inner core, and the energy required to sustain the geomagnetic field. The estimated core–mantle boundary (CMB) heat flux is 6–12 TW (Buffett, 2003) and controls the rate of core cooling. Recent studies (Buffett, 2003; Nimmo et al., 2004; Costin and Butler, 2006) have examined the question of how best to reconcile the CMB heat flux with the size of inner core, the ~3.5 Ga history of the magnetic field, and the CMB temperature. These authors note that the presence of K in the core at ~200–400 ppm is in best accord with the present size of the inner core and the power needs of a geomagnetic dynamo. Using somewhat different parameters, similar conclusions have been reached by others (for example, Labrosse, 2003; Buffett, 2003; Butler et al., 2005; Costin and Butler, 2006). Thus, it appears that a variety of thermal models of the core call for an additional heat source in the core. In view of the recent experimental data, it is reasonable to attribute this at least in part to the radioactivity of K. However, no precise estimates are possible at present. A more definitive evaluation of the influence of temperature, pressure, and other variables on K partitioning into metal alloys is needed. There is also an uncertainty of the mantle K content at the time of core separation (Murthy, 2006) although the convention is to use the model BSE content of ~240 ppm. Nor are the S and O contents of the core known. Given these caveats, a core K content of up to 250 ppm of K seems tenable. The corresponding radioactive heat production in the core, especially with any accompanying U and Th (see discussion below) would be ~2 TW, a significant fraction of the CMB heat flux. However, the uncertainty of these figures cannot be overemphasized. Table 2 summarizes the current estimates of K-radioactivity in the core from standard geochemical models, thermal modeling of the core, and recent high P, T partition experiments.

Radioactivity in Earth's Core, Table 2 A comparison of inferred values of potassium abundance in the core from geochemical models, theoretical calculations, and recent experiments, and the corresponding heat production in terawatts (TW = 10^{12} W) today

Method	K abundance in Core (ppm)	Heat production in Core (TW)	Reference
Geochemical Models			
BSE Models	0	0	McDonough (1999, 2003)
E-chondrite	550 ± 260	~4–5	Lodders (1995)
Geophysical and	200–400	–	Buffett (2003)
Geodynamo Models	250–750	–	Labrosse (2003)
	Up to 1,420	9	Roberts et al. (2003)
	400	~3	Nimmo et al. (2004)
	300	–	Costin and Butler (2006)
Experiments			
Sulfur-bearing alloys	<1	0.01	Chabot and Drake (1999)
	100–250	0.8–2.0	Gessman and Wood (2002)
	60–130	0.4–0.8	Murthy et al. (2003)
	~250	~1.5	Bouhifd et al. (2007)
	~25–80	0.2–0.5	Corgne et al. (2007)
Fe-metal			
	~25	0.17	Bouhifd et al. (2007)
	~7,000[a]	–	Lee and Jeanloz (2003)
	35	0.23	Hirao et al. (2006)

[a]For realistic conditions of core segregation, K in core will be much less (see discussion in Lee and Jeanloz, 2003, 2004)

U (and th) radioactivity in the Earth's core

There have been some attempts to evaluate the presence of U (and Th) in the core. Earlier experiments were largely at low pressure and low temperature. Furst et al. (1982) noted that under the highly reduced conditions relevant to core formation in the Earth, U and Th tend to become chalcophile (affinity for sulfur) and might enter a sulfur-bearing core. The chalcophile tendency of U under these conditions is confirmed by Murrell and Burnett (1986) and in the recent high P, T experiments described later. Feber et al. (1984) noted the significant solubility of UO_2 in impure Fe-metal at 1 bar pressure and at temperature >3,000 K suggesting that significant U could be present in the core. In a series of detailed experiments, Murrell and Burnett (1986) measured the partitioning of K, U, and Th between silicate and sulfide liquids at low pressures (1 bar to 1.5 GPa) and found that U and Th tend to be less lithophile under reducing conditions and are markedly chalophile and partition at greater levels than K into Fe–FeS liquids.

The high P, T partitioning experiments of U described below yield somewhat ambiguous results. Bao et al. (2005, 2006) investigated the partitioning of U between silicate melt and pure Fe liquids at pressures of 3.0–14.5 GPa and temperatures of 1,660–2,500°C. They found that partition of U into the metal is dependent on whether the silicates were molten or solid and increases with both P and T. They found a similar P and T dependence of U in Fe-10 wt% S and Fe-35wt% S liquids, with a positive dependence on S content of the metal phase. Inferred values for U content in core are in the range of 1–4 ppb for pure Fe and at least 10 ppb for 10wt% S in core. At these latter levels, U alone can produce 1–2 TW of heat energy. Together with K, radioactivity in the core can then account for a significant fraction of the CMB heat flux. In contrast to these observations, Wheeler et al. (2006) found that U entry into core is not supported in their experiments for conditions up to 10 GPa, 2,300°C and 28 wt% S at fO_2 about 2 log units below IW. The discrepancy between these two studies is not understood.

Malavergne et al. (2007) found that U partitioning into core metal is dependent on S content and highly reducing conditions (fO_2 about 4–5 log units below IW buffer) and at temperatures up to 2,400°C and 20 GPa. These experiments permit a range of 0.0003–0.63 ppb U in core depending on whether the metal is S free or S rich. In experiments at temperatures of 2,200–2,400°C and pressures up to 8 GPa, under very low oxygen fugacity (~4 log units below IW), Murthy et al. (2007) found U partitioning into metal to be positively correlated with S content. The data permit U concentration in the core of 1–2 ppb with a heat production of 0.1–0.3 TW.

In a series of papers, Herndon (see for example, Herndon, 1998, 2006, and references to his work cited therein) has argued that gravitationally segregated U in the core functions as a nuclear reactor to provide substantial energy. In order to segregate uranium into the core, the core–mantle differentiation needs to have occurred at fO_2 of about 4–5 log units below IW buffer (Malavergne et al., 2007). Such a scenario would have reduced most iron in the planet to metallic form, leaving little oxidized iron in the mantle. This is not supported by the fact that the mantle contains ~8% of oxidized iron. The mantle would also have been strongly depleted in Ta under such a low oxidation, which is not the case (Mann et al., 2009). Current ideas of core formation based on siderophile element partitioning in the mantle suggest a much higher oxygen fugacity of about 2 log units below IW. Schuiling (2006) discusses other difficulties of the Herndon proposal.

Conclusions

There have been many theoretical discussions of radioactivity in the Earth's core in the past. But it is the new high P, T experiments of metal-silicate partitioning of K and U that urge a serious consideration of radioactivity in the Earth's core. The experimental data could permit up to 250 ppm of K and up to 10 ppb of U in the core at the upper end. The corresponding heat production due to K and U (and associated Th) could be ~2–4 TW, a significant fraction of the CMB heat flux. As mentioned in the section on K-radioactivity, the uncertainty in these numbers is very large. It may not be possible to precisely determine the radioactivity of the core by currently available geochemical and geophysical analyses except to note the general possibility of radioactivity in the core.

A way out of this impasse may now be available. Radioactive β^- decays of K, U, and Th in the interior of the Earth produce electron antineutrinos ($\overline{\nu_e}$).

These are termed geoneutrinos, and can serve as probes to directly measure the radioactivity of the Earth's mantle and core (e.g., Raghavan et al., 1998; Rothschild et al., 1998; Mantovani et al., 2004; Enomoto et al., 2005; Fiorentini et al., 2005; Giammarchi and Miramonti, 2006). Geoneutrino flux measurements use the antineutrino inverse β-decay reaction with protons, $\overline{\nu}_e + p \rightarrow e^+ + n$ in an appropriate liquid in a scintillator chamber. The resulting positron and neutron emissions are then measured by usual scintillation techniques. Geoneutrinos from each radioactive isotope can be distinguished by their characteristic energy spectrum. Neutrino detectors exist now in several countries, and new ones less susceptible to interference by nuclear reactors and other near surface radioactive sources are planned. These developments in neutrino geophysics should be able to measure the U and Th abundances in the deep Earth and the core. Present detection techniques do not allow measurement of K due to the lower energy of its antineutrinos compared to those from U and Th decay. Further developments are needed for K determination. Geoneutrino measurements are likely to provide a definitive answer about the level of radioactivity in the core and in addition clarify several questions regarding the Earth's formation, differentiation, chemical composition, heat budget, and geodynamics (Sleep, 2006).

Bibliography

Badro, J., Fiquet, G., Guyot, F., Gregoryanz, E., Occelli, F., Antonangeli, D., and d'Astuto, M., 2007. Effect of light elements on the sound velocities in solid iron: implications for the composition of the Earth's core. *Earth and Planetary Science Letters*, **254**, 233–238.

Bao, X., Secco, R. A., Gagnon, J. E., and Fryer, B. J., 2005. Experiments of U solubility in Earth's Core. *EOS Transactions*, American Geophysical Union. 86 (18): Jt. Assem. Suppl. (abstr).

Bao, X., Secco, R. A., Gagnon, J. E., and Fryer, B. J., 2006. Uranium partitioning between liquid iron and silicate melt at high pressures: implications for uranium solubility in planetary cores. arXiv:astro-ph/0606614.

Bouhifd, M. A., and Jephcoat, A. P., 2003. The effect of pressure on partitioning of Ni and Co between silicate and iron-rich metal liquids: a diamond-anvil cell study. *Earth and Planetary Science Letters*, **209**, 245–255.

Bouhifd, M. A., Gautron, L. A., Bolfan-Casanova, N., Malavergne, V., Hammouda, T., Andrault, D., and Jephcoat, A. P., 2007. Potassium partitioning into molten iron alloys at high pressure: implications for Earth's core. *Physics of the Earth and Planetary Interiors*, **160**, 22–33.

Buffett, B. A., 2003. The Thermal State of Earth's Core. *Science*, **299**, 1675–1676.

Bukowinski, M. S. T., 1976. The effect of pressure on the physics and chemistry of potassium. *Geophysical Research Letters*, **3**, 491–503.

Butler, S. L., Peltier, W. R., and Costin, S. O., 2005. Numerical models of the Earth's thermal history: effects of inner-core solidification and core potassium. *Physics of the Earth and Planetary Interiors*, **152**, 22–42.

Chabot, N. L., and Drake, M. J., 1999. Potassium solubility in metal: the effects of composition at 15 kbar and 1900°C on partitioning between iron alloys and silicate melts. *Earth and Planetary Science Letters*, **172**, 323–335.

Chabot, N. L., and Agee, C. B., 2003. Core formation in the Earth and Moon: new experimental constraints from V, Cr, and Mn. *Geochimica et Cosmochimica Acta*, **67**, 2077–2091.

Chabot, N. L., Draper, D. S., and Agee, C. B., 2005. Conditions of core formation in the Earth: constraints from nickel and cobalt partitioning. *Geochimica et Cosmochimica Acta*, **69**, 2141–2151.

Corgne, A., Kesav, S., Fei, Y., and McDonough, W., 2007. How much potassium is in the Earth's core? New insights from partitioning experiments. *Earth and Planetary Science Letters*, **256**, 567–576.

Costin, S. O., and Butler, S. L., 2006. Modelling the effects of internal heating in the core and lowermost mantle on the earth's magnetic history. *Physics of the Earth and Planetary Interiors*, **157**, 55–71.

Enomoto, S., Ohtani, E., Inoue, K., and Suzuki, A., 2005. Neutrino geophysics with KamLAND and future prospects. arXiv:hep-ph/0508049.

Feber, R. C., Wallace, T. C., and Libby, L. M., 1984. Uranium in the Earth's core. *EOS. Transactions of the American Geophysical Union*, **65**, 785.

Fiorentini, G., Lissia, M., Mantovani, F., and Vanucci, R., 2005. Geo-neutrinos: a new probe of Earth's interior. *Earth and Planetary Science Letters*, **238**, 235–247.

Furst, M. J., Stapanian, M. I., and Burnett, D. S., 1982. Observation of non-lithophile behavior for U. *Geophysical Research Letters*, **9**, 41–44.

Ganguly, J., and Kennedy, G. C., 1977. Solubility of K in Fe-S liquid, Silicate-K-(FeS)liq equilibria, and their planetary implications. *Earth and Planetary Science Letters*, **35**, 411–420.

Gessmann, C. K., and Rubie, D. C., 1998. The effect of temperature on the partitioning of nickel, cobalt, manganese, chromium, and vanadium at 9GPa and constraints on formation of the Earth's core. *Geochimica et Cosmochimica Acta*, **62**, 867–882.

Gessmann, C. K., Rubie, D. C., and McCammon, C. A., 1999. Oxygen fugacity dependence of Ni, Co, Mn, Cr, V, and Si partitioning between liquid metal and magnesiow¨ustite at 9–18 GPa and 2200°C. *Geochimica et Cosmochimica Acta*, **63**, 1853–1863.

Gessmann, C. K., and Rubie, D. C., 2000. The origin of the depletions of V, Cr, Mn in the mantles of the Earth and Moon. *Earth and Planetary Science Letters*, **184**, 95–107.

Gessman, C. K., and Wood, B. J., 2002. Potassium in the Earth's core? *Earth and Planetary Science Letters*, **200**, 63–78.

Giammarchi, M. G., and Miramonti, L., 2006. Borexino: Geoneutrinos in Borexino. *Earth, Moon and Planets*, **99**, 207–220.

Goettel, K. A., 1976. Models for the origin and composition of the Earth, and the hypothesis of potassium in the earth's core. *Geophysical Surveys*, **2**, 369–397.

Gubbins, D., Masters, T. G., and Jacobs, J. A., 1979. Thermal evolution of the Earth's core. *Geophysical Journal of the Royal Astronomical Society*, **59**, 57–99.

Hall, H. T., and Murthy, V. R., 1971. The early chemical history of the earth: some critical elemental fractionations. *Earth and Planetary Science Letters*, **11**, 239–244.

Herndon, J. M., 1998. Composition of the deep interior of the earth: divergent geophysical development with fundamentally different geophysical implications. *Physics of the Earth and Planetary Interiors*, **105**, 1–4.

Herndon, J. M., 2006. Solar system processes underlying planetary formation, geodynamics, and the georeactor. *Earth, Moon and Planets*, **99**, 53–89.

Hillgren, V. J., Drake, M. J., and Rubie, D. C., 1996. High pressure and high temperature metal-silicate partitioning of siderophile elements: the importance of silicate liquid composition. *Geochimica et Cosmochimica Acta*, **60**, 2257–2263.

Hirao, N., Ohtani, E., Kondo, T., Endo, N., Kuba, T., Suzuki, T., and Kikegawa, T., 2006. Partitioning of potassium between iron and silicate at the core-mantle boundary. *Geophysical Research Letters*. **33**, L08303, doi:10.1029/2005GL025324.

Hofmeister, A. M., and Criss, R. E., 2005. Earth's heat flux revised and linked to chemistry. *Tectonophysics*, **395**, 159–177.

Ito, E., Morooka, K., and Ujike, O., 1993. Dissolution of K in molten iron at high pressure and temperature. *Geophysical Research Letters*, **20**, 1651–1654.

Jana, D., and Walker, D., 1997. The influence of silicate melt composition on distribution of siderophile elements among metal and silicate liquids. *Earth and Planetary Science Letters*, **150**, 463–472.

Javoy, M., 1995. The integral enstatite chondrite model of the Earth. *Geophysical Research Letters*, **22**, 2219–2222.

Javoy, M., 1999. Chemical Earth models. In *Earth and Planetary Sciences*. C. R. Academy Science, 329, pp. 537–555.

Labrosse, S., 2003. Thermal and magnetic evolution of the Earth's core. *Physics of the Earth and Planetary Interiors*, **140**, 127–143.

Labrosse, S., and Macouin, M., 2003. The inner core and the geodynamo. *Comptes Rendus Geoscience*, **335**, 37–50.

Labrosse, S., Poirier, J.-P., and Le Mouel, J.-L., 2001. The age of the inner core. *Earth and Planetary Science Letters*, **190**, 111–123.

Lee, W. H. K., 1970. On the global variations of terrestrial heat-flow. *Physics of the Earth and Planetary Interiors*, **2**, 332–341.

Lee, K. K. M., and Jeanloz, R., 2003. High-pressure alloying of potassium and iron: radioactivity in the Earth's core? *Geophysical Research Letters*, **30**, 2212, doi:10.1029/2003GL018515.

Lee, K. K. M., Steinle-Neumann, G., and Jeanloz, R., 2004. Ab-initio high-pressure alloying of iron and potassium: implications for the Earth's core. *Geophysical Research Letters*, **31**, L11603, doi:10.1029/2004GL019839, 2004.
Lewis, J. S., 1971. Consequences on the presence of sulfur in the core of the Earth. *Earth and Planetary Science Letters*, **11**, 130–134.
Li, J., and Agee, C. B., 2001. The effect of pressure, temperature, oxygen fugacity and composition on partitioning of nickel and cobalt between liquid Fe-Ni-S alloy and liquid silicate: implications for the Earth's core formation. *Geochimica et Cosmochimica Acta*, **65**, 1821–1832.
Lodders, K., 1995. Alkali elements in the Earth's core: evidence from enstatite chondrites. *Meteoritics*, **30**, 93–101.
Lodders, K., and Fegley, B. J., Jr., 1998. *The Planetary Scientist's Companion*. Oxford: Oxford University Press.
Malavergne, V., Tarrida, M., Combes, R., Bureau, H., and Jones, J., 2007. New high-pressure and high-temperature metal/silicate partitioning of U and Pb: Implications for the cores of the Earth and Mars. *Geochimica et Cosmochimica Acta*, **71**, 2637–2655.
Mann, U., Frost, D. J., and Rubie, D. C., 2009. Evidence for high-pressure core-mantle differentiation from the metal-silicate partitioning of lithophile and weakly-siderophile elements. *Geochimica et Cosmochimica Acta*, **73**, 7360–7386.
Mantovani, F., Carmignani, G. L., Fiorentini, G., and Lissia, M., 2004. Antineutrinos from the Earth: a reference model and its uncertainties. *Physical Review D*, **69**, 297–314.
McDonough, W. F., 1999. Earth's core. In Marshall, C. P., and Fairbridge, R. W. (eds.), *Encyclopedia of Geochemistry.* Dordrecht: Kluwer Academic.
McDonough, W. F., 2003. Compositional model for the Earth's core. In Carlson, R. W. (ed.), *Treatise on Geochemistry.* The Mantle and the Core, Vol. 2, pp. 547–568.
Murrell, M. T., and Burnett, D. S., 1986. Partitioning of K, U, and Th between sulfide and silicate liquids: Implications for radioactive heating of planetary cores. *Journal of Geophysical Research*, **91**, 8126–8136.
Murthy, V. R., and Hall, H. T., 1970. The chemical composition of the Earth's core: possibility of sulphur in the core. *Physics of the Earth and Planetary Interiors*, **2**, 276–282.
Murthy, V. R., 1991. Early differentiation of the earth and the problem of mantle siderophile elements: a new approach. *Science*, **253**, 303–306.
Murthy, V. R., van Westrenen, W., and Fei, Y., 2003. Experimental evidence that potassium is a substantial radioactive heat source in planetary cores. *Nature*, **423**, 163–165.
Murthy, V. R., 2006. Radioactivity of the Earth and the case for potassium in the Earth's core. *Earth, Moon and Planets*, **99**, 23–32.
Murthy, V. R., Draper, D., and Agee, C., 2007. Uranium in the Earth's core? Metal-silicate partitioning of Uranium at High Pressure and Temperature and Highly Reducing Conditions. In *Workshop on Early Planetary Differentiation*. *Lunar Planetary Institute Contribution*. 1355, pp. 78–79.
Nimmo, F., Price, G. D., Brodholt, J., and Gubbins, D., 2004. The influence of potassium on core and geodynamo. *Geophysical Journal International*, **156**, 363–376.
Oversby, V. M., and Ringwood, A. E., 1972. Potassium distribution between metal and silicate and its bearing on the occurrence of potassium in the earth's core. *Earth and Planetary Science Letters*, **14**, 345–347.
Pollack, H. N., Hurter, S. J., and Johnson, J. R., 1993. Heat flow from the Earth's interior: analysis of the global data set. *Reviews of Geophysics*, **31**, 267–280.
Parker, L. J., Atou, T., and Badding, J. V., 1996. Transition element-like chemistry for potassium under pressure. *Science*, **273**, 95–97.
Raghavan, R. S., Schoenert, S., Enomoto, S., Shirai, S., Suekane, F., and Suzuki, A., 1998. Measuring the global radioactivity in the earth by multidetector antineutrino spectroscopy. *Physical Review Letters*, **80**, 636–638.
Righter, K., Drake, M. J., and Yaxley, G., 1997. Prediction of siderophile element metal/silicate partition coefficients to 20 GPa and 2800°C: the effects of pressure, temperature, oxygen fugacity, and silicate and metallic melt composition. *Physics of the Earth and Planetary Interiors*, **100**, 115–134.
Roberts, P. H., Jones, C. A., and Calderwood, A. R., 2003. Energy fluxes and Ohmic dissipation in the Earth's core. In Jones, C. A., Soward, A. M., and Zhang, K. (eds.), *Earth's Core and Lower Mantle*. Taylor: London.
Rothschild, C. G., Chen, M. C., and Calaprice, F. P., 1998. Antineutrino geophysics with liquid scintillation detectors. *Geophysical Research Letters*, **25**, 1083–1086.
Rubie, D. C., Melosh, H. J., Reid, J. E., Liebske, C., and Righter, K., 2003. Mechanisms of metal–silicate equilibration in the terrestrial magma ocean. *Earth and Planetary Science Letters*, **205**, 239–255.
Schuiling, R. D., 2006. Is there a nuclear reactor at the center of the earth? *Earth, Moon and Planets*, **99**, 33–49.
Sleep, N. H., 2006. Strategy for applying neutrino geophysics to the earth sciences including planetary habitability. *Earth, Moon and Planets*, **99**, 343–358.
Stevenson, D. J., Spohn, T., and Schubert, G., 1983. Magnetism and thermal evolution of the terrestrial planets. *Icarus*, **54**, 466–489.
Van Schmus, W. R., 1995. Natural radioactivity of the crust and mantle. In Ahrens, T. J. (ed.), *Global Earth Physics: A Handbook of Physical Constants, AGU Reference Shelf 1*. Washington, DC: American Geophysical Union, pp. 283–291.
Verhoogen, J., 1973. Thermal regime of the earth's core. *Physics of the Earth and Planetary Interiors*, **7**, 47–58.
Wade, J., and Wood, B. J., 2005. Core formation and the oxidation state of the Earth. *Earth and Planetary Science Letters*, **236**, 78–95.
Walter, M. J., and Thibault, Y., 1995. Partitioning of tungsten and molybdenum between metallic liquid and silicate melt. *Science*, **270**, 1186–1189.
Wheeler, K. T., Walker, D., Fei, Y., Minarik, W., and McDonough, W., 2006. Experimental partitioning of uranium between liquid iron sulfide and liquid silicate: implications for radioactivity in the Earth's core. *Geochimica et Cosmochimica Acta*, **70**, 1537–1547.

Cross-references

Core Dynamo
Core-Mantle Coupling
Energy Budget of the Earth
Geodynamics
Mantle Convection
Radiogenic Heat Production of Rocks

RADIOGENIC HEAT PRODUCTION OF ROCKS

Christoph Clauser
Institute for Applied Geophysics and Geothermal Energy, E.ON Energy Research Center, RWTH Aachen University, Aachen, Germany

Synonyms

Radiogenic heat generation

Definition

Radiogenic heat production rate. Physical property defining the amount of heat liberated in unit time in a unit volume of rock by the decay of unstable radiogenic isotopes; dimension: W m^{-3}.

Geoneutrino. An electron antineutrino emitted in β-decay of nuclei during radiogenic heat production caused by the decay of the unstable isotopes ^{238}U, ^{232}Th, and ^{40}K.

eV (electron Volt). A non-SI unit of energy in nuclear physics, defined as the kinetic energy gained by an electron of elementary charge when accelerating through an electric potential difference of 1 V. Thus, one electron Volt equals one Volt, which is one Joule per Coulomb, multiplied by the electron charge of $e = 1.602\,176\,487(40) \times 10^{-19}$ C. Therefore, $1\ eV = 1.602\,176\,487 \times 10^{-19}$ J.

ppm (parts per million). A non-SI unit of relative frequency (or abundance) in 10^{-6}, similar to % (percent) or ‰ (per mil) in 10^{-2} and 10^{-3}, respectively.

Radiogenic heat generation

The main interior sources of heat in the Earth are the heat content of the infant Earth immediately after formation due to gravitational contraction and the decay of unstable, radioactive isotopes. The potential energy released as heat during the gravitational accretion of the infant Earth dwarfs all other heat sources. This energy is calculated from gravitational contraction. The mass accumulated in a spheroidal planet of constant density ρ and radius r equals $m = (4\pi/3)\, r^3\rho$. Addition of a further shell of thickness dr and mass $dm = 4\pi\, r^2\rho\, dr$ results in the release of additional potential gravitational energy $dE_p = G\, m\, dm/r = (16/3)\, \pi^2\, G\, \rho^2\, r^4\, dr$, where $G = 6\,673 \times 10^{-11}\ m^3\ kg^{-1}\ s^{-2}$ is the universal gravitational constant. Integration yields:

$$E_p = \int_0^{r_E} dE_p = \int_0^{r_E} \frac{G\,m}{r}\,dm = \frac{16\,\pi^2\,\rho^2\,G}{3}\int_0^{r_E} r^4\,dr = \frac{16\,\pi^2\,\rho^2\,G}{3}\,\frac{r^5}{5} = \underbrace{\left(\frac{4\pi\,r_E^3\,\rho}{3}\right)}_{M_E}^{2}\,\frac{3\,G}{5\,r_E} = \frac{3}{5}\,\frac{G\,M_E^2}{r_E}, \quad (1)$$

where $M_E = 5.976 \times 10^{24}$ kg and $r_E = 6.371 \times 10^6$ m are mass and radius of the Earth, respectively. Accordingly, the accretion of the Earth results in the release of heat on the order of $E_p = 2.24 \times 10^{32}$ J. A refined analysis considering the radial variation of density in the Earth modifies this value by only about 10% to $E_p = 2.49 \times 10^{32}$ J (Stacey and Davis, 2008).

An additional, similarly large energy input is assumed to have occurred in the wake of an impact into the young infant Earth by a Mars-sized protoplanet orbiting the sun close to the Earth's orbit. The terrestrial Moon is generally believed to be the result of this impact, which released energy on the order of 10^{31} J (Melosh, 1990; Canup and Righter, 2000; Canup and Asphaug, 2001). However, most of the vast energy liberated during accretion of the Earth or by this impact was radiated back into space already during accretion or immediately after the impact (Stacey and Davis, 2008).

For an average specific heat capacity of 1 088 J kg^{-1} K^{-1} and a cooling by 650 K over a lifetime of 4.6 billion years, the mean rate by which the Earth is losing its original heat is 29 TW (Vacquier, 1991, 1992). This corresponds to a thermal energy of 4.2×10^{30} J. As the heat loss of the young and hot infant Earth was larger than today, the contribution of original heat to today's heat loss is certainly smaller. A larger cooling by 1 175 K would result in a mean cooling rate of 52 TW corresponding to a thermal energy of 7.6×10^{30} J (Stacey and Davis, 2008).

The annual production of radiogenic heat in the Earth, 6.3×10^{20} J (Jaupart et al., 2007), corresponds to more than twice the global production of primary energy in the year 2000. This huge energy source by itself clearly exceeds the world's energy annual demands predicted through the year 2030 (IEA, 2008). If it were used at great scale, it may satisfy a large proportion of the primary energy demand of the entire twenty-first century.

Apart from the heat content of the infant Earth immediately after formation, the radiogenic decay of the unstable isotopes of uranium (^{238}U; ^{235}U), thorium (^{232}Th), and potassium (^{40}K) provides the largest internal source of heat. Most of these isotopes are enriched in the Earth's crust and mantle (Table 1). During radioactive decay, mass is converted into energy. Except for the tiny amount associated with the antineutrinos and neutrinos generated in β^-- and β^+-decay or electron capture, respectively, all of this energy is converted into heat. Certain peaks in the corresponding γ-spectra are characteristic for the different decay series while the continuous background spectrum is due to Compton scattering and photoelectric absorption.

The two uranium isotopes, ^{238}U and ^{235}U, decay into lead, ^{206}Pb and ^{207}Pb, with a half-life of 4.5 and 0.71 billion years, respectively. Characteristic lines for the γ-spectrum of the ^{238}U decay series, for example, are produced by the daughter element bismuth, ^{214}Bi, at 609 keV, 1 120 keV, and 1 764 keV. Uranium is an abundant mobile trace element in many rocks. With a half-life of 14.1 billion years, thorium ^{232}Th decays also into lead ^{208}Pb. Characteristic lines for the γ-spectrum of the thorium decay series are produced by the daughter elements thallium, ^{208}Tl, or actinium, ^{228}Ac, at 584 keV and 2 615 keV or 912 keV and 966 keV, respectively. Thorium also occurs as a trace element, is relatively inactive chemically and frequently bound to clay minerals. The only unstable isotope of potassium is ^{40}K. It disintegrates by electron capture or emission into argon, ^{40}Ar, or calcium, ^{40}Ca, respectively, with a corresponding characteristic line in the

Radiogenic Heat Production of Rocks, Table 1 Average radiogenic heat generation rate per mass, A′, in geologic materials (data: McDonough and Sun, 1995; Stacey and Davis, 2008; Jaupart et al., 2007)

Material		Concentration (ppm by mass)				Heat generation (10^{-12} W kg^{-1})
		C_U	C_{Th}	C_K	C_K/C_U	A′
Igneous rocks	Granites	4.6	18	33 000	7 000	1 050
	Alkali basalt	0.75	2.5	12 000	16 000	180
	Tholeiitic basalt	0.11	0.4	1 500	13 600	27
	Eclogite	0.035	0.15	500	14 000	9.2
	Peridotite, dunite	0.006	0.02	100	17 000	1.5
Meteorites	Carbonaceous chondrite	0.0074–0.0080	0.029–0.030	544–550	20 000	5.2
Moon	Apollo samples	0.23	0.85	590	2 500	47
Earth	Average crust (2.8×10^{22} kg)	1.2–1.3	4.5–5.6	15 500	13 000	293–330
	Average mantle (4.0×10^{24} kg)	0.013–0.025	0.040–0.087	70–160	2 800	2.8–5.1
	Average core	0	0	29	–	0.1
	Bulk silicate Earth (BSE)	0.020 ± 20%	0.081 ± 15%	118 ± 20%	5 400	4.7 ± 0.08

γ-spectrum of the potassium decay at 1 460 keV. Potassium occurs in many clay minerals at concentrations of several percent.

Tabulated data

The compilations of measured radiogenic heat generation rate which are reported in several research papers and reference books will not be duplicated here. Čermák and Rybach (1982) present a large collection of data arranged according to rock type. Van Schmus (1984, 1995) presents data on the abundance of radiogenic isotopes in various minerals and rocks of the Earth's crust and mantle and discusses the variation of radiogenic heat generation over the lifetime of the Earth. Stacey and Davis (2008) present data on average mass specific heat generation rates for various geologic materials, and Jaupart et al. (2007) discuss radiogenic heat sources in the Earth's crust and mantle. They provide an in-depth discussion on current models for the composition of a bulk silicate Earth (BSE), that is, the Earth's crust and mantle without the core: These are based on data from samples of either (1) meteorites considered representative for the starting material and high-temperature processes in the early solar nebula when the Earth accreted or (2) upper mantle rocks formed by low-temperature processes. As can be expected, both show large variations in composition due to their different provenance and history.

Chondrites are considered representative of undifferentiated silicate material from the solar system prior to melting and the formation of a (mainly) iron core. Different classes of chondrites correspond to perturbations in elemental abundance in the gas state caused mainly by volatility or condensation temperature. The important elements with respect to radiogenic heat generation are uranium, thorium, and potassium. As elaborated by Jaupart et al. (2007), the first two condensate at very high temperature, why they are called "refractory lithophile" elements. They show the same ratio in all types of chondritic meteorites, which demonstrates that they behave similarly in the early solar system. In contrast, potassium is a "moderately volatile" element with a lower condensation temperature. The best agreement with solar concentration ratios is found for carbonaceous chondrites of the CI type of which, however, McDonough and Sun (1995) report only five finds. Compositional data exist only for three of them, in particular for the 700 g meteorite found near Ivuna in Tanzania in 1938 from which the name "CI" is derived for this type of carbonaceous chondrite. However, chondrites are known to be more enriched in volatiles such as H_2O and CO_2 than the Earth.

Samples from the Earth's mantle are called pyrolites, a contraction of the names of two principal mantle minerals, *pyr*oxene and *oli*vine, making up the complementary rocks peridotite and basalt. Because the former is the solid residue of the partial melting event which produced basalts, a mixture between the two was considered a suitable starting material, the pyrolytic mantle. Generally, pyrolite data suffer from leaching of uranium during low-temperature alteration. As most samples derive from the upper mantle, it has been questioned whether the pyrolytic composition adequately represents a bulk silicate Earth.

A third approach discussed by Jaupart et al. (2007) avoids a specific choice of a starting composition but determines it from the intersection of the two compositional trends of chondritic meteorites and peridotites. A source of error in this approach derives from the scatter in the two trends.

A fourth method proceeds from the elemental ratios of uranium and thorium and determines the primitive bulk Earth abundances from measurements on peridotites. Again, Jaupart et al. (2007) provide a concise introduction into this method. Table 1 summarizes mass-specific radiogenic heat generation rates for terrestrial rocks, meteorites, Moon surface samples, and average values for the Earth's crust, mantle, and a bulk silicate Earth (BSE).

Calculated heat generation rate

The energy emitted by all of these decay processes comprises the kinetic energy of the emitted particles and the γ-radiation associated with the different decay processes. It is absorbed in the rocks and finally transformed into heat. In general, the total heat generation rate A of a rock is the sum of the individual contributions A_U, A_{Th}, and A_K by uranium, thorium, and potassium, respectively:

$$A = \rho\left(C_U A'_U + C_{Th} A'_{Th} + C_K A'_K\right), \quad (2)$$

where ρ is rock density, and A′ and C are heat generation rate per mass and concentration, respectively, of the corresponding element in the rock. Table 2 shows values for A′ reported by different authors. The variation is less than 3 %, at most.

Thus, if a rock's density ρ and its concentrations in uranium (C_U), thorium (C_{Th}), and potassium (C_K) are known, its radiogenic heat generation rate A can be determined. Inserting the values, for example, of Rybach (1988) from Table 2 into Equation 2 yields:

$$A\left[\mu W\,m^{-3}\right] = 10^{-5} \times \rho\left[kg\,m^{-3}\right] \times (9.52 \times C_U[ppm] + 2.56 \times C_{Th}[ppm] + 3.48 \times C_K[\%]), \quad (3)$$

where concentrations are given in weight-ppm (i.e., 10^{-6} kg kg^{-1}) for uranium and thorium and in weight-% for potassium.

The natural γ-radiation of rocks can be measured, for instance, by spectrometry on rock samples in the laboratory (see Measuring Techniques). An alternative source of γ-spectra is the natural gamma spectrometer (NGS) borehole tool which yields as output the three logs URAN [weight-ppm uranium], THOR [weight-ppm thorium], and POTA [weight-% potassium]. In combination with density RHOB from the compensated density log [g cm^{-3}], this information can be used directly to obtain the bulk heat generation rate A from Equation 2 using any of the values for A'_U, A'_{Th}, and A'_K from Table 2.

Alternatively, for lack of spectra, the total γ-ray emission GR may be used, measured either in the laboratory or by suitable logging tools in boreholes. A γ-ray (GR) log is often included in most logging runs as one of the basic geophysical measurements in boreholes. Its reading reflects the combined radioactive decay of uranium, thorium, and potassium. However, it does not yield information on the individual contributions as it records the total number of γ-rays detected by the tool during a time interval. The relative gamma activities of uranium, thorium, and potassium, that is, the relative number of γ-rays emitted by the same mass during the same time interval are shown in Table 3.

Therefore, if POTA, THOR, and URAN are recorded as percent, ppm, and ppm, respectively, the gamma ray log's reading GR is proportional to the corresponding contents (Beardsmore and Cull, 2001):

$$GR = X \times (POTA + 0.13 \times THOR + 0.36 \times URAN), \quad (4)$$

where the proportionality X varies with the radius of sensitivity of the log for detecting γ-rays. This varies with a number of factors, but lies within a few decimeters, at most, and can be approximated as constant for a given borehole (Beardsmore and Cull, 2001). Then, bulk heat generation rate A can be related to GR by combining Equations 2 and 4 using the data of, for example, Rybach (1988) from Table 2:

$$\frac{A[\mu W m^{-3}]}{GR[API]} = \frac{10^{-3} \times RHOB[g cm^{-3}] \times (34.8 \times POTA[\%] + 25.6 \times THOR[ppm] + 95.2 \times URAN[ppm])}{X[API] \times (POTA[\%] + 0.13 \times THOR[ppm] + 0.36 \times URAN[ppm])},$$

that is:

$$A[\mu W m^{-3}] = \frac{0.0348}{X[API]} \times RHOB[g cm^{-3}] \times GR[API] \underbrace{\frac{(POTA[\%] + 0.736 \times THOR[ppm] + 2.736 \times URAN[ppm])}{(POTA[\%] + 0.130 \times THOR[ppm] + 0.360 \times URAN[ppm])}}_{Y} = 0.0348\left(RHOB \times GR \times \frac{Y}{X}\right). \quad (5)$$

Y in Equation 5 still contains the unknown relative elemental abundances. However, it turns out that the range for Y is limited between $1 \leq Y < 7.6$ for the two extreme cases when either 0 = THOR = URAN and POTA > 0, or 0 = POTA = THOR and URAN > 0. Beardsmore and Cull (2001) argue that in regions in which sediments have been derived from a common source through time the relative proportions of the elements in Equation 5, and hence Y, should remain relatively constant with depth. Thus, Beardsmore and Cull (2001) maintain that a plot of A (derived, e.g., from a NGS log) versus RHOB × GR should yield a gradient, which depends only on the relative

Radiogenic Heat Production of Rocks, Table 2 Heat generation rate per mass, A′, for uranium, thorium, and potassium (Clauser, 2009)

A'_U (μW kg^{-1})	A'_{Th} (μW kg^{-1})	A'_K (μW kg^{-1})	Data source
97.0	27.0	0.0036	Birch (1954) cited in Jessop (1990)
97.7	26.3	0.0034	Hamza and Beck (1972) cited in Jessop (1990)
96.7	26.3	0.0035	Emsley (1989) and Jessop (1990) cited in Beardsmore and Cull (2001)
95.2	25.6	0.00348	Rybach (1988)

Radiogenic Heat Production of Rocks, Table 3 Relative gamma activity of uranium, thorium, and potassium (data source: Adams and Weaver (1958) and Emsley (1989) referred to in Beardsmore and Cull (2001))

Element	^{238}U	^{232}Th	^{40}K
Half-life (Ga)	4.47	14.1	1.28
Relative γ-activity	3 600	1 300	1

proportions of the radioactive elements in the sediments and should be similar over any region containing sediment derived from the same source.

An alternative empirical relationship was published by Bücker and Rybach (1996), which does not require information on the relative proportions of the radioactive elements in the rock. It relates bulk heat generation rate A to total γ-ray emission GR recorded by borehole tools as

$$A\left[\mu W\,m^{-3}\right] = 0.0158(GR[API] - 0.8). \quad (6)$$

This relation has been successfully tested on a number of data sets where both NGS and GR logs were available. As the relations in Equations 3 and 6 are based on measurements mostly on igneous rocks, caution is required when applying them to sedimentary rocks.

Measuring techniques

A number of different analytical methods are available for determining the concentrations of uranium, thorium, and potassium. An inter-laboratory comparison of different methods on selected materials yielded consistent results within a few percent (Rybach, 1988). However, among all other methods, γ-ray spectroscopy is the only one which enables determining all three concentrations simultaneously. It implies secular equilibrium within the uranium and thorium decay series and a constant ^{40}K/K ratio, both of which are satisfied for most rock types (Rybach, 1988). The method is based on counting the number of decays per energy channel. Commercial spectrometers will offer, for instance, 2 048 channels for measuring decays in the energy range of 0–3 000 keV, yielding a resolution, in this particular case, of 1.46 keV per channel. In performing measurements, it is important to realize that radioactive decay is a stochastic process. Therefore, decays need to be recorded over a sufficiently long time. In view of the minute U-, Th-, and K-concentrations, and depending on crystal size and sensitivity, this may imply recording times on the order of several hours. Additionally, a possible sampling bias needs to be considered when selecting samples for measurements.

The interpretation of γ-ray spectroscopy relates the elemental concentrations C linearly to the corresponding counting rates N in the energy channel x by the corresponding coefficient h(x): N(x) = h(x) C. Thus, the total count rate due to the decay of uranium, thorium, and potassium in the energy channel x is given by:

$$N(x) = h_U(x)\,C_U(x) + h_{Th}(x)\,C_{Th}(x) + h_K(x)\,C_K. \quad (7)$$

The coefficients h depend mainly on the characteristic spectra of the radioactive elements, but also on the geometry of the sample, the geometric configuration of the measurement, and on the efficiency and sensitivity of the detectors. Thus, γ-spectroscopic measurements require calibration based on samples of known elemental concentrations. This enables accounting for the influence of the geometry of the detectors. Additionally, possibly variable sample geometry needs to be addressed separately. Three different effects require attention (Adams and Gasparini, 1970): (1) *Variation of the continuous background spectrum* may occur if the background radiation was determined with an empty sample chamber in which no absorption of the natural background radiation occurs. This can be avoided by measuring the background radiation using a nonradioactive sample of equal size and similar density, such as water-filled plastic tubes; (2) *Deformation of the spectrum* may occur if a larger sample volume increases the likelihood for internal Compton scattering. Thus, larger samples are characterized by a decreased peak size and an increased level of the continuous spectrum; (3) *Radiation self-absorption* of a fraction of the γ-radiation emitted within the sample reduces the γ-radiation recorded at the detector. Self-absorption increases with sample thickness d and reduces the count rate N by a factor

$$F = \left(1 - e^{-\mu d}\right)/\mu\,d, \quad (8)$$

where μ is the absorption coefficient (Watt and Ramsden, 1964). The joint effect of spectrum deformation and self-absorption is quantified by measurements on potash salt standards. While the absolute ^{40}K-concentrations are unknown, the relative variations reveal the influence of variable sample diameters. Theoretically, the total activity is proportional to the sample cross section, that is, to the square of the sample diameter. In plotting the count rate versus sample diameter one finds, however, a smaller exponent of only 1.5. Thus, the remainder of the radiation is either absorbed within the sample or transferred to other energy bands.

The energy peaks in the spectrum are analyzed for finding the elemental concentrations. Theory predicts a Gaussian normal distribution for the energy peaks. The count rate N at energy x in the local neighborhood of such a Gauß peak comprises a contribution of the Gauß peak itself, G(x), and a linear trend ax + b, describing the background radiation and the Compton scattering (Adams and Gasparini, 1970):

$$N(x) = G(x) + ax + b. \quad (9)$$

Such linear trends need to be removed prior to further interpretation, for example, by linear regression. Then,

the area under the peak above the base line corresponds to the sought total energy. In practice, the count rates within this area are summed up, or a Gaussian distribution function

$$G(x) = A\, e^{-(x-x_0)^2/2\sigma^2} \quad (10)$$

is fitted to the corrected peak, where x_0 and σ are mean energy and standard deviation of the peak, and A is a calibration constant. For a Gauß distribution, the peak's full width at half maximum (FWHM) is related to σ by

$$FWHM = 2\sigma\sqrt{\log 2}. \quad (11)$$

The summed up count rates analyzed this way then need to be calibrated by corresponding measurements on samples of known concentrations. Finally, this yields the concentrations of the radiogenic isotopes of uranium, thorium, and potassium.

Heat generation and geoneutrinos

Currently, general hypotheses on possible distributions of radiogenic heat generating elements in the mantle and core of the Earth are mainly based on analogies with chondritic material and direct evidence is lacking. However, as radiogenic sources produce heat, electron antineutrinos $\bar{\nu}$ are emitted in β^--decay of nuclei in the decay chains of ^{238}U, ^{232}Th, and ^{40}K. Thus, the heat produced in nuclear decay is directly related to the antineutrino flux, as shown in Table 4 (Fiorentini et al., 2003).

Neutrinos are elementary particles that travel close to the speed of light, are electrically neutral, and can pass through ordinary matter almost undisturbed. They have a very small, but nonzero mass. Therefore, neutrinos are extremely difficult to detect. As the emission of geoneutrinos is coupled to radiogenic heat generation, this enables, at least theoretically, a geoneutrino tomography of radiogenic heat distribution within the Earth. But this concept is as difficult to implement as it is intriguing. First, there is a huge flux of cosmic neutrinos which traverses the Earth from which geoneutrinos need to be distinguished: The neutrino flux at the Earth's surface from the sun alone is estimated at about 60×10^9 s^{-1} cm^{-2} (e.g., Bahcall, 1969). The flux of geoneutrinos is still on the order of $(60–15) \times 10^6$ s^{-1} cm^{-2} (Fiorentini et al., 2003).

In spite of this immense flux density, the vanishing mass of neutrinos requires exceptional sensors for their detection: (1) At the Gran Sasso observatory in Italy, the detector is a tank of 100 t of liquid gallium trichloride ($GaCl_3$) containing the equivalent of 30 t of gallium at a subsurface laboratory. Here, antineutrinos are observed indirectly through the conversion of a gallium nucleus into a germanium isotope by neutrino capture. For all the high neutrino flux through the tank, an interaction of a neutrino and a gallium nucleus occurs only every 35 h. (2) In the *Kamioka Liquid Scintillator Antineutrino Detector* (*KamLAND*) experiment (Araki et al., 2005), the detector comprises of a 13 m diameter nylon balloon filled with 1 kt of liquid scintillator consisting of mineral oil, benzene and fluorescent chemicals. KamLAND is an experiment at the Kamioka Observatory, an underground neutrino observatory near Toyama, Japan, built to detect electron antineutrinos. The experiment is located in the old Kamiokande cavity in a horizontal mine drift in the Japanese Alps. The detector is housed in an 18 m diameter stainless steel spherical vessel with 1 879 photomultiplier tubes mounted on the inner surface. KamLAND is the first detector to conduct an investigation on geoneutrinos and may yield important geophysical information. It has the sensitivity to detect electron antineutrinos produced by the decay of ^{238}U and ^{232}Th within the Earth. Earth composition models suggest that the radiogenic power from these isotope decays is 16 TW, approximately half of the total measured heat dissipation rate from the Earth.

A successful antineutrino tomography requires global sampling and distinguishing between cosmic neutrinos traversing the Earth and geoneutrinos generated within the Earth. Recently, a project was proposed to deploy considerably smaller neutrino detection sensors in a network of boreholes around the globe (de Meijer et al., 2006). Time will show whether the significant technical challenges associated with the construction of smaller sensors than those in the experiments at the Gran Sasso and Kamioka observatories can be overcome and a road opened for an antineutrino tomography of the Earth.

Radiogenic Heat Production of Rocks, Table 4 Maximum antineutrino energy and heat production rates in natural decay processes (de Meijer et al., 2006; Fiorentini et al., 2005)

Decay reaction	E_{max} (MeV)	Heat (μW kg^{-1})
$^{238}U \rightarrow {}^{206}Pb + 8\,{}^4He + 6\,e + 6\bar{\nu}$	3.25	95
$^{232}Th \rightarrow {}^{208}Pb + 6\,{}^4He + 4\,e + 4\bar{\nu}$	2.25	27
$^{40}K \rightarrow {}^{40}Ca + e + \bar{\nu}$	1.31	0.0036

Summary

The heat produced in the radioactive decay of the unstable isotopes of uranium (^{238}U; ^{235}U), thorium (^{232}Th), and potassium (^{40}K) is the largest internal heat source of the Earth. During radioactive decay, mass is converted into energy. Except for the tiny amount associated with the antineutrino and neutrinos generated in β^-- and β^+-decay or electron capture, respectively, all of this energy ends up as heat. The annual production of radiogenic heat in the Earth equals 8.6×10^{20} J, which is more than twice the global production of primary energy in the year 2000. The distribution of radiogenic isotopes in the Earth controls to a large extent the thermal regime of the Earth. Unfortunately, this distribution is known only with great uncertainty. Recently, it has become possible to detect geoneutrinos, that is, antineutrinos emitted during

radioactive decay of unstable isotopes, in large detectors. With decreased size and improved accuracy of detectors together with directional resolution power, an antineutrino tomography of the Earth appears to become possible. Ideally, this would enable locating and quantifying the distribution of unstable isotopes in the Earth, thus helping to resolve a number of open questions with regard to the state and evolution of the Earth's thermal regime.

Bibliography

Adams, J. S., and Gasparini, P., 1970. *Gamma-Ray Spectrometry of Rocks*. Amsterdam: Elsevier.

Adams, J. A. S., and Weaver, C. E., 1958. Thorium to uranium ratios as indicators of sedimentary processes: examples of the concept of geochemical facies. *Bulletin of the American Association of Petroleum Geologists*, **42**, 387–430.

Araki, T., et al., 2005. Experimental investigation of geologically produced antineutrinos with KamLAND. *Nature*, **436**, 499–503.

Bahcall, J. N., 1969. Neutrinos from the Sun. *Scientific American*, **221**(1), 28–37.

Beardsmore, G. R., and Cull, J. P., 2001. *Crustal Heat Flow*. Cambridge: Cambridge University Press.

Birch, F., 1954. Heat from radioactivity. In Faul, H. (ed.), *Nuclear Geology*. New York: Wiley, pp. 148–174.

Bücker, C., and Rybach, L., 1996. A simple method to determine heat production from gamma-ray logs. *Marine and Petroleum Geology*, **13**, 373–375.

Canup, R. M., and Asphaug, E., 2001. Origin of the Moon in a giant impact near the end of the Earth's formation. *Nature*, **412**, 708–712.

Canup, R. M., and Righter, K. (eds.), 2000. *Origin of the Earth and Moon*. Tucson: University of Arizona Press.

Čermák, V., and Rybach, L., 1982. Radioactive heat generation in rocks. In Angenheister, G. (ed.), *Landolt-Börnstein, Group V: Geophysics and Space Research*. Heidelberg/Berlin: Springer. Physical Properties of Rocks, Subvol. A, Vol. l, pp. 353–371.

Clauser, C., 2009. Heat transport processes in the Earth's crust. *Surveys in Geophysics*, **30**, 163–191, doi:10.1007/s10712-009-9058-2.

De Meijer, R. J., Smit, F. D., Brooks, F. D., Fearick, R. W., Wörtche, H. J., and Mantovani, F., 2006. Towards Earth AntineutRino TomograpHy (EARTH). *Earth, Moon and Planets*, **99**(1–4), 193–206.

Emsley, J., 1989. *The Elements*. Oxford: Clarendon.

Fiorentini, G., Mantovani, F., and Ricci, B., 2003. Neutrinos and energetics of the Earth. *Physics Letters B*, **557**, 139–146.

Fiorentini, G., Lissia, M., Mantovani, F., and Vanucci, R., 2005. Geo-neutrinos: a new probe of Earth's interior. *Earth and Planetary Science Letters B*, **557**, 139–146.

Hamza, V. M., and Beck, A. E., 1972. Terrestrial heat flow, the neutrino problem, and a possible energy source in the core. *Nature*, **240**(5380), 343–344.

IEA, 2008. *World Energy Outlook 2008*. Paris: International Energy Agency (IEA). http://www.iea.org/textbase/nppdf/free/2008/weo2008.pdf. Retrieved 10 July 2010.

Jaupart, C., Labrosse, S., and Mareschal, J.-C., 2007. Temperatures, heat and energy in the mantle of the Earth. In Bercovici, D. (ed.), *Mantle Dynamics – Treatise on Geophysics*. Amsterdam: Elsevier, Vol. 7, pp. 253–303.

Jessop, A. M., 1990. *Thermal Geophysics*. Amsterdam: Elsevier.

McDonough, W. F., and Sun, S.-S., 1995. The composition of the Earth. *Chemical Geology*, **120**, 223–253.

Melosh, H. J., 1990. Giant impacts and the thermal state of the early Earth. In Newsom, H. E., and Jones, J. H. (eds.), *Origin of the Earth*. New York: Oxford University Press, pp. 69–83.

Rybach, L., 1988. Determination of heat production rate. In Hänel, R., Rybach, L., and Stegena, L. (eds.), *Handbook of Terrestrial Heat Flow Density Determination*. Dordrecht: Kluwer, pp. 125–142.

Stacey, F. D., and Davis, P. M., 2008. *Physics of the Earth*, 4th edn. Cambridge: Cambridge University Press.

Vacquier, V., 1991. The origin of terrestrial heat flow. *Geophysical Journal International*, **106**(1), 199–202.

Vacquier, V., 1992. Corrigendum to 'The origin of terrestrial heat flow'. *Geophysical Journal International*, **111**(3), 637–638.

Van Schmus, W. R., 1984. Radioactivity properties of minerals and rocks. In Carmichael, R. S. (ed.), *Handbook of Physical Properties of Rocks*. Boca Raton: CRC Press, Vol. III, pp. 281–293.

Van Schmus, W. R., 1995. Natural radioactivity in crust and mantle. In Ahrens, T. J. (ed.), *Global Earth Physics – A Handbook of Physical Constants. AGU Reference Shelf 1*. Washington, DC: American Geophysical Union, pp. 283–291.

Watt, D. E., and Ramsden, D., 1964. *High Sensitivity Counting Techniques*. London: Pergamon.

Cross-references

Energy Budget of the Earth
Heat Flow Measurements, Continental
Heat Flow, Continental
Heat Flow, Seafloor: Methods and Observations
Radioactivity in Earth's Core
Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity

REMANENT MAGNETISM

Laurie Brown[1], Suzanne McEnroe[2]
[1]Department of Geosciences, University of Massachusetts, Amherst, MA, USA
[2]Norwegian Geological Survey, Trondheim, Norway

Definition

Remanent magnetization. Permanent magnetization held in rocks and other earth materials, commonly, though not always, dating from cooling to below the Curie temperature (Tc) or Néel temperature (Tn) of the magnetic minerals in the rock, or the time of formation of the rock unit.

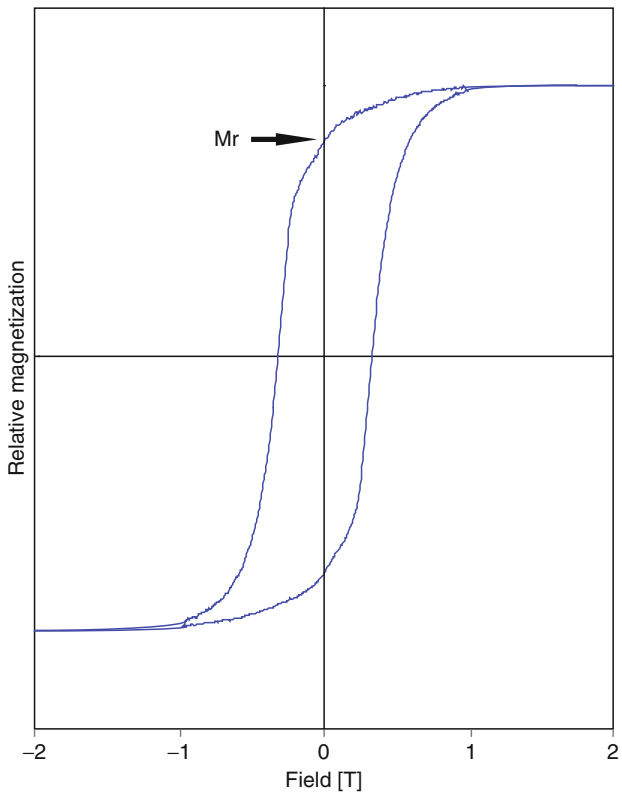
Introduction

The fact that earth materials, particularly rocks, possess a permanent magnetization has been known for some time. Early observations of rocks producing large anomalies were commonly ascribed to lightning strikes. By the mid-nineteenth century remanence was accepted as a property of some rocks, and by the early twentieth century crude measurements of the direction of magnetization were possible. During the second half of the twentieth century, the measurement of remanent magnetization (RM) became a well-established research field (see entries under *Paleomagnetism, Principles*). RM is a vector quantity,

having both direction, usually related to the direction of the Earth's field at the time of origin, and intensity, related to both mineral phase and the external field. Descriptions of remanent magnetization, associated paleomagnetic techniques, and results are detailed in standard paleomagnetic textbooks such as Butler (1992), McElhinny and McFadden (2000), and Tauxe (2010). More detailed discussion of the theory and development of remanent magnetization can be found in rock magnetism books, including those by Nagata (1961), Stacey and Banerjee (1974), O'Reilly (1984), and Dunlop and Özdemir (1997).

Source of RM

Remanence is a property unique to a small group of minerals possessing ferromagnetic or antiferromagnetic properties, mostly iron oxides and iron sulfides. Remanence is observable in hysteresis behavior, when, after saturation, the applied field is reversed and decreased to zero. If the magnetization does not also return to zero, but retains a measurable quantity, a remanent magnetization is present (Figure 1). This characteristic is temperature dependent, with only some of the mineral phases possessing remanence at earth surface temperatures.



Remanent Magnetism, Figure 1 Hysteresis loop of hematite-dominated sample of the El Laco deposit, northern Chile. Horizontal axis is the external field, in Tesla; vertical axis is relative magnetization. Mr indicates the remanent magnetization left after the sample is saturated and returned to a zero field.

The most common and important magnetic minerals on Earth belong to the cubic oxide series magnetite ($Fe^{2+}Fe_2^{3+}O_4$) – ulvöspinel ($Fe_2^{2+}Ti^{4+}O_4$), and the rhombohedral oxide series hematite ($Fe_2^{3+}O_3$) – ilmenite ($Fe^{2+}Ti^{4+}O_3$). Both series have important solid solution versus temperature relationships, as well as magnetic complexities of key interest in the rock magnetism. Additionally, both series show remanence properties that vary with composition. Furthermore, the oxide bulk compositions in the vicinity of equilibrium tie lines between ilmenite and magnetite are found in a majority of igneous and metamorphic rocks, especially basalts. In these assemblages, the slopes of the tie lines are a function temperature, whereas tie line positions are a function of the oxygen fugacity of equilibration according to the oxygen thermobarometer (Buddington and Lindsley, 1964).

End-member hematite is antiferromagnetically ordered with equal and opposite magnetic moments in alternate cation layers. However, at room temperature, a weak ferromagnetic moment is produced because the alternate moments are ~0.13° away from being perfectly antiparallel (Dzialoshinskii, 1957). This effect is commonly referred to as "spin-canting" and such hematite is termed "canted-antiferromagnetic." Many oxidized sediments contain hematite, and carry this weak, but very stable magnetization, which is mainly due to its high coercivity.

The temperature and composition phase relations of the hematite-ilmenite series are a result of a complex interaction of Fe-Ti ordering and magnetic ordering at high temperature. Slow cooling to intermediate to low temperatures produces exsolution of discrete ilmenite and hematite (Burton, 1991; Harrison et al., 2000; Ghiorso and Evans, 2008). Exsolved grains contain two phases, a host (hematite or ilmenite) and lamellae of the second phase. Ilmeno-hematite (hematite with ilmenite exsolution), or hemo-ilmenite (ilmenite with hematite exsolution) commonly have a strong, and very stable magnetization, referred to as "lamellar magnetism." Lamellar magnetism is due to a defect magnetic moment at the interface (or contact layer) between the two phases (Robinson et al., 2002, 2004) where a ferrimagnetic substructure is present because of the different quantity of Fe^{2+} and Fe^{3+} ions in contact Fe layers, as compared to standard Fe^{3+} layers in adjacent hematite.

Intermediate members of this series rapidly cooled from high temperature produce intermediate metastable phases that are ferrimagnetic and show complicated magnetic properties near and below room temperature (Ishikawa and Akimoto, 1957; Burton et al., 2008; Robinson et al., 2010) with some compositions with varied Fe-Ti ordering showing magnetic self-reversal (Harrison et al., 2005; Ishikawa and Syono, 1963).

Natural remanent magnetization (NRM)

The magnetization held by a rock sample prior to subjection to any laboratory procedures is referred to as Natural Remanent Magnetization (NRM). It is the sum total of all

remanences held in the rock and may include primary as well as secondary remanences arising from different processes. The NRM is differentiated from induced magnetization, which a rock may acquire as it sits in an external field, and loses when the field is removed. Induced magnetization is dependent on the magnetic susceptibility of the material as well as the strength of the external field. NRM is traditionally described by the manner of acquisition of the remanences involved, leading to a number of both natural and laboratory magnetizations, as described below.

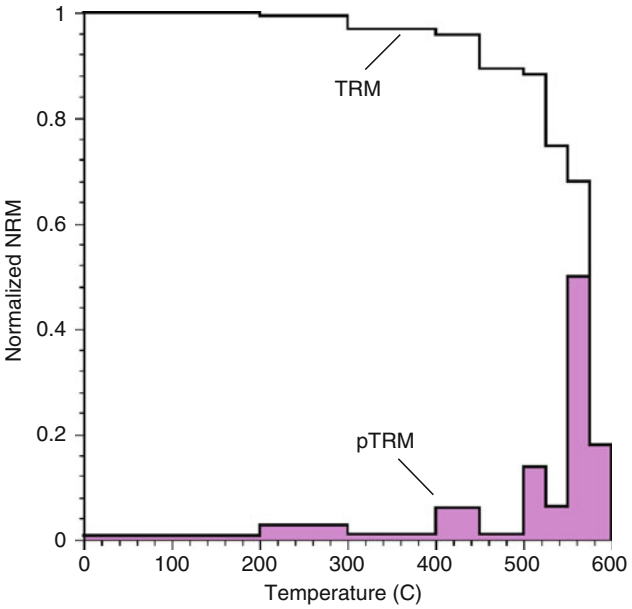
Thermoremanent magnetization (TRM)

Igneous rocks, cooling down from a molten state through the Curie or Néel temperature (Tc or Tn), obtain a thermoremanent magnetization (TRM) parallel to the ambient field. For magnetite this temperature is 580°C while for hematite it is 675°C. At these temperatures the spontaneous nature of the ferromagnetic material occurs and remanence is blocked just below this temperature. The material must cool slightly further to the blocking temperature (Tb) before this magnetization becomes locked in with a relaxation time equivalent to geologic time spans. The direction of remanence, so important to the whole field of paleomagnetism, becomes permanent at this point, with a net direction parallel to the ambient field. The intensity of the magnetization is influenced by the size of the external field, and by the phase, composition, and grain size of the ferromagnetic minerals present.

The TRM preserved in igneous rocks is a summation of the remanences of all the ferromagnetic grains present. This means there may be a range of Curie temperatures and related blocking temperatures, giving rise to partial thermal remanence (pTRM). As a simple case, if a rock had both pure hematite and pure magnetite grains present, the hematite grains would lock in the field slightly below 675°C while the magnetite grains would provide additional remanence once the rock cooled to below 580°C. Commonly, a combination of oxide minerals of varying composition provide for remanence acquired at a range of temperatures. In Figure 2, a plot of pTRM is shown for a basalt with low Ti-magnetite. Metamorphic rocks heated to above their Curie temperatures will have the remanence reset as the material cools down. High grade metamorphic rocks, such as granulite- and amphibolite facies rocks, heated to above 600–700°C will have all the existing magnetic minerals remagnetized in the external field at the time of cooling, whereas rocks of lower metamorphic grade will have only minerals with Tc below the peak metamorphic temperature completely reset. Néel (1955) introduced the theory of thermoremanence; Dunlop (1977) and Dunlop and Özdemir (1997) provide detailed treatment of the theory.

Detrital remanent magnetization (DRM)

Sedimentary rocks obtain a remanent magnetization in a very different way. Here the detrital grains settling in a water column are influenced by the external magnetic field producing a depositional remanent magnetization (DRM). These grains are also influenced by procedures in the sedimentary environment imparting a post-depositional remanence (pDRM). The general term of detrital remanent magnetization includes both processes of depositional as well as post-depositional magnetization.



Remanent Magnetism, Figure 2 Thermoremanent magnetization plot from a Tertiary basalt, Meseta Lago Buenos Aires, Argentina. *Upper curve* labeled TRM represents the amount of magnetization left after increasing heatings up to 600°C, with nearly 75% of the remanence being removed between 550°C and 600°C. *Lower curve* (pTRM) indicates the percentage of TRM residing within certain blocking temperatures.

When eroded ferromagnetic grains enter the sedimentary regime, they retain their magnetic moments even though the original orientation of that moment is lost. These grains, as they settle through a water column, respond to the external field, and in a mechanical response, become aligned with the influencing field over a very short time period. As redeposition experiments in the laboratory have shown, this theory is not so simple as projected (Verosub, 1977). There are a number of changes that can occur once the grain is deposited, but before magnetization is locked in. These effects lead to a pDRM, or post-depositional remanent magnetization, a prevalent but complicated process (Shcherbakov and Shcherbakova, 1983). Common effects are the dewatering of the sediment, compaction of the grains, and bioturbation, all of which allow for ferromagnetic particles, particularly small ones, to realign. Magnetization is not considered final until the grains are no longer able to move about or realign, occurring at a depth referred to as the lock-in depth. This gives grain size an importance in DRM, with smaller grains more susceptible to realignment, but also

yielding stronger and more stable remanence. To this end fine-grained sediments, as mudstones and siltstones, retain a remanence better than coarser-grained rocks such as sandstones or conglomerates.

The notion of individual grains settling in a water body has recently been investigated (Tauxe et al., 2006) and found to be more complicated. In marine environments in particular, fine-grained sediments do not sink as discrete particles, but rather flocculate with coexisting nonmagnetic particles. The entire floc needs to be aligned with the external field, leading to variations in remanence, especially with respect to intensity, from that predicted by theory and redeposition experiments.

Chemical remanent magnetization (CRM)

Commonly rocks undergo chemical changes after formation, usually at temperatures below the Curie temperature of the minerals involved. These changes can be alterations to existing magnetic minerals and/or the creation of new iron oxides within the rock unit. These changes can result in new or altered remanences; this process is commonly referred to chemical remanent magnetization (CRM).

A classic example of CRM is "red beds" or sedimentary rocks with strong reddish hues due to the ubiquitous presence of hematite. Such rocks may have a DRM from the time of formation, due to the presence of detrital grains of magnetite and/or hematite. After deposition, changes brought on by differing environments, a range of aqueous solutions, or variations in oxidation state may result in weathering of existing ferromagnetic minerals, or in the deposition by precipitation of new ferromagnetic minerals. Existing magnetic minerals may break down to weaker or nonmagnetic minerals, such as hematite altering to goethite or pyrrhotite to pyrite. Ferromagnetic minerals may also alter to other magnetic minerals with different remanence characteristics. One example is the oxidation of magnetite to maghemite. Another alteration common is that of nonmagnetic grains to magnetic ones, such as the case of Fe-bearing silicate grains altering to discrete hematite grains. New ferromagnetic minerals develop, such as the formation of hematite cements and coatings, which is pervasive in red beds develops.

CRM can also be produced during progressive metamorphism, for example, when Fe-silicates breakdown to produce magnetite, in hematite-ilmenite solid solutions, or when Fe-bearing silicates react with pyrite (nonmagnetic FeS_2) to produce pyrrhotite (magnetic FeS). Lamellar magnetism mentioned earlier is a CRM because it is produced during the chemical reaction of exsolution and not at the Curie temperature of the minerals; with slow cooling this can occur hundreds of degrees below the Tc.

CRM is also recognized in igneous rocks, such as in the formation of secondary hematite in cracks and factures in basalts or around the rims of discrete magnetite grains. In this case, the CRM is often hard to distinguish from the original TRM, as the CRM is quite stable and many of the magnetic properties of the two remanences are similar.

Viscous remanent magnetization (VRM)

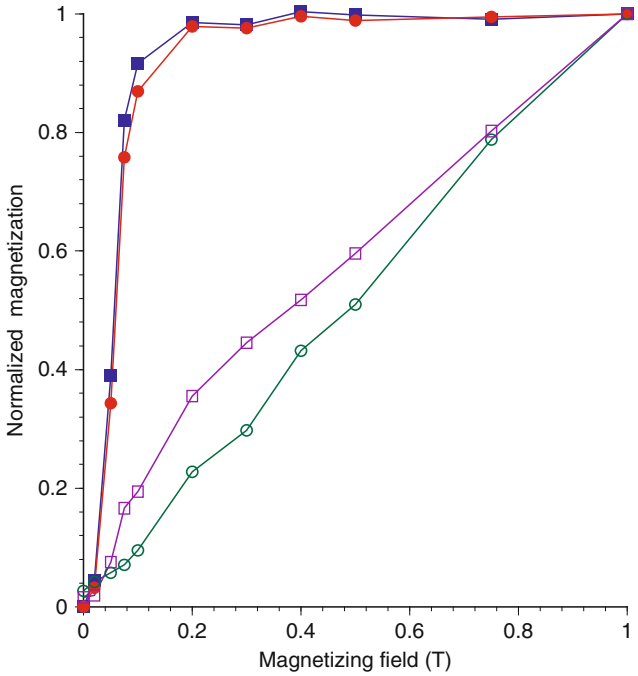
At room temperature, in low ambient fields, rocks may acquire a component of magnetization in the direction of the "present" field. This remanence acquired over long time periods in low fields, such as the Earth's field, is referred to as viscous remanent magnetization (VRM). It is generally assumed that the strength of the VRM is related to the log of the time involved, although this is a gross simplification of the situation. The importance of other factors, including grain size, initial magnetic state, and the angle of the ambient field relative to the original remanence vector have recently been investigated and shown to also affect the strength of VRM (Yu and Tauxe, 2006). Although VRM is ubiquitous in geologic examples and is found associated with many different rocks types, in many cases it is small and can easily be removed by demagnetization techniques. Due to its soft nature, VRM is often referred to as an "overprint" commonly representing the present magnetic field and must be removed prior to investigating the primary remanence.

In natural situations where temperatures may be elevated for a considerable time, as in low-grade metamorphism, or in young oceanic lithosphere, there is commonly a thermal viscous remanent magnetization (TVRM) acquired. Here, increased temperature over extended time periods enhance the viscous remanence and produce a magnetization that may eventually replace the original NRM. Pullaiah et al. (1975) have investigated such magnetization changes over a range of temperatures and timescales, with additional discussion provided by Dunlop and Özdemir (1997).

Isothermal remanent magnetization (IRM)

The presence of a large magnetic field over a short time period may also produce a remanent magnetization, referred to as isothermal remanent magnetization (IRM). An IRM can be imparted in the laboratory and used to help identify properties of magnetic materials. Samples are subjected to increasingly large external fields, usually to 1 T or larger, and the resulting magnetization recorded. Differences in composition, concentration, grain size, and saturation state effect the acquisition of IRM. For example, magnetite-bearing samples will saturate in fields of 0.1–0.3 mT, while hematite-dominated rocks will require fields larger than 1 T to reach saturation (Figure 3).

In the natural environment, a typically cited example of an IRM is the magnetization gained in a rock after being struck by lightning. The large magnetization produced by the electric currents in lightning imparts an intense, but chaotic remanence on surface rocks. First studied by Cox (1961), IRM is common in regions subject to intense thunderstorms, such as the desert southwest of the North America. In some cases, the effect of the IRM can be removed by demagnetization; in other samples, the



Remanent Magnetism, Figure 3 Normalized IRM acquisition in fields up to 1 T; *closed symbols* are Jurassic diabase dikes with magnetite; *open circles* are Jurassic mudstones with hematite.

magnetization has been totally reset, giving rise to paleomagnetic sites with random directions (Tauxe et al., 2003).

However, this type of IRM is commonly accompanied by high heat; therefore, the term is slightly misused.

Applications of remanent magnetization

RM is used extensively in the disciplines of magnetism, including paleomagnetism, rock magnetism, and the study of magnetic anomalies. The study of paleomagnetism centers on the identification and description of the remanent field held in rocks. Results from the measurement of RM are used widely in the study of plate tectonics, structure, regional tectonics, stratigraphy, volcanology, and archeology. Positions of the earth's magnetic field at specific times in the geologic past provide us invaluable information on the position of continental bodies in the past, the interrelationship between different continents, and the paths they have followed in subsequent motions across the earth's surface. RM can provide age information by comparison to established reversal timescales, as well as relative ages within a region. Regional tectonics, local rotations, and timing of folding can all be investigated using the remanence recorded in suitable rock units. Stratigraphic studies, especially in the wealth of deep-sea sediment cores from international drilling projects, lake sediments, or in thick sedimentary sequences on land, can be established and compared using the recorded remanence.

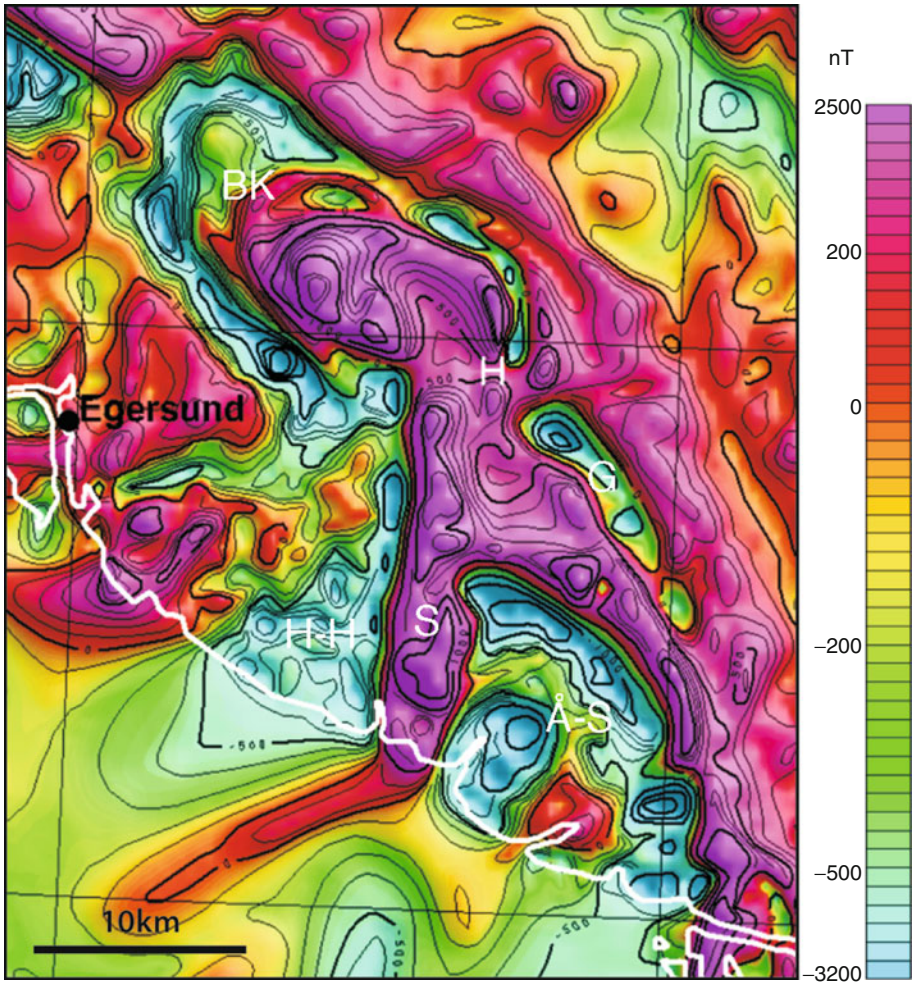
Experimental studies in rock magnetism make excellent use of the remanent magnetization properties of rocks and minerals. Studies of hysteresis behavior, magnetization at high and low temperatures, and responses to acquisition and removal of laboratory remanences all provide important information on oxide composition, grain size, and magnetic properties of a wide range of natural and synthetic materials.

Remanent magnetism plays an important part in the interpretation of magnetic anomalies. Well established in the 1960s in the study of marine magnetic anomalies from the seafloor, remanent polarity of extruded basalts played a pivotal role in the interpretation of observed anomalies (Vine and Matthews, 1963; Heirtzler et al., 1963). The use of the polarity sequence preserved in the ocean crust allows for dating of the seafloor as well as calculations of spreading rates and investigations of plate histories. Anomalies on land show considerable complexities and many are assumed to be entirely related to induced fields due to the present magnetic field. But, in many cases, there is a significant component of remanent magnetization interacting with the induced field, and depending on the remanent direction and intensity, enhancing or at times annihilating the induced field (McEnroe et al., 2009). An excellent case of remanent magnetization producing strong aeromagnetic anomalies comes from the Rogaland Igneous Complex in southern Norway (McEnroe et al., 2008), where anorthosites and a layered intrusion have intense remanence roughly antiparallel to the present field, producing large negative anomalies (Figure 4).

Exploration for natural resources by aeromagnetic surveys has been widely used since the 1950s. The classic work by Balsley and Buddington (1958) and later by McEnroe and Brown (2000) in the oxide rich area of the Adirondack Mountains of New York required a detailed knowledge of the RM for accurate interpretation of the data. Understanding the contribution of RM to magnetic anomalies will be crucial for future exploration as the accuracy and resolution of these surveys increase. Magnetic surveys will also be at the forefront for geological mapping of Earth and other planets.

Summary

Remanent magnetization is a permanent magnetization residing in earth materials that is gained when the material is formed or altered in the presence of the Earth's magnetic field. Measurable RM is founding in most rocks, ranging from metamorphic gneisses to lava flows to marine and lake sediments. The remanence is carried by iron oxides and sulfides, but magnetite and hematite are by far the most common and the most important remanence carriers. RM is categorized and described by various methods of acquisition; TRM for remanence produced as minerals cool through their Curie temperatures, DRM for the remanence obtained as detrital grains of magnetic minerals align during the deposition process, and CRM for the process where remanence develops from chemical changes in



Remanent Magnetism, Figure 4 Aeromagnetic anomaly map from a fixed-wing survey over the Rogaland region in southwestern Norway. The magnetic total field was reduced to anomaly values by subtracting the International Geomagnetic Reference Field from the total field of 1965. Color shades: *pink*, large positive magnetic anomalies; *blue*, strong negative anomalies; *yellow and green*, intermediate values. *Letters* refer to specific bodies within the Rogaland Igneous Complex (Redrawn from McEnroe et al., 2001).

the rock. RM forms the basis for research in paleomagnetism, where the direction, intensity, and age of the remanence are determined and interpreted. It is also useful in experimental studies of rock magnetism, and in the investigation of anomalies in global, regional, and local magnetic field measurements.

Bibliography

Balsley, J. R., and Buddington, A. F., 1958. Iron-titanium oxides minerals, rocks and aeromagnetic anomalies of the Adirondack area, New York. *Economic Geology*, **53**, 777–805.

Buddington, A. F., and Lindsley, D. H., 1964. Iron-titanium oxide minerals and synthetic equivalents. *Journal of Petrology*, **5**, 310–357.

Burton, B. P., 1991. Interplay of chemical and magnetic ordering. In Lindsley, D. H. (ed.), *Oxide Minerals*. Washington: Mineralogical Society of America, pp. 303–321.

Burton, B. P., Robinson, P., McEnroe, S. A., Fabian, K., and Boffa Ballaran, T., 2008. A low-temperature phase diagram for ilmenite-rich compositions in the system Fe_2O_3-$FeTiO_3$. *American Mineralogist*, **93**, 1260–1272.

Butler, R. F., 1992. *Paleomagnetism: Magnetic Domains to Geologic Terranes*. Boston: Blackwell Scientific. On-line version: http://www.geo.arizona.edu/Paleomag/book/

Cox, A., 1961. Anomalous remanent magnetization of basalt. *United States Geological Survey Bulletin*, **1083E**, 131–160.

Dunlop, D. J. (ed.), 1977. *Origin of Thermoremanent Magnetization*. Tokyo: Center for Academic Publications Japan.

Dunlop, D. J., and Özdemir, Ö., 1997. *Rock Magnetism – Fundamentals and Frontiers*. Cambridge: Cambridge University Press.

Dzialoshinskii, I. E., 1957. Thermodynamic theory of "weak" ferromagnetism in antiferromagnetic substances. *Soviet Journal of Experimental and Theoretical Physics*, **5**, 1259–1272.

Ghiorso, M. S., and Evans, B., 2008. Thermodynamics of rhombohedral oxide solid solutions and a revision of the Fe-Ti two-oxide geothermometer and oxygen-barometer. *American Journal of Science*, **308**, 957–1039.

Harrison, R. J., Becker, U., and Redfern, S. A. T., 2000. Thermodynamics of the *R* to *Rc* phase transition in the ilmenite-hematite solid solution. *American Mineralogist*, **85**, 1694–1705.
Harrison, R. J., Kasama, T., White, T., Simpson, E. T., and Dunn-Borkowski, R. E., 2005. Origin of self-reversed thermoremanent magnetization. *Physical Review Letters*, **95**, doi:10.1103/PhysRevLett95.268501.
Heirtzler, J. R., Dickson, G. O., Herron, E. M., Pitman, W. C., III, and Le Pichon, X., 1968. Marine magnetic anomalies, geomagnetic field reversals, and motions of the ocean floor and continents. *Journal of Geophysical Research*, **73**, 2119–2136.
Ishikawa, Y., and Akimoto, S., 1957. Magnetic properties of the $FeTiO_3$-Fe_2O_3 solid solution series. *Journal of the Physics Society of Japan*, **12**, 1083–1098.
Ishikawa, Y., and Syono, Y., 1963. Order-disorder transformation and reverse thermoremanent magnetization in the $FeTiO_3$-Fe_2O_3 system. *Journal of Physics and Chemistry of Solids*, **24**, 517–528.
McElhinny, M. W., and McFadden, P. L., 2000. *Paleomagnetism: Continents and Oceans*. San Diego: Academic.
McEnroe, S. A., and Brown, L. L., 2000. A closer look at remanence-dominated anomalies: Rock-magnetic properties and magnetic mineralogy of the Russell Belt microcline-sillimanite gneisses, Northwest Adirondacks Mountains, New York. *Journal of Geophysical Research*, **105**, 16,437–16,456.
McEnroe, S. A., Robinson, P., and Panish, P., 2001. Aeromagnetic anomalies, magnetic petrology and rock magnetism of hemo-ilmenite- and magnetite-rich cumulates from the Sokndal Region, South Rogaland, Norway. *American Mineralogist*, **86**, 1447–1468.
McEnroe, S. A., Brown, L. L., and Robinson, P., 2008. Remanent and induced magnetic anomalies over a layered intrusion: Effects from crystal fractionation and magma recharge. *Tectonophysics*, doi:10.1016/j.tecto.2008.11.021.
McEnroe, S. A., Fabian, K., Robinson, P., Gaina, C., and Brown, L. L., 2009. Crustal magnetism, lamellar magnetism and rocks that remember. *Elements*, doi:10.2113/gselements.5.4.241.
Nagata, T., 1961. *Rock Magnetism*. Tokyo: Maruzen Limited.
Néel, L., 1955. Some theoretical aspects of rock magnetism. *Advances in Physics*, **4**, 191–243.
O'Reilly, W., 1984. *Rock and Mineral Magnetism*. Glasgow: Blackie.
Pullaiah, G. E., Irving, E., Buchan, K. L., and Dunlop, D. J., 1975. Magnetization changes caused by burial and uplift. *Earth and Planetary Science Letters*, **28**, 133–143.
Robinson, P., Harrison, R. J., McEnroe, S. A., and Hargraves, R., 2002. Lamellar magnetism in the hematite-ilmenite series as an explanation for strong remanent magnetization. *Nature*, **418**, 517–520.
Robinson, P., Harrison, R. J., McEnroe, S. A., and Hargraves, R., 2004. Nature and origin of lamellar magnetism in the hematite-ilmenite series. *American Mineralogist*, **89**, 725–747.
Robinson, P., Fabian, K., and McEnroe, S. A., 2010. The geometry of ionic arrangements and magnetic interactions in ordered ferri-ilmenite solid solutions and its effect on low-temperature magnetic behavior. *Geochemistry, Geophysics, Geosystems*, doi:10.1029/2009GC002858.
Shcherbakov, V., and Shcherbakova, V., 1983. On the theory of depositional remanent magnetization in sedimentary rocks. *Geophysical Surveys*, **5**, 369–380.
Stacey, F. D., and Banerjee, S. K., 1974. *The Physical Principles of Rock Magnetism*. Amsterdam: Elsevier.
Tauxe, L., 2010. *Essentials of Paleomagnetism*. Berkeley: University of California Press. On-line version: http://magician.ucsd.edu/Essentials/index.html.
Tauxe, L., Constable, C., Johnson, C., Miller, W., and Staudigel, H., 2003. Paleomagnetism of the Southwestern U.S.A. recorded by 0–5 Ma igneous rocks. *Geochemistry, Geophysics, Geosystems*, doi:10.1029/2002GC000343.
Tauxe, L., Steindorf, J. L., and Harris, A., 2006. Depositional remanent magnetization: toward an improved theoretical and experimental foundation. *Earth and Planetary Science Letters*, **244**, 515–529.
Verosub, K. L., 1977. Depositional and post-depositional processes in the magnetization of sediments. *Reviews of Geophysics and Space Physics*, **15**, 129–143.
Vine, F. J., and Matthews, D. H., 1963. Magnetic anomalies over ocean ridges. *Nature*, **199**, 947–949.
Yu, Y., and Tauxe, L., 2006. Acquisition of viscous remanent magnetization. *Physics of the Earth and Planetary Interiors*, **159**, 32–42.

Cross-references

Magnetic Anomalies, Interpretation
Paleomagnetic Field Intensity
Paleomagnetism, Magnetostratigraphy
Paleomagnetism, Measurement Techniques and Instrumentation
Paleomagnetism, Polar Wander
Paleomagnetism, Principles

REMOTE SENSING AND GIS TECHNIQUES FOR TECTONIC STUDIES

Semere Solomon[1], Woldai Ghebreab[2]
[1]Det Norske Veritas, DNV Research and Innovation, Høvik, Norway
[2]Department of Geology and Environmental Science, University of Akron, Akron, OH, USA

Synonyms

Remote sensing data; Space imagery

Definition

Remote sensing. Refers to small- or large-scale acquisition of information of an object or phenomenon (e.g., earth surface features), by the use of either recording or real-time sensing device(s) such as by way of aircraft, spacecraft, satellite, buoy, or ship that are not in physical contact with the object.

Geographic information system. Refers to spatial data management and analysis tools that can assist users in organizing, storing, editing, analyzing, and displaying positional and attribute information about geographical data.

Lineaments. Any linear features that can be picked out as lines (appearing as such or evident because of contrasts in terrain or ground cover, tone, pattern, size, etc.) in aerial or space imagery.

Tectonics. A field of study within geology concerned generally with the structures within the lithosphere of the Earth (or other planets) and particularly with the forces and movements that have operated in a region to create these structures.

Neotectonics. A subdiscipline of tectonics which deals with the current or recent geologic deformations.
Active fault. A planar fracture in rocks with visible displacement or seismic activity that has occurred during the geologically recent period, as opposed to passive faults that show no movements since their formation.
Lithology. The scientific study and description of rocks/rock types in terms of compositional and physical character of the rock.
Geomorphic features. Landform features such as rivers and peneplains formed by geomorphic processes (e.g., erosion and mass-wasting) that shape them.
Strain meter. An instrument used by geophysicists to measure the deformation of the Earth in the field.

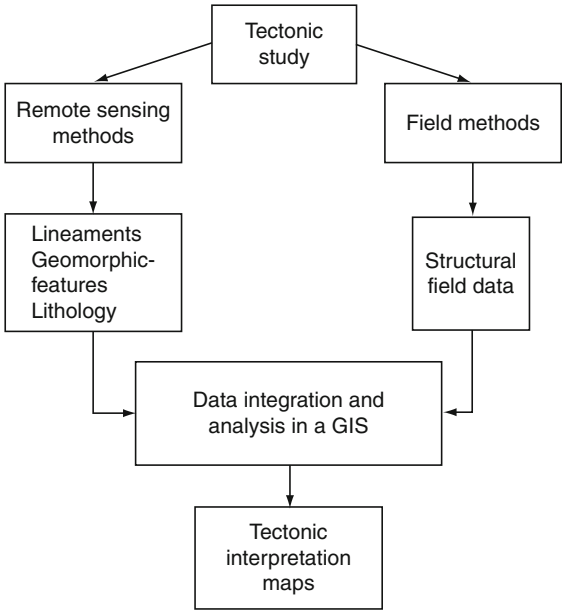
Introduction

Mapping of lineaments from various space imagery are a commonly used initial step in tectonic studies. The surface expression of geological structures such as faults, joints, dykes, veins, and straight routes of rivers are often displayed or represented in the form of lineaments in aerial photographs or on remotely sensed images. In addition, remote sensing data are useful for detecting lithology and geomorphic features, which are, in turn, important in assessing the mechanisms of faulting. The most common and widely used remote sensing techniques include: Satellite Pour l'Observation de la Terre (SPOT), which is a high-resolution optical imaging Earth Observation Satellite (EOS), Landsat satellite both multispectral scanner (MSS), and thematic mapper (TM) including enhanced thematic mapper (ETM), Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), Shuttle Radar Topographic Mission (SRTM) and/or *SAR Interferometry* (InSAR), and Global Positioning System (GPS). Data acquired from optical, thermal, radar images or digital terrain elevation at the same or different sensor platforms or devices are all useful for tectonic studies.

Lineament mapping and analysis is a combined effort of technology and field evidence (Figure 1). The availability of multispectral and multi-sensor data with synoptic coverage and use of different image enhancement techniques provides an opportunity to prepare more reliable and comprehensive lineament maps useful for tectonic studies. Field studies or ground truthing of orientation measurements of structural features and identifying rock types provide support for verifying remote sensing derived tectonic interpretation. Geographic information system (GIS) facilitates integration of all data types and provides a strong basis for tectonic interpretation of structural features.

Data and method

Mapping of geological structures from various space imageries is a commonly used initial step in tectonic studies. The routine procedure for extracting tectonic features from digital remote sensing data usually involves initial digital image enhancement followed by manual interpretation (Suzen and Toprak, 1998). There have been significant approaches for the evaluation and automatic detection of lineaments and curvilinear features from satellite images (Karnieli et al., 1996). However, the human expert judgment still remains to be an asset for lineament detection and interpretation.



Remote Sensing and GIS Techniques for Tectonic Studies, Figure 1 A simplified flow chart showing the methodology of tectonic study using remote sensing and field data in a GIS framework.

Satellite-based remote sensing data may be optical image (e.g., SPOT and Landsat), which usually senses the behavior of visible, ultraviolet, and infrared light components of the electromagnetic energy used in imaging. The data could also be thermal image in which the devices sense the thermal energy of the electromagnetic spectrum. The Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) is an example of modern remote sensing device that senses the behavior of visible, ultraviolet, infrared light, and thermal infrared components of the electromagnetic energy spectrum. The remote sensing satellite may also acquire radar images such as the Shuttle Radar Topographic Mission (SRTM) and/or Interferometric Synthetic Aperture Radar (InSAR) in which the system has a transmitter that emits microwaves or radio waves of the electromagnetic energy to detect objects. The Global Positioning System (GPS) is a US space-based global navigation satellite system. GPS is made up of three parts: satellites orbiting the Earth; controlling and monitoring stations on Earth; and the GPS receiver that reads broadcasted signals from space sent by GPS satellites. Each GPS receiver then provides three-dimensional location based on latitude, longitude, and altitude, and the time that are useful for tectonic studies.

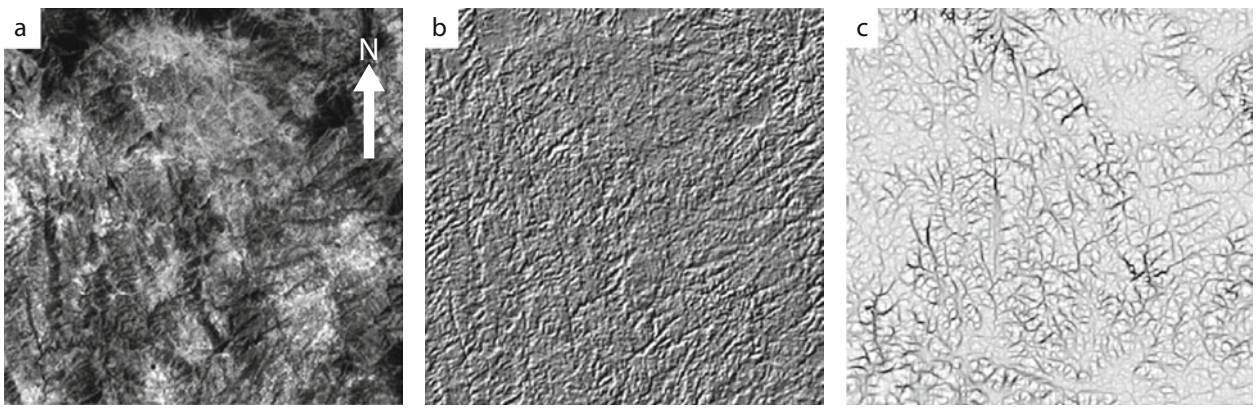
Digital elevation models (DEM) may be prepared in a number of ways, but they are frequently obtained by remote sensing rather than by direct survey. One powerful technique for generating digital elevation models is InSAR: two passes of a radar satellite (such as RADARSAT-1 or TerraSAR-X), or a single pass if the satellite is equipped with two antennas (like the SRTM instrumentation), suffice to generate a digital elevation map with extensive areal coverage and high resolution (10–30 m). Alternatively, other kinds of stereoscopic pairs can be employed using routine photogrammetric techniques, where two optical images acquired with different angles taken from the same pass of an airplane or an EOS such as the high-resolution stereo of SPOT5 or the visible and near-infrared band of ASTER (e.g., Hiranoa et al., 2003).

Remote sensing

The surface expression of lineaments displayed in aerial photographs or remote sensing data (Figure 2) may be passive faults, showing no movements since their formation or may be active faults with movements created by either earthquakes or ongoing uplifting. Earthquakes and uplifts are related to tectonic movements in the subsurface. The surface expression of these movements as structures visible on satellite imageries can be mapped and used to investigate the tectonic movements. Time series thermal satellite remote sensing datasets were used to detect pressure built-up due to tectonic activities and associated subsurface degassing that created changes in the thermal regime prior to an earthquake event (Saraf and Choudhury, 2005). Such studies are useful in locating earthquakes' epicenters and understanding the fault mechanics involved in the deformation during earthquake occurrences and uplifting processes.

DEM acquired from satellites such as SRTM data or optical images are useful for derivation of topographic parameters such as slope and surface curvatures (geomorphic indices) for mapping drainage networks and lineament extraction (Zizioli, 2008; Demirkesen, 2008). These morphotectonic parameters can be used as indicators of active tectonics in a region. For instance, rivers are sensitive to changes in tectonic deformation, adjusting their routes over different periods of time depending on the physical properties of the host rocks, climatic effects, and tectonic activity. Thus, the drainage system of a region records the evolution of tectonic deformation. Quantitative measurements of a number of geomorphic indices are commonly used as a reconnaissance tool in tectonic geomorphology studies to identify areas experiencing tectonic deformation. Combinations of different geomorphic indices derived from DEM data allow the quantification of surface deformation both on maps and on stream profiles. It is possible to assess and quantify the rate of uplift by studying the influence of faults on stream geometry and changes in direction as a result of neo-tectonic activity (Gloaguen et al., 2008). Such thematic geomorphic indices can easily be incorporated into a GIS tool for further analysis and interpretation.

Interferometric synthetic aperture radar (InSAR) is a radar technique used in geodesy and remote sensing. This geodetic method measures the line-of-sight displacements and uses two or more synthetic aperture radar (SAR) images to generate maps of surface deformation or digital elevation, using differences in the phase of the waves returning to the satellite, or aircraft. The technique can potentially measure millimeter-scale or even lower changes in deformation over time spans of days to years. InSAR can be used to measure ground movements due to earthquakes and to monitor creep and strain accumulation on faults (e.g., Stramondo et al., 2005). Due to its easy integration within a GIS environment, DEM data is important in such analyses. Time series data, a sequence of space imagery shots taken at different times, is a requirement in such studies. Then the changes are detected at observation points that have a fixed geographical location for later comparison and monitoring purposes.



Remote Sensing and GIS Techniques for Tectonic Studies, Figure 2 Different types of remote sensing data showing conspicuous NE-SW trending lineaments from 12 km × 12 km area (**a**) SPOT band 3, (**b**) digitally enhanced Landsat TM band 5, and (**c**) minimum curvature image derived from SRTM digital elevation model. The lineaments represent both extensional and strike-slip tectonics.

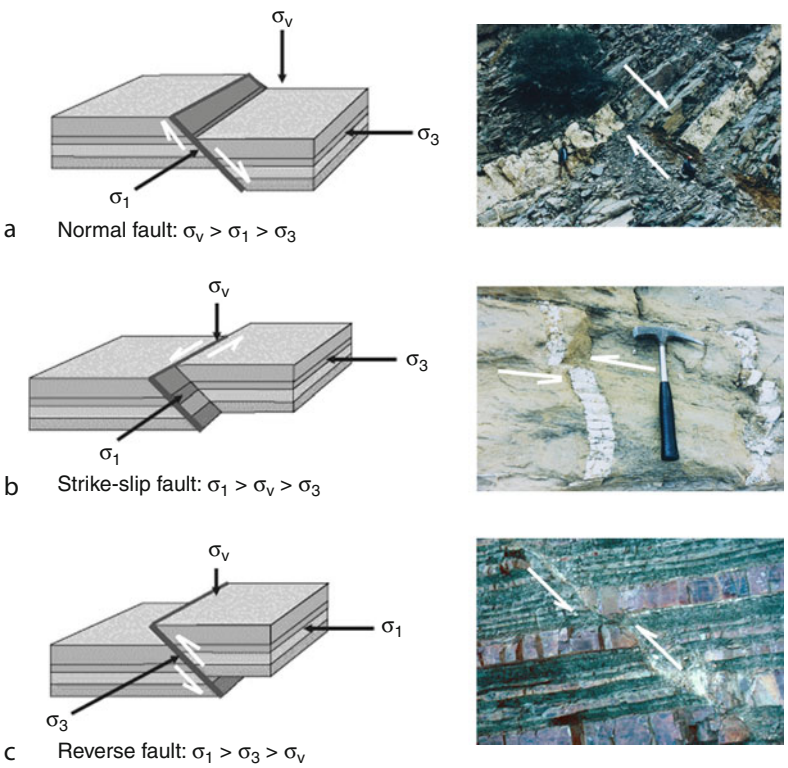
GPS measurements, among other applications, are now in use to determine the motion of the Earth's tectonic plates and deformation around active faults and volcanoes. It provides a powerful means to directly measure the kinematic pattern of present-day crustal deformation by setting up several regional GPS network stations and monitoring the changes in geographical locations and heights at the control points (Jouanne et al., 1998). GPS can measure movement of faults or tectonic plates to within the precision of a few millimeters to approximately one centimeter over baseline separations of hundreds of meters to thousands of kilometers (Segall and Davis, 1997). The three-dimensional nature of GPS measurements allows one to determine vertical as well as horizontal displacement at the same time and place. Field methods such as the use of strain meters can provide far greater strain sensitivity than does GPS. However, strain meters cannot offer the spatial coverage and long-term stability as GPS do. InSAR measurements are tremendously exciting because of their unparalleled spatial coverage. InSAR and GPS are complementary in that GPS provides long-term stability, vector displacements, and better temporal coverage as compared to the extensive spatial coverage provided by SAR/InSAR and thus results from one system are directly relevant to the others (Segall and Davis, 1997).

Tectonic studies

Tectonics deals with large-scale deformations of the Earth's lithosphere. The lithosphere is broken up into several tectonic plates that move relative to one another. Plate tectonics theory explains the formation of the plates and the origin of the forces responsible for the motion of these plates. There are three types of plate boundaries characterized by their relative motion: each having different types of surface phenomena including volcanic activities, earthquakes, and faults. The tectonic styles at the different plate boundaries are:

- Extensional tectonics that results in the formation of normal faults (Figure 3a) and associated tectonic processes due to the thinning/stretching of the lithosphere.
- Strike-slip tectonics that results in strike-slip faults (Figure 3b) in zones of lateral displacement within the lithosphere.
- Compressional tectonics that results in the formation of thrust faults (Figure 3c) and associated tectonic activities due to the shortening and thickening of the lithosphere.

The difference between structural geology and tectonics is the scale of observation. Structural geology deals with small-scale rock deformation, while tectonics is more concerned with larger features. Thus, remote sensing



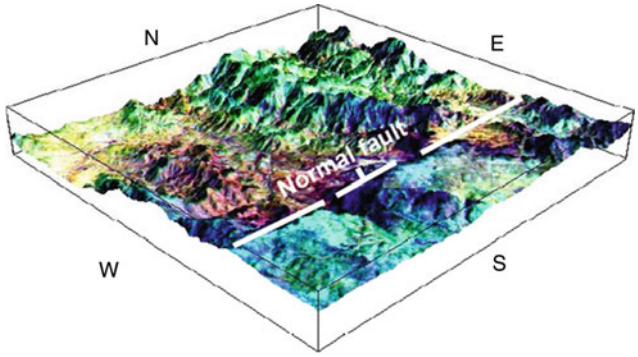
Remote Sensing and GIS Techniques for Tectonic Studies, Figure 3 Fault block models and their equivalent outcrop-scale structural features from field study formed by different tectonic processes: (**a**) Normal faults, (**b**) strike-slip faults, and (**c**) thrust faults. In the sketches the maximum principal stress (σ_1), the vertical stress (σ_v), and minimum principal stress, (σ_3) are shown. Field structures in (**a**) and (**c**) are cross-sectional views and (**b**) is plain view.

methods help in detecting large-scale tectonic features while field studies are necessary in investigating how the mapped lineaments control the tectonics of a region by looking into the details at mesoscopic-scale. All remote sensing methods used for lineament extraction cannot describe the nature of the tectonics in a region without ground truthing. It is generally recognized that lineament or fracture trace analysis used in conjunction with tectonic investigations is not credible without field verification. In regional scale studies, although field checking of each lineament is impossible due to significant cost and time involved (Mabee et al., 1994), systematic field checking is still necessary to decipher and derive confidence estimates of the nature of lineaments. The field data include observation and measurement of orientations of joints, dykes, and faults. Field evidences presented in the form of figures and sketches (e.g., Figure 3a–c) can provide good support for tectonic interpretation. The structural data collected from the field need to be grouped according to rock type and proximity of sampling points to take into account the geographic locations of observation points. In order to understand the structural–geological significance of fracture arrays, it is a common practice to subdivide them into separate sets on the basis of orientation. Rose diagrams are commonly used to reveal the orientations of vertical and steeply dipping joints, faults, and dykes observed at different scales. Many workers (e.g., Arlegui and Soriano, 1998; Solomon and Ghebreab, 2006) have effectively used remote sensing data supported by field studies to map regional spatial distributions of lineaments toward understanding their tectonic origin.

Geographic information system (GIS)

GIS is a tool in spatial data management and analysis that can assist users in organizing, storing, editing, analyzing, and displaying positional and attribute information about geographical data. GIS facilitates integration of all data types and allows a better understanding of the tectonic interpretation of a region under study. The full potential of remote sensing and GIS can be utilized when an integrated approach is adopted. Integration of the two technologies has proven to be an efficient tool in many studies, specifically in tectonic studies (e.g., Ratzinger et al., 2006; Gloaguen et al., 2008).

For instance, different geomorphic indices derived from DEM and a number of factors influencing each index can be integrated in a GIS that will allow a better interpretation of the results. Backed with remote sensing data and ground-checking, it is possible to outline the deformed area and the local strain intensity (relative uplift rates) and thus infer the neotectonic history of a region. Jain and Verma (2006) applied a combination of remote sensing and GIS for mapping active tectonic intensity zone in India. Superimposing a lineament map over a lithological map in a GIS environment provides an insight into the nature of tectonics of a region. Visible horizontal displacement along lineaments (e.g., Figure 3b) as evidenced from movement in rock units can be used to depict strike-slip fault tectonism.



Remote Sensing and GIS Techniques for Tectonic Studies, Figure 4 Three-dimensional synoptic view of a digitally processed remote sensing data created by draping from ASTER bands (13, 5, and 3) in red green and blue order over the SRTM DEM data in a GIS environment. Note the east–west normal fault that dips to the north.

Draping the lithological map or digitally enhanced color image over a three-dimensional DEM data in a GIS environment can also reveal the faulting mechanisms such as normal faulting due to visible subvertical displacement of lithological units (Figure 4). In order to classify the fault mechanics involved in deformation, an overlay of historical earthquake data on mapped tectonic features or lineaments can be integrated in a model derived from remote sensing data with results such as from differential radar interferometry in a GIS environment. A GPS-derived parameters can be easily integrated in a GIS framework to provide a powerful means of directly measuring the dynamics of present-day crustal deformation by monitoring the changes in geographical (horizontal) locations and heights (vertical displacements) at well-positioned control points. All these examples demonstrate the application of remote sensing and GIS techniques for mapping both active and passive faults. However, to better understand the types of faults involved, field checking still remains a requirement.

Summary

Remote sensing and GIS are complementary to each other and efficient techniques for lineament mapping, and hence for tectonic studies. Advanced remote sensing techniques such as InSAR and GPS have revolutionized tectonic studies by directly measuring the motion of faults associated with earthquakes and rates of uplifts. Field studies are, however, needed to identify the types of lineaments and correlate them to the remotely acquired data. Comparison of orientations of major structures or lineaments such as joints, dykes, and faults obtained from field measurements, and plotted on rose diagrams facilitates their tectonic interpretation. GIS is a very time-saving and cost-effective tool once the database is established. Integrating data of different layers such as lineament and lithological maps, and geomorphic indices

acquired, for example, from DEM, GPS measurements and field studies in a GIS environment followed by spatial analysis of the data allows correlation between different parameters for unraveling the nature of the tectonics of a region. Moreover, integrating historical dataset from past earthquake events, if any, including an understanding of the regional tectonic setting is essential in facilitating interpretation of results. Overall, combining remote sensing, GIS, and field techniques provides a powerful tool for understanding the tectonic framework of a region, and assessing the resulting fault mechanisms.

Bibliography

Arlegui, L. E., and Soriano, M. A., 1998. Characterizing lineaments from satellite images and field studies in the central Ebro basin (NE Spain). *International Journal of Remote Sensing*, **19**, 3169–3185.

Demirkesen, A. C., 2008. Digital terrain analysis using Landsat-7 ETM+ imagery and SRTM DEM: a case study of Nevsehir province (Cappadocia), Turkey. *International Journal of Remote Sensing*, **29**, 4173–4188.

Gloaguen, R., Käßner, A., Wobbe, F., Shazah, F., Mahmood, A., 2008. Remote sensing analysis of crustal deformation using river networks. *IEEE International Geoscience & Remote Sensing Symposium*, July 6–11, Boston, MA, USA, Paper: TH1.101.1.

Hiranoa, A., Welcha, R., and Langb, H., 2003. Mapping from ASTER stereo image data: DEM validation and accuracy assessment. *Journal of Photogrammetry & Remote Sensing*, **57**, 356–370.

Jain, S., and Verma, P. K., 2006. Mapping of active tectonics intensity zones using remote sensing and GIS. *Journal of the Indian Society of Remote Sensing*, **34**, 131–142.

Jouanne, F., Genaudeau, N., Ménard, G., and Darmendrail, X., 1998. Estimating present-day displacement fields and tectonic deformation in active mountain belts: an example from the Chartreuse Massif and the southern Jura Mountains, Western Alps. *Tectonophysics*, **296**, 403–419.

Karnieli, A., Meiseis, A., Fisher, L., and Arkin, Y., 1996. Automatic extraction and evaluation of geological linear features from digital remote sensing data using the Hough transform. *Photogrammetric Engineering & Remote Sensing*, **62**, 525–531.

Mabee, S. B., Hardcastle, K. C., and Wise, D. W., 1994. A method of collecting and analyzing lineaments for regional-scale fractured bedrock aquifer studies. *Ground Water*, **32**, 884–894.

Ratzinger, K., Neuhäuser, B., and Papthoma, M., 2006. Hazard mapping of earthquake triggered landslides. In *First European Conference on Earthquake Engineering and Seismology*, September 3–8, Geneva, Switzerland, p. 1405.

Saraf, A. K., and Choudhury, S., 2005. Thermal remote sensing technique in the study of pre-earthquake thermal anomalies. *Journal of Indian Geophysical Union*, **9**, 197–207.

Segall, P., and Davis, J. L., 1997. GPS applications for geodynamics and earthquake studies. *Annual Review Earth Planetary Science*, **25**, 301–336.

Solomon, S., and Ghebreab, W., 2006. Lineament characterization and their tectonic significance using Landsat TM data and field studies in the central highlands of Eritrea. *Journal of African Earth Sciences*, **46**, 371–378.

Stramondo, S., Moro, M., Tolomei, C., Cinti, F. R., and Doumaz, F., 2005. InSAR surface displacement field and fault modelling for the 2003 Bam earthquake (southeastern Iran). *Journal of Geodynamics*, **40**, 347–353.

Suzen, M. L., and Toprak, V., 1998. Filtering of satellite images in geological lineament analysis: application to a fault zone in Central Turkey. *International Journal of Remote Sensing*, **19**, 1101–1114.

Zizioli, D., 2008. DEM-based morphotectonics analysis of Western Ligurian Alps. *Scientifica Acta*, **2**, 44–47.

Cross-references

Remote Sensing, Applications to Geophysics
SAR Interferometry

REMOTE SENSING, APPLICATIONS TO GEOPHYSICS

Hojjatollah Ranjbar
Department of Mining Engineering, Shahid Bahonar University of Kerman, Kerman, Iran

Definition

Remote sensing is defined as the acquisition of data using a remotely located sensing device, that relies on the measurements of electromagnetic energy (EM) in the range of 0.4 μm to 1 m, and extraction of information from the data (Gupta, 2003; McCloy, 2006). Remote sensing can be multilevel (i.e., sensing the surface of material from a distance of few centimeters to millions of kilometers). Taking a spectrum of a leaf by using a spectroradiometer to taking an image by Hubble telescope from a star at a distance of few light years fall in the realm of remote sensing. However, here, we mean the imaging of the Earth's surface.

Introduction

The remote sensing systems can acquire images from the surface of the Earth by either passive or active methods. Passive systems use the electromagnetic energy that is produced by the sun or any other process like anomalous heat buildup within the Earth. Active remote sensing systems create their own electromagnetic energy that (1) is transmitted from the sensor toward the terrain, (2) interacts with the terrain producing a backscatter of energy, and (3) is recorded by the remote sensor's receiver (Jensen, 2000). The sending and receiving of signals are largely unaffected by atmospheric conditions. Therefore, radar satellites can record the details of the Earth's surface in cloudy weather condition and any time of the day and night. LIDAR (LIght Detection and Ranging) and SONAR (SOund NAvigation Ranging) are also the active remote sensing systems. It should be noted that remote sensing is a geophysical method as it uses the electromagnetic spectrum as a medium for surveying the Earth's surface. It is now recognized as a separate branch of science and engineering.

The electromagnetic spectrum that is used by remote sensing systems is named according to the wavelength ranges. Gupta (2003) has classified the electromagnetic spectrum into visible (0.4–0.7 μm), near infrared (0.7–1.0 μm), Shortwave infrared (1.0–3.0 μm),

mid-infrared (3.0–35 μm), far-infrared (35–1.0 mm), and RADAR (1–100 cm).

A remote sensing system that allows sensing the Earth in any wavelength and in any time makes an ideal remote sensing system. An ideal remote sensing system cannot be achieved due to the absorptions of the electromagnetic energy from the Earth's atmosphere.

Each object has unique sets of absorption and reflectance features concerning the wavelengths. The plot is known as spectral signature. Figure 1 shows the spectra of the common surface materials. Although each material at the surface of the Earth has many absorption and reflection features, the present remote sensing systems can only image these materials in limited wavelengths because the atmosphere is not clear for all the wavelengths.

There are different sensors that are used onboard the satellites or in an aeroplane. The present remote sensing systems that orbit the Earth are imaging the Earth's surface in visible, near infrared, shortwave infrared, thermal infrared, and RADAR regions of EM spectrum. The reader can refer to Jensen (2000) and Lillesand et al., (2004) for different sensors and platforms specifications.

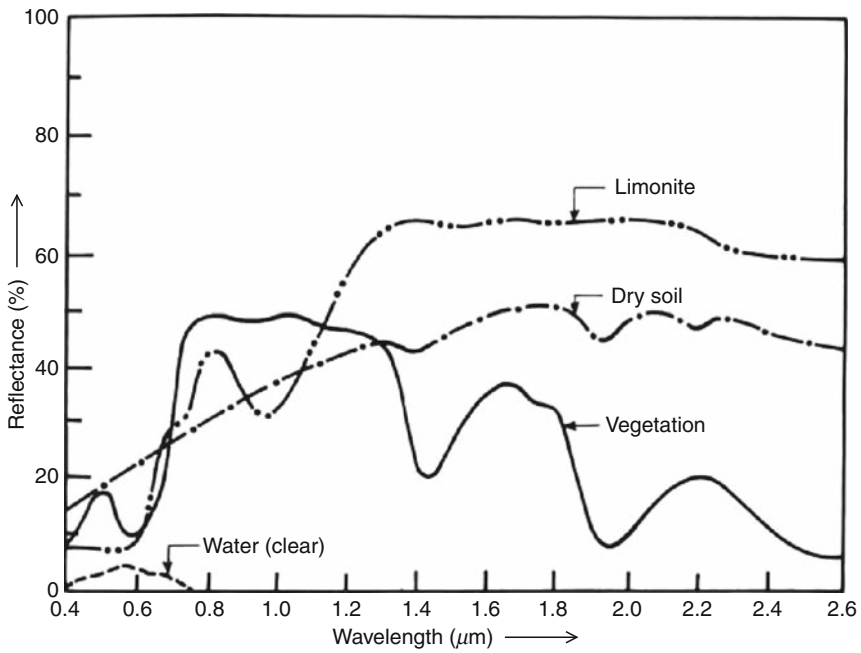
A digital image consists of pixels. Each pixel is representative of a surface area that defines the spatial resolution of the system. Enhanced Thematic Mapper Plus (ETM^+) multispectral data has a resolution of 30 m (each pixel is equivalent to 900 m^2 ground area), while Quickbird has a resolution of 60 cm for its panchromatic image (each pixel is equivalent to 0.36 m^2 on the ground).

As these images are in digital formats, several simple or complex mathematical operations can be applied on them in order to enhance specific features from these images. This is called image processing operation. Image filtering (Mather, 2001) is a common image processing technique that enhances the linear features. Figure 2 shows the enhanced faults in the southern part of Iran in a filtered image.

Application of remote sensing to geophysics

Remote sensing data can be used for designing the survey grid and line in geophysical surveys, verifying, and helping a better interpretation of the geophysical methods. Remote sensing images can be combined to form color images in various scales that depend on the spatial resolution of the remote sensing system. These color images can be draped over a digital elevation model (DEM) to give a three-dimensional (3-D) model of the area (Figure 3). These 3-D models can be used in both land and airborne geophysical surveys for designing the survey lines. The 3-D model can provide information such as the terrain ruggedness, man-made installations, surface conditions, transportation routes, etc.

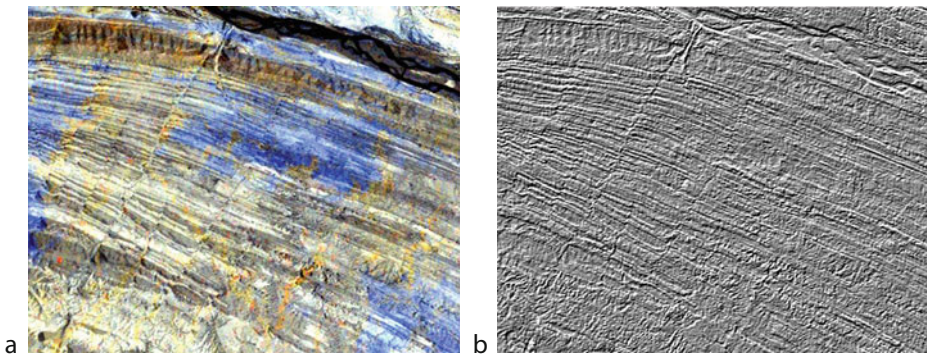
The surface of the Earth is continuously deformed due to the tectonic activities over geological time that causes the formation of structures such as fold and fault. Geophysical methods such as magnetic, magneto-telluric, gravity, and seismic methods are used for studying these structures. These methods are subsurface methods.



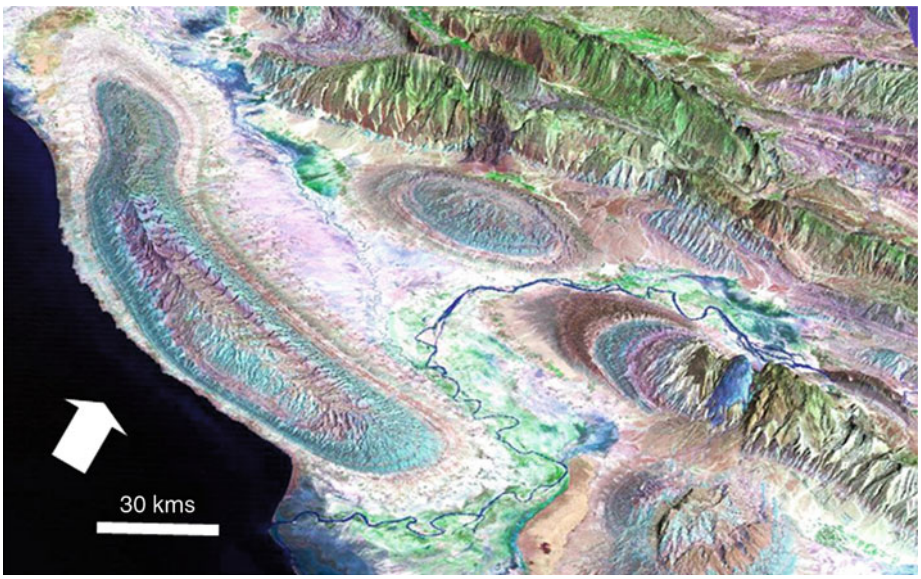
Remote Sensing, Applications to Geophysics, Figure 1 Typical spectral reflectance curves for selected common natural objects (Gupta, 2003).

Remote sensing techniques are useful in order to study these structures and to interpret the tectonic activities in the regional and local scale at the surface of the Earth. In order to obtain a clear picture of the structures, the combined interpretation of remote sensing and geophysical data is helpful, if the geological structures have any trace in the satellite images. Remote sensing applications to tectonic and structural studies are reported by many researchers (e.g., Harris, 1991; Philip, 1996; Saintot et al., 1999; Chernicoff et al., 2002; Chatterjee, 2003; Raharimahefa and Kusky, 2006; Walker, 2006). Figure 3 shows a part of Zagros simply folded belt that the folds axis directions indicate a force direction because of the subduction of Arabian plate beneath the Iranian plate. These structures are also studied by using gravitational, magnetic, and seismic methods.

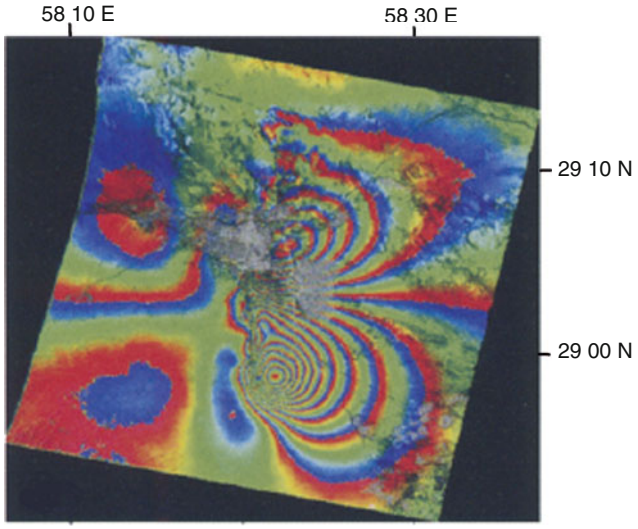
Remote sensing is also useful for interpretation of geophysical data. Alteration process causes the changes in mineralogical content of rocks due to the reaction of hot solution with the rocks. Porphyry copper mineralization is associated with these alteration zones (Evans, 1993). Both radiometric and remote sensing methods can be used for alteration detection (Ranjbar et al., 2001; Ranjbar and Honarmand, 2004; Ranjbar et al., 2010). The minerals present in sericite and k-feldspar have a higher gamma radiation. We cannot differentiate these minerals by geophysical methods. Remote sensing method is capable of recognizing the minerals responsible for the higher



Remote Sensing, Applications to Geophysics, Figure 2 Enhancement of faults and other linear features by using directional filter. (**a**) Original satellite color image, (**b**) filtered image.



Remote Sensing, Applications to Geophysics, Figure 3 3-D view of folding due to the movement of Arabian plate against Iranian microplate. The force direction is shown by an *arrow*. Landsat color image (Band7 in *red*, band 4 in *green*, and band 2 in *blue*) is draped over DEM of the area.



Remote Sensing, Applications to Geophysics, Figure 4 An interferogram that shows the deformation along a fault due to the Bam earthquake in Iran. The rings that are closer, indicating more deformation (After Ye, 2005).

radiation in geophysical data. There are also minerals in the alteration zones that have very low radiation to be detected by geophysical methods. Kaolinite, Calcite, epidote, and chlorite are such minerals. These minerals are recognized by remote sensing methods. Remote sensing techniques can also be used in combination with other geophysical methods such as magnetic, electromagnetic, and electrical methods in helping with the interpretation.

Another area that received much attention is the field of active faults study by the combined use of interferometry and geophysical data. RADAR (Radio Detection And Ranging) images are more useful for detection of the geological structures than the images acquired in the optical ranges of EM spectrum, because the RADAR pulses are illuminating the ground from a specific direction. The RADAR wavelengths are in order of few tens of centimeters. Synthetic Aperture Radar interferometry (INSAR) is a branch of remote sensing that has emerged in the recent years that deals with the deformation in the Earth's crust. There are several papers published about INSAR method in structural and tectonic fields (e.g., Klees and Massonnet, 1999; Wright et al., 2001, 2004; Ye, 2005; Bürgmann et al., 2006; Taylor and Peltzer, 2006). Figure 4 shows an interferogram of a fault in Bam area that triggered an earthquake in 2003. This fault is detected by remote sensing and geophysical methods.

Summary

Remote sensing images due to their multi-temporal, multispectral, and synoptic views are extensively used for solving problems related to mineral exploration, tectonics, vegetation cover studies, surface changes, earthquake studies, weather and volcanic eruption forecasting, etc. The images can be acquired in different wavelength ranges such as visible, near infrared, thermal, and Radar ranges of electromagnetic spectrum. We are now able to recognize many minerals from the space by using their spectral properties. Remote sensing data can be used for designing the survey grid and line in geophysical surveys, verifying, and helping a better interpretation of the geophysical methods. Small and large structures are easily recognizable in the images, although we often need to enhance the images for recognizing these structures.

Bibliography

Bürgmann, R., Hilley, G., Ferretti, A., and Novali, F., 2006. Resolving vertical tectonics in the San Francisco Bay Area from permanent scatterer InSAR and GPS analysis. *Geology*, **34**, 221–224.

Chatterjee, R. S., 2003. Structural pattern of Holenarsipur Supracrustal Belt, Karnataka, India as observed from digitally enhanced high-resolution multi-sensor optical remote sensing data aided by field survey. *International Journal of Applied Earth Observation and Geoinformation*, **4**, 195–215.

Chernicoff, C. J., Richards, J. P., and Zappettini, E. O., 2002. Crustal lineament control on magmatism and mineralization in northwestern Argentina: geological, geophysical, and remote sensing evidence. *Ore Geology Reviews*, **21**, 127–155.

Evans, A. M., 1993. *Ore geology and industrial minerals: an introduction*. Oxford: Blackwell.

Gupta, R. P., 2003. *Remote Sensing Geology*. Berlin: Springer.

Harris, J. R., 1991. Mapping of regional structure of eastern Nova Scotia using remotely sensed imagery: implications for regional tectonics and gold exploration. *Canadian Journal of Remote Sensing*, **17**, 122–136.

Jensen, J. R., 2000. *Remote sensing of the environment: an Earth resource perspective*. New jersey: Prantice Hall.

Klees, R., and Massonnet, D., 1999. Deformation measurements using SAR interferometry: potential and limitations. *Geologie en Mijnbouw*, **77**, 161–176.

Lillesand, T., Keifer, R. W., and Chipman, J. W., 2004. *Remote Sensing and Image Interpretation*. New York: Wiley.

Mather, P. M., 2001. *Computer Processing of Remotely-Sensed Images. An Introduction*. Chichester: Wiley.

McCloy, K. R., 2006. *Resource Management Information Systems: Remote Sensing, GIS and Modelling*, 2nd edn. Boca Raton: Taylor and Francis.

Philip, G., 1996. Landsat thematic mapper data analysis for quaternary tectonics in parts of the Doon valley, NW Himalaya, India. *International Journal of Remote Sensing*, **17**, 143–153.

Raharimahefa, T., and Kusky, T. M., 2006. Structural and remote sensing studies of the southern Betsimisaraka Suture, Madagascar. *Gondwana Research*, **10**, 186–197.

Ranjbar, H., and Honarmand, M., 2004. Integration and analysis of airborne geophysical and ETM+ data for exploration of porphyry type deposits in the Central Iranian Volcanic Belt, using fuzzy classification. *International Journal of Remote Sensing*, **25**, 4729–4741.

Ranjbar, H., Hassanzadeh, H., Torabi, M., and Ilaghi, O., 2001. Integration and analysis of airborne geophysical data of the Darrehzar area, Kerman province, Iran, using principal component analysis. *Journal of Applied Geophysics*, **48**, 33–41.

Ranjbar, H., Masoumi, F., and Carranza, E. J. M., 2010. Evaluation of geophysics and spaceborne multispectral data for alteration mapping in Sar Cheshmeh mining area, Iran. *International Journal of Remote Sensing*, doi:10.1080/01431161003745665.

Saintot, A., Angelier, J., and Jean Chorowicz, J., 1999. Mechanical significance of structural patterns identified by remote sensing studies: a multiscale analysis of tectonic structures in Crimea. *Tectonophysics*, **313**, 187–218.

Taylor, M., and Peltzer, G., 2006. Current slip rates on conjugate strike-slip faults in central Tibet using synthetic aperture radar interferometry. *Journal of Geophysical Research*, **111**, 1–16.

Walker, R. T., 2006. A remote sensing study of active folding and faulting in southern Kerman province, S.E. Iran. *Journal of Structural Geology*, **28**, 654–668.

Wright, T., Fielding, E., and Parsons, B., 2001a. Triggered slip: observations of the 17 August 1999 Izmit (Turkey) earthquake using radar interferometry. *Geophysical Research Letters*, **28**, 1079–1082.

Wright, T., Parsons, B., and Fielding, E., 2001b. Measurement of interseismic strain accumulation across the North Anatolian Fault by satellite radar interferometry. *Geophysical Research Letters*, **28**, 2117–2120.

Wright, T. J., Lu, Z., and Wicks, C., 2004. Constraining the slip distribution and fault geometry of the *M*w 7.9, 3 November 2002, Denali Fault Earthquake with interferometric synthetic aperture radar and global positioning system data. *Bulletin of the Seismological Society of America*, **94**, S175–S189.

Ye, X., 2005. Bam earthquake: surface deformation measurement using radar interferometry. *Acta Seismological Sinica*, **18**, 451–459.

Cross-references

Earthquakes and Crustal Deformation
Earthquake Precursors and Prediction
Earthquake Rupture: Inverse Problem
Gravity Method, Airborne
Magnetic Anomalies, Interpretation
Magnetic Methods, Airborne
Plate Tectonics, Precambrian
Remote Sensing and GIS Techniques for Tectonic Studies
SAR Interferometry
Seismic Data Acquisition and Processing
Subduction Zones
Very Long Baseline Interferometry

S

SAR INTERFEROMETRY

Masato Furuya
Department of Natural History Sciences, Hokkaido University, Sapporo, Japan

Synonyms

Differential InSAR (abbreviated as D-InSAR); Interferometric SAR (abbreviated as InSAR); Radar interferometry; SAR interferometry

Definition

Radar. Acronym standing for Radio Detection and Ranging. A technique to detect any targets and measure the distance to them, based on the round-trip time of microwave (radio wave) pulses between the antenna and the targets.
SAR. Acronym standing for Synthetic Aperture Radar. A technique to image any ground surfaces, using airborne or spaceborne radar sensor. Its high spatial resolution is achieved by collecting numerous return pulses from each target in sight and by effectively synthesizing large antenna size.
InSAR. Acronym standing for Interferometric SAR. A technique to image surface topography and ground displacements, using phase values of two or more SAR images.

Introduction

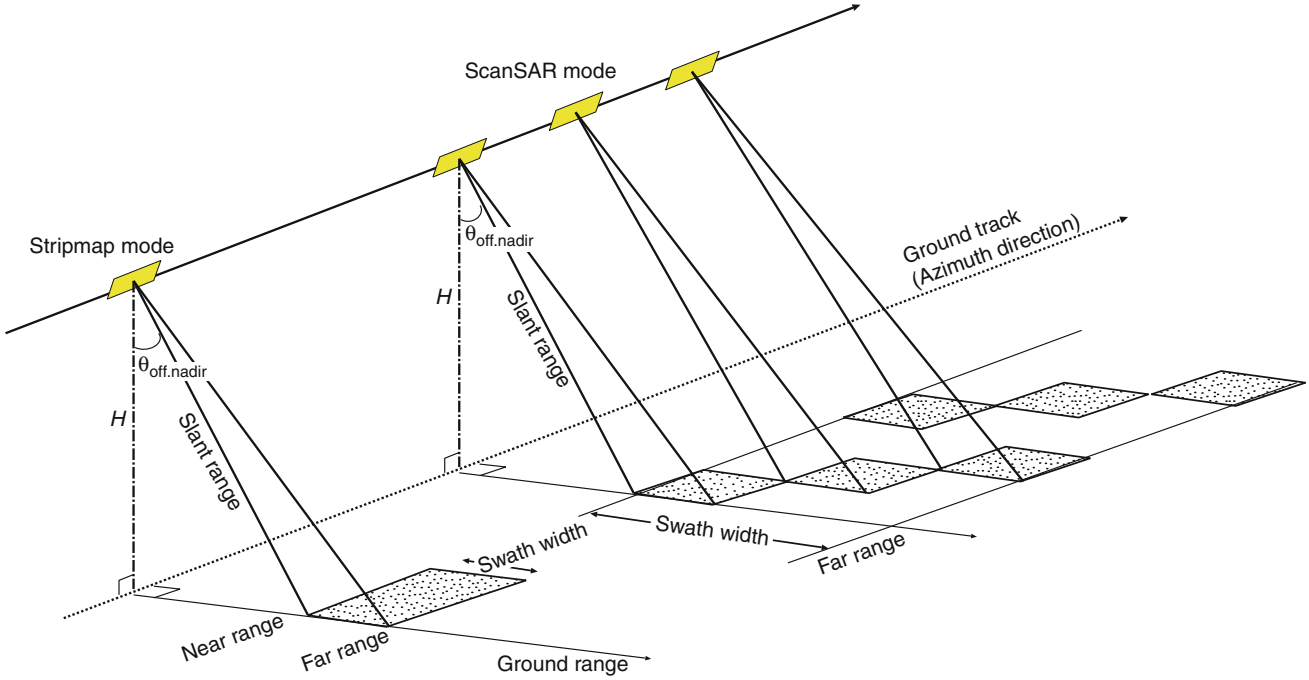
Crustal deformation data have been traditionally acquired by ground-based geodetic techniques such as leveling, triangulation, and electro-optic distance measurement. More recently, global positioning system (GPS) has become a standard tool for high-precision crustal deformation measurement, and provided us with a wealth of data to study plate tectonics, earthquakes, volcanic activities, and atmospheric and hydrological loading deformation. All these techniques, however, require in situ benchmarks, and thus prevent us from observing inaccessible areas. Interferometric SAR (InSAR) was, therefore, regarded as a surprising and revolutionary technique when Massonnet et al. (1993) first showed an image of the co-seismic deformation associated with the 1992 M7.3 Landers earthquake, because the raw data was completely acquired on a spaceborne sensor. Another big surprise for the community was its incredibly high spatial resolution, which no other geodetic techniques were possible to achieve in practice.

Nowadays, InSAR users have proliferated in a worldwide community and applied to a variety of geophysical problems. A number of excellent and extensive reviews for advanced students and researchers are already published (Bamler and Hartl, 1998; Massonnet and Feigl, 1998; Burgmann et al., 2000; Hanssen, 2001; Pritchard, 2006; Simons and Rosen, 2007; Zhou et al., 2009). I therefore tried to make this article much shorter and more introductory, but it still includes necessary and useful concepts, ranging from the fundamentals of SAR/InSAR imagery to more up-to-date topics.

Fundamentals of SAR imaging and SAR data

SAR satellite flies over at an altitude of hundreds of km, repeating transmission and reception of microwave pulses. The along-track and across-track axes are almost identical to the *azimuth* and *range* axis in the acquired radar image. The area illuminated on the ground is called *swath*, whose width spans roughly 50–100 km in the standard *stripmap* (or *strip*) mode with an incidence angle of 20–50° (Figure 1). While previous SAR applications are mostly derived from the stripmap mode, another imaging mode, *ScanSAR*, is also promising because it covers much wider swath width, 300–500 km, by illuminating multiple swaths at the expense of reducing the resolution. ScanSAR is useful for imaging long-wavelength signals

Harsh K. Gupta (ed.), *Encyclopedia of Solid Earth Geophysics*, DOI 10.1007/978-90-481-8702-7,



SAR Interferometry, Figure 1 Geometry of SAR imaging. SAR sensor transmits microwave pulses in slant range direction, and receives their reflected pulses. While stripmap mode achieves high spatial resolution with a fixed off-nadir angle, ScanSAR mode achieves wider imaged area (swath) with multiple off-nadir angles at the expense of the resolution.

associated with, for instance, a magnitude-8-class earthquake (Motagh et al., 2008).

Although it was not strictly necessary, satellite-based SAR system has been often placed on a sun-synchronous and near-polar orbit with an inclination angle of slightly greater than 90°. When the satellite moves to the north (south), we call it is in *ascending* (*descending*) orbit.

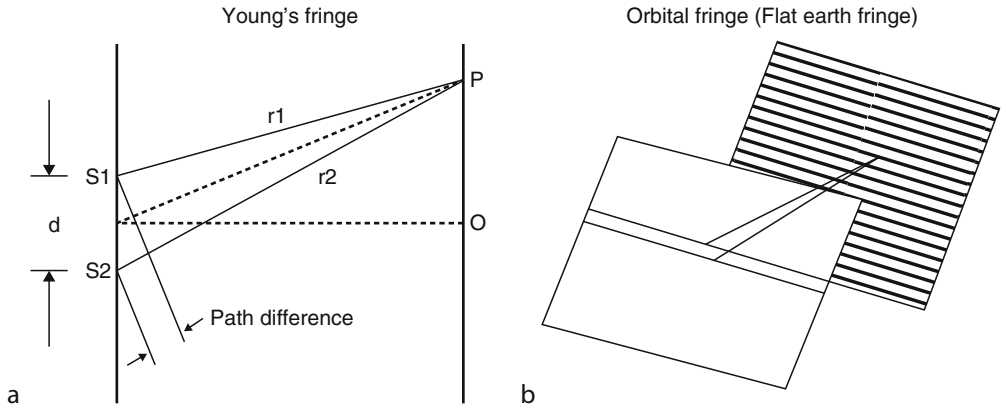
The raw data acquired on SAR sensor are impossible to visually interpret, and require a bit involved processing algorithms; those algorithms are detailed in a couple of text books (e.g., Curlander and McDonough, 1991; Cumming and Wong, 2005). The first interpretable SAR image is a single-look-complex (SLC) image, whose important difference from other optical images is that each pixel consists of a complex (real and imaginary) value, i.e., amplitude and phase. This is because the waveform of each repeated pulse is precisely controlled to be identical, and hence the received pulse provides us with not only a scattering (reflection) *intensity* but also a *phase*. The phase data do contain the geometric information from the antenna to the ground targets, and are fully exploited in generating InSAR image as discussed later. However, the phase image itself is usually not as useful as the intensity image because it is again impossible to visually interpret the physical meaning. Meanwhile, the intensity image is often useful and derived from a square-root magnitude of SLC data with spatial averaging called *multi-looking*. By single-look, it means the finest spatial resolution for both range and azimuth axis. In the standard stripmap mode, the range and azimuth resolutions are derived as,

$$\Delta r = \frac{c}{2B}, \quad and \quad \Delta a = \frac{L}{2}, \tag{1}$$

respectively; the c, B, and L are the speed of light, the frequency bandwidth of the microwave pulse, and the antenna length along azimuth axis, respectively (Curlander and McDonough, 1991; Cumming and Wong, 2005). The waveform of each microwave pulse is called *chirp signal*, whose instantaneous frequency linearly changes by as much as the frequency bandwidth B over the duration of each pulse. It should be noted that the spatial resolution depends neither on the sensor altitude nor the carrier frequency of microwave. Intensity images are often shown in gray scale images, in which strongly (weakly) reflected objects/areas are usually colored as bright (dark). Although they simply look like black-and-white photographs, we should keep in mind that they could be acquired regardless of weather and time because SAR is actively transmitting and receiving microwaves. Also, intensity images are indispensable for high-precision image matching prior to a generation of InSAR image.

Fundamental principles of InSAR

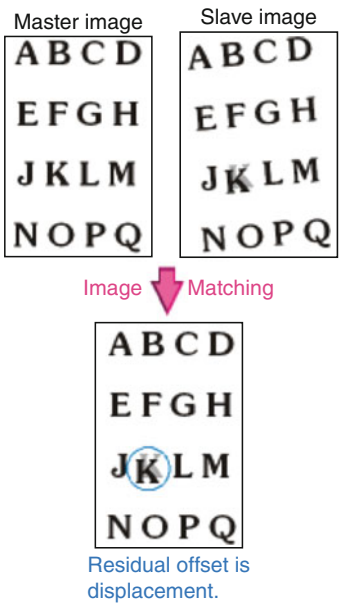
Interferometric SAR (InSAR) is a technique to generate a digital elevation model (DEM) or a ground displacement



SAR Interferometry, Figure 2 (**a**) Geometry of the Young's experiment. Depending on the path difference, the two coherent waves from the slit, S1 and S2, are in-phase or out-of-phase on the screen, and interference fringes are observed on the right screen. (**b**) Orbital fringe (flat earth fringe) can be regarded as a 3-D analogue of the Young's experiment.

image from a *pair* of SLC images. The term *interferogram* is often used to represent InSAR image. We can understand the principle of InSAR, recalling the classical Young's experiment that is known to be a proof of the wave characteristics of the light (Ghilia and Pritt, 1998). Two coherent waves out of the slits will generate "stripes" on the wall, called interference *fringe* (Figure 2a). We can simulate the fringe if we know the separation of the slits, the distance from each slit to the wall, and the wavelength of the coherent wave. Depending on the path difference, the two coherent waves are in-phase or out-of-phase when they reach the screen. Namely, the difference of the phases generates the interference fringe. We may regard the imaging geometry of InSAR as the 3-D Young's experiment (Figure 2b). The repeat orbit tracks, the ground surface, and the microwave correspond to the double slits, the screen, and the coherent wave, respectively. Once we get two SLC images, we can generate an initial interferogram, multiplying one SLC image with the complex conjugate of the other SLC image. We then observe similar fringes in the initial interferogram as illustrated in Figure 2b, which is literally a map of the difference of two SLC phases. For descriptive purposes, the former SLC image is often denoted as *master*, and the latter SLC image is called *slave*. At this moment, the slave image must be precisely co-registered (or matched) to the master image (Figure 3); we will come back to this *image co-registration* (or *image matching*) procedure later on.

While Figure 2b shows an initial interferogram over flat areas with parallel orbits, the fringe will appear undulated if the areas are not flat. The fringe over flat areas is called *flat Earth fringe* (or, *orbital fringe*), and can be precisely simulated from the pair of orbit data. If we subtract the flat Earth fringes from the initial interferogram, we can extract *topographic fringe* that can be used to generate DEM. The Shuttle Radar Topography Mission (SRTM) was carried out along this idea in 2001, and generated 3-s resolution DEM over +/− 60° latitudes (Farr et al., 2007). In the case



SAR Interferometry, Figure 3 Image registration (matching) of the master and slave images prior to interferogram generation, and the principle of pixel-offset technique to derive large displacements. While long-wavelength distortion can be corrected, localized huge displacement remains as residual offset. Courtesy of Tobita et al. (2001a).

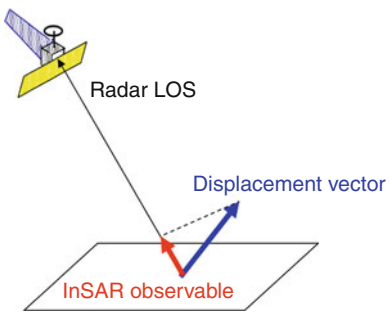
of SRTM, they carried two SAR antennas on the same platform, and thus were able to generate DEM without repeating the previous orbit track. In contrast, all the present SAR satellite systems carry only one antenna with a repeat-pass period of several weeks, which are 11 days for TerraSAR-X, 16 days for COSMO-SkyMed, 24 days for Radarsat-1/2, 35 days for Envisat, and 46 days for ALOS. Therefore, if ground surface undergoes significant deformation during the repeat orbit cycles due, for

instance, to earthquake and volcanic eruption, the interferogram will include *deformation fringe* as well. To extract deformation fringe, we must take out both orbital fringe and topographic fringe, which can be simulated from satellite orbit data and DEM. The deformation fringes represent slant range changes along the radar line-of-sight (LOS), and thus projections of the 3-D displacement vector on the ground along the unitary vectors toward the radar LOS (Figure 4). The range changes should be interpreted as relative displacements to the reference point(s) inside each interferogram. Depending on literatures, they denote differential interferometric SAR (D-InSAR) when the technique is used to detect deformation signals. Recently, however, the term InSAR is often and simply used to represent D-InSAR.

Even if no significant ground displacements take place during the repeat-pass period, however, we usually encounter other non-negligible fringes due to the spatial heterogeneities in the propagation delay of microwaves through the atmosphere, the errors in satellite orbit data, and those in DEM. Because these fringes limit the precision and accuracy of SAR-based crustal deformation measurement, a couple of correction approaches have been proposed. More advanced time-series analysis techniques have also been developed to overcome the issues, which will be introduced in the last section.

InSAR processing

Image registration (Matching): Before we get an initial interferogram, we must register (or, match) each imaged target in one SLC image to the same target in the other SLC image with a sub-pixel level accuracy, because any ground objects do not usually locate at the same pixel coordinates in each SLC image. This pre-processing is called *image registration* (or *image matching*) and prerequisite to be performed prior to generating an initial interferogram. Although a simple polynomial transformation between the range and azimuth coordinates of two SLC images is sufficient in most cases, we need to take into account the effects of 3-D topography when the terrain surface is rugged to eliminate a stereoscopic effect (Michel et al., 1999).



SAR Interferometry, Figure 4 InSAR observable is a projection of the displacement vector along the radar line-of-sight (LOS) direction.

When large ground displacements on the order of meters or more take place locally, and if we correct for the long-wavelength image distortion using the polynomial transformation, we can detect and quantify those localized displacements as a by-product of image registration without viewing InSAR image (Figure 3; Tobita et al., 2001a). This approach to detect large displacements is called *pixel offset* or *feature tracking* technique, and has been applied to earthquakes, volcanic eruptions, and glacier movements. The advantages of pixel-offset data are twofolds. First, pixel-offset data can quantify large displacements even in such areas that completely loses interferometric coherence, where InSAR data cannot be unwrapped; we describe *coherence* and *unwrapping* later below. Secondly, in contrast to InSAR data, pixel-offset data provide us with not only range offset but also azimuth offset component. While the range offset has the same sensitivity to the 3-D displacement vector as InSAR data (Figure 4), the azimuth offset is a projection of the displacement vector onto the unitary vector perpendicular to the LOS. Hence, the azimuth offset data are complementary to the range offset or InSAR data. Taking advantage of this property, Fialko et al. (2001) derived a full 3-D displacement map for the 1999 M7.1 Hector Mine earthquake, combining the InSAR data from both ascending and descending track with the azimuth offset data. Using pixel-offset data from both descending and ascending track, Tobita et al. (2001a, b) inferred a 3-D displacement map associated with the 2,000 eruption episode at Usu volcano.

Interferometric phase and its relation to geometry: Suppose we have two co-registered SLC images, E_1 and E_2, acquired from different ranges r_1 and r_2:

$$E_1 = e^{j\phi_{Scatter}} e^{-\frac{4\pi r_1}{\lambda}} \tag{2a}$$

$$E_2 = e^{j\phi_{Scatter}} e^{-\frac{4\pi r_2}{\lambda}} \tag{2b}$$

Here we assume that the reflection magnitude and scattering phase are constant during the data acquisition time. Then, the interferometric phase ϕ is derived as

$$E_1 E_2^* = e^{-\frac{4\pi(r_1 - r_2)}{\lambda}} \tag{3}$$

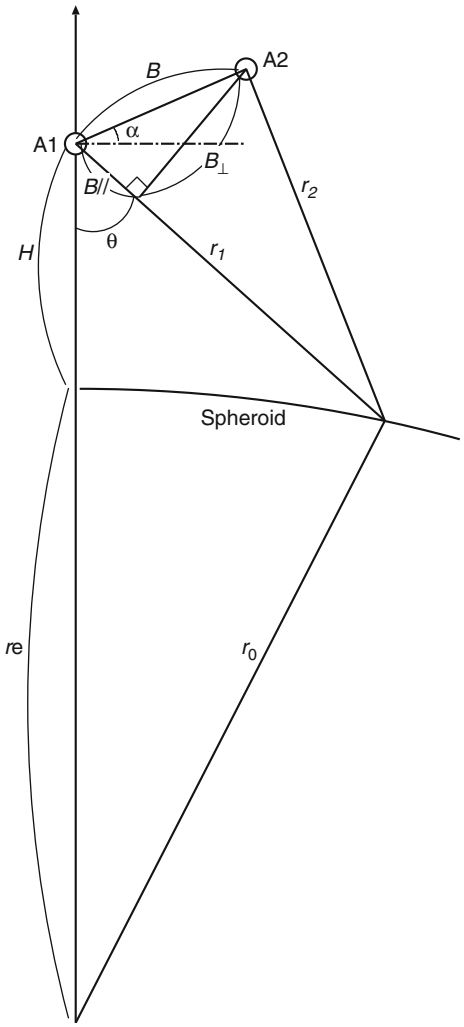
or

$$\phi = \frac{4\pi}{\lambda}(r_1 - r_2) \tag{4}$$

The last one is the fundamental equation for InSAR, which describes "unwrapped" phase in the initial interferogram. The actual phase in the initial interferogram is "wrapped" into an interval $[-\pi, \pi]$, and thus has ambiguities of $2\pi N$; N is integer. In order to quantify the ground displacement along radar LOS, we have to perform 2-D *phase unwrapping* on the interferogram, which is not necessarily straightforward (Bamler and Hartl, 1998; Ghilia and Pritt, 1998). While the interferometric phase is strictly a phase "difference" of two SLC phases, it is conventional

to simply call phase. The factor 4 is to take into account the round-trip distances.

Figure 5 is a cross section that is perpendicular to the satellite repeat tracks and passes through the Earth's center, and shows a geometry of InSAR data acquisition. The spatial separation of the repeating orbits is termed *baseline* (or *spatial baseline*), B; the temporal separation of data acquisition is sometimes called *temporal baseline*. Because the baseline B is usually much shorter than the ground range distance R, a parallel ray approximation holds (Zebker et al., 1994) and the fundamental Equation 4 can be approximated as follows:



SAR Interferometry, Figure 5 Geometry of InSAR data acquisition and its relation to the baseline. The A1 and A2 are the satellite's repeat orbits, and the spatial distance between the A1 and A2 is the baseline B. The initial InSAR phase is proportional to the difference between the ranges, r_1 and r_2, and hence the B_{para} (eq. 5). The fringe rate (density) along the range axis is proportional to the B_{perp} (eq. 6).

$$\phi = \frac{4\pi}{\lambda}(r_1 - r_2) \approx -\frac{4\pi}{\lambda}B_{//} = B\sin(\theta - \alpha), \qquad (5)$$

where θ and α are defined in Figure 5, and $B_{//}$ (or B_{para}) is a baseline component parallel to the radar LOS. The angle θ is called *off-nadir angle*, and is identical to incidence angle if the Earth's curvature is negligible. The other baseline component $B_\perp$ (or B_{perp}) is perpendicular to radar LOS and gives us an important criterion for successful InSAR processing as we discuss below.

Decorrelation (Coherence): In the actual InSAR data processing, we do not necessarily get clear fringes over the entire area. Depending on the data pairs and places, it is not uncommon that no fringes are observed. To detect clear fringes, the reflected waves received at master and slave acquisitions must be more or less correlated to each other. The degree of correlation is quantified as *coherence*, and there are two independent decorrelation sources.

The first source of decorrelation originates in the imaging geometry. As Figure 6 indicates, we observed higher (fewer) fringe density as becomes longer (shorter); imagine the case of zero baseline length. The fringe density can be derived from the gradient of phase (Equation 5) along the range axis:

$$\frac{\partial\phi}{\partial R} = -\frac{4\pi B_\perp}{\lambda R\tan\theta} + \frac{4\pi B_\perp}{\lambda(r_e + H)\sin\theta} \approx -\frac{4\pi B_\perp}{\lambda R\tan\theta}. \qquad (6)$$

Namely, the fringe density is proportional to the perpendicular baseline $B_\perp$, and inversely proportional to the wavelength λ; see Simons and Rosen (2007) for the case with topography. If the fringe density becomes too high to be counted within a range resolution of SAR image, we will not be able to identify any orbital fringes. This type of decorrelation is termed *baseline decorrelation* (or *spatial decorrelation*). The *critical baseline* is given as such a perpendicular baseline that gives a phase gradient 2π over the range resolution Δr;

$$B_\perp^c = \frac{\lambda R\tan\theta}{2\Delta r}.$$

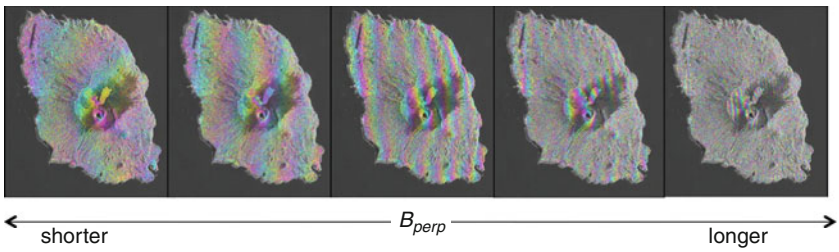
For a typical value of ALOS/PALSAR with $\lambda = 23(cm)$, $R = 870(km)$, $\theta = 34°$, $\Delta r = 5(m)$, the critical baseline becomes $B_\perp^c = 135,000(m)$, which gives an upper limit of $B_\perp$. However, we practically prefer much shorter $B_\perp$, generally less than ~2,000 m for ALOS/PALSAR, because in more realistic situations the effect of topography also comes in. The longer the $B_\perp$, the more sensitive to rugged terrain as Figure 6 indicates. To eliminate topographic fringes, we need more accurate and higher resolution DEM if the $B_\perp$ becomes longer. Massonnet et al. (1996) proposed an alternative approach that could effectively reduce the $B_\perp$ by a combination of integer multiplied (wrapped) interferograms. For instance, if one interferogram with perpendicular baseline of 300 m is combined with the other interferogram with perpendicular baseline of 290 m with factors 1 and −1, the effective perpendicular baseline becomes 10 m.

The scaling operation, however, also scales the amount of noise, and thus the approach is limited to small integer numbers.

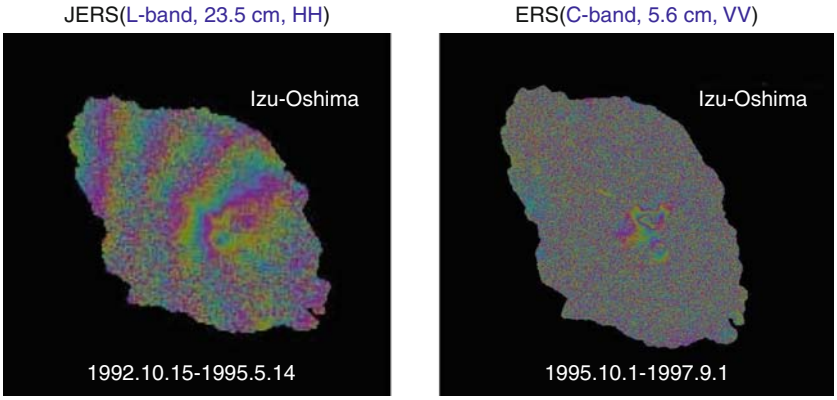
The second type of decorrelation is termed *temporal decorrelation*, which is related to the scattering phase in the Equation 2a, and originates in how the microwave pulses interact with the physical objects near the ground. We often encounter the temporal decorrelation problem over vegetated areas with C-band (shorter-wavelength) SAR data and/or snow-covered areas; see Figure 7. It should be recalled that each pixel value in SLC image is a superposition of all the reflected microwaves from all scatterers inside each resolution cell ($\sim 5 \times \sim 10$ m). Short-wavelength microwave pulses tend to be reflected on the vegetation canopies before reaching the ground surface, and their random motion will result in different scattering phases at different acquisition time, causing temporal decorrelation. On the contrary, long-wavelength microwave pulses can more easily reach the ground, which does not move as rapidly as vegetations, and thus the resulting scattering phases will be also stable over time. Besides the selection of wavelength, the polarization of microwave is also essential for better coherence over time. While, most presently, operated satellite-SAR sensors are capable of multi-polarization modes, it was shown that HH-polarization gives better coherence than VV-polarization (Cloude and Papathanassiou, 1998). This is because the HH-polarized pulses can more easily penetrate through vegetations.

Outlook for InSAR geodesy

Limitations of present InSAR: Although it has a potential to detect tens of km-scale or even larger-scale secular deformation signals on the order of mm/year, InSAR technique has been most successfully applied to detection of spatially localized signals on the order of centimeters or more, such as those associated with earthquakes, volcanic eruptions, and ground subsidence. This is because the artifacts due to inaccurate satellite orbit data and/or microwave propagation delays (advances) in the troposphere (ionosphere) can mask small-amplitude, long-wavelength deformation signals that are similar in both their amplitude and the spatial scale.



SAR Interferometry, Figure 6 The fringe rate (density) depends on the B_{perp}; see eq (6). The shorter the B_{perp}, the fewer the observed fringes, and thus better to detect deformation signals. In order words, there is a limit in the B_{perp} over which we cannot count the number of fringes. The InSAR image is based on JERS data over Izu-Oshima volcano island, Japan. Original SAR data is copyrighted by JAXA and MITI, Japan.



SAR Interferometry, Figure 7 Comparison of two interferograms at Izu-Osima volcano, derived from (left) L-band HH JERS data and (right) C-band VV ERS data. While clear fringes are observed to the left even with 2.5 years temporal baseline, we can recognize the fringes only around the caldera that are covered with few vegetations.

Although high-precision orbit data are indispensable to correct for the orbital fringes in the initial interferograms, their errors even on the order of 10 cm or less will generate non-negligible long-wavelength artifacts, which usually look like curved surfaces in the entire interferogram (e.g., Hanssen, 2001). Conventionally, they are fitted with low-order polynomials and simply taken out unless any sort of stacking or time-series analysis discussed below is applied. While this procedure works to eliminate the artifacts due to orbit errors, it will also take out any long-wavelength geophysically interesting signals such as the inter-seismic, post-seismic, ocean tidal loading, solid-Earth tide, and post-glacial rebound signals. Alternatively, if the ground control points (GCP) are available, where the precision ground deformation data are available, we can reestimate the baseline, based on those GCP data (e.g., Rosen et al., 1996), but such data are often unavailable in remote areas.

One approach to correct for the tropospheric delay signals is to employ the other independent estimates derived from either the GPS-based tropospheric delay estimates (e.g., Onn and Zebker, 2006) or the output results from high-resolution numerical weather model (e.g., Foster et al., 2006). These so-called calibration approaches are, however, not easily applicable. The dense ground-based GPS network is limited to a few areas in the world. Also, high-resolution numerical weather model still needs significant computational resources.

Besides the tropospheric delay problem, the effects of ionosphere on both interferograms and pixel-offset images were clearly recognized in the results of the 2008 Wenchuan earthquake (Mw7.9), China, derived from ALOS/PALSAR (Kobayashi et al., 2009; Raucoules and de Michele, 2010), although they were pointed out in polar region many years ago (e.g., Matter and Gray, 2002). It is well known that the lower the carrier frequency is, the more significant the ionospheric dispersion impacts on the propagation delay. Thus, in many of the previous applications of C-band SAR data, the effects of ionosphere could have been neglected. While GPS also employs L-band, the high-precision GPS geodetic survey corrects for the ionospheric effect with the use of dual frequency, L1 and L2, observation data. In contrast, PALSAR is a single frequency SAR sensor and incapable of the standard ionosphere-correction approach. Empirically, however, we will encounter the ionospheric signals more frequently in the ascending data acquired in the local nighttime than in the descending data acquired in the local daytime. We also recall that the JERS, the other L-band SAR operated during 1992–1998, did not reveal any significant ionospheric signals at least in mid-latitude regions, and that most of the JERS data were acquired in the descending track. Besides the latitude, the effects of ionosphere on SAR image might, therefore, significantly depend on the data acquisition time. Like the tropospheric effects, detailed studies of ionospheric impacts on the SAR data are also currently underway.

A simple approach to eliminate those noises is *stacking*, in which several interferograms are stacked to isolate small-amplitude signals, because those noises can be regarded as temporally random, whereas the deformation signals are spatially persistent. Two important prerequisites for successful stacking are: (1) the data acquisition dates of those interferograms should not be overlapped, in order not to enhance the noises of any particular acquisition date(s), and (2) each temporal baseline should be as long as possible so that each interferogram can include as much deformation signals as possible. In reality, it is not easy to gather many independent interferograms that have desirably long temporal baselines because the available data often encounter the spatial and temporal decorrelation. Also, the simple stacking approach inherently assumes temporally linear evolution in the ground deformation, preventing us from deriving time-series data.

Time-Series Analysis: Ferretti et al. (2000, 2001) proposed a new analysis technique called *Permanent Scatterer InSAR* (*PS-InSAR*), in which they take advantage of even such data pairs whose spatial baselines are longer than the critical values. Thereby, they could expand the temporal coverage, and thus could estimate the long-term deformation signals on the order of mm/year. Key idea of PS-InSAR is to pick up only such pixels that will exhibit long-term coherence due to the existence of corner-reflector-like targets, which Ferretti et al. called "permanent scatterers." Based on those pixels alone, they generate a stack of differential interferograms, using available DEM and orbit data. The phase values include not only deformation signals, but also such topographic signals that were not initially taken into account, because the longer spatial baseline pairs are so sensitive to the topography that the available DEM could not account for. In PS-InSAR and its variants (Werner et al., 2003; Hooper et al., 2004), they fit the differential interferogram stack to a phase model that describes not only temporal evolution of deformation but also corrections to the available DEM. Deviations from the phase model can be filtered into either non-linear deformation or atmospheric signals because the former signals are correlated and thus low-pass filtered along temporal axis, while the latter signals are temporally random; the orbit data must be assumed to be correct. A known limitation of PS-InSAR is its rather lower sampling density over non-urban areas. However, despite a lack of man-made objects, Furuya et al. (2007) succeeded in detecting active salt tectonic motion, applying a similar technique to Canyonlands National Park, Utah, presumably because the area was non-vegetated and the exposed surface rocks behaved like corner-reflector-like targets.

Another time-series analysis approach was devised and known as small baseline subset (*SBAS*) algorithm (Berardino et al., 2002). Key idea of the SBAS algorithm is *least-squares inversion* of unknown deformation at each SAR data acquisition epoch, based on the available unwrapped differential interferograms (e.g., Lundgren

et al., 2001; Schmidt and Burgmann, 2003). Using small baseline interferometric pairs, the SBAS approach is free from spatial decorrelation and allows us to take advantage of the fine spatial resolution of InSAR data. If the number of interferograms is greater than or equal to the number of SAR acquisitions, the inversion problem becomes an over-determined or well-determined problem, and can be easily solved in a least-squares approach. It is uncommon, however, that all the available interferometric pairs have short baselines, and accordingly the temporal sampling rate will decrease. Berardino et al. (2002) proposed to employ several groups of "small baseline subset" to overcome the lower temporal resolution issue, and solved the rank-deficient problem with the use of singular value decomposition (SVD) technique. The SVD gives the minimum-norm least-squares solution, which is equivalent to minimizing the estimated velocities at any time intervals.

Time-series analysis of SAR data is a promising technique, but almost all previous analyses are based on the C-band ERS1/2 and Envisat data, because not only more-than-decade-long data but also high-precision, well-controlled satellite orbits are available for these satellites. As noted before, not all geophysically interesting phenomena could be detected by C-band and shorter-wavelength SAR data. If the L-band ALOS/PALSAR data are archived for a much longer time, and if the follow-on ALOS-2 and the DESDynI are launched as scheduled, the time-series analysis of SAR data will become feasible even in areas that have never been monitored before. The time-series analysis with ScanSAR data should also be possible. Long-term continuous monitoring with L-band SAR will provide us with more opportunities for new discoveries.

Bibliography

Bamler, R., and Hartl, P., 1998. Synthetic aperture radar interferometry. *Inverse Problems*, **14**, R1.

Berardino, P., Fornaro, G., Lanari, R., and Sansosti, E., 2002. A new algorithm for surface deformation monitoring based on small baseline differential SAR interferograms. *IEEE Transactions on Geoscience Remote Sensing*, **40**, 2375.

Burgmann, R., Rosen, P. A., and Fielding, E. J., 2000. Synthetic aperture radar interferometry to measure Earth's surface topography and its deformation. *Annual Review of Earth and Planetary Sciences*, **28**, 169.

Cloude, S. R., and Papathanassiou, K. P., 1998. Polarimetric SAR Interferometry. *IEEE Transactions on Geoscience Remote Sensing*, **36**, 1551.

Cumming, I. G., and Wong, F. H., 2005. *Digital Processing of Synthetic Aperture Radar Data: Algorithm and Implementation*. Boston: Artech House.

Curlander, J. C., and McDonough, R. N., 1991. *Synthetic Aperture Radar: Systems and Signal Processing*. New York: Wiley interscience.

Farr, T. G., et al., 2007. The shuttle radar topography mission. *Reviews of Geophysics*, **45**, RG2004.

Ferretti, A., Prati, C., and Rocca, F., 2000. Nonlinear subsidence rate estimation using permanent scatterers in differential SAR interferometry. *IEEE Transactions on Geoscience and Remote Sensing*, **38**, 2202.

Ferretti, A., Prati, C., and Rocca, F., 2001. Permanent scatterers in SAR interferometry. *IEEE Transactions on Geoscience and Remote Sensing*, **39**, 8.

Fialko, Y., Simons, M., and Agnew, D., 2001. The complete (3-D) surface displacement field in the epicentral area of the 1999 Mw7.1 hector mine earthquake, California, from space geodetic observations. *Geophysical Research Letters*, **28**, 3063.

Foster, J., Brooks, B., Cherubini, T., Shacat, C., Businger, S., and Werner, C. L., 2006. Mitigating atmospheric noise for InSAR using a high resolution weather model. *Geophysical Research Letters*, **33**, L16304.

Furuya, M., Mueller, K., and Wahr, J., 2007. Active salt tectonics in the Needles District, Canyonlands (Utah) as detected by interferometric synthetic aperture radar and point target analysis: 1992–2002. *Journal of Geophysical Research*, **112**, B06418.

Ghilia, D. C., and Pritt, M. D., 1998. *Two Dimensional Phase Unwrapping: Theory, Algorithms, and Software*. New York: Wiley.

Hanssen, R. F., 2001. *Radar Interferometry: Data Interpretation and Error Analysis*. Dordrecht: Kluwer.

Hooper, A., Zebker, H., Segall, P., and Kempes, B., 2004. A new method for measuring deformation on volcanos and other natural terrains using InSAR persistent scatterers. *Geophysical Research Letters*, **31**, L23611.

Kobayashi, T., Takada, Y., Furuya, M., and Murakami, M., 2009. Location and types of ruptures involved in the 2008 Sichuan Earthquake inferred from SAR image matching. *Geophysical Research Letters*, **36**, L07302.

Lundgren, P., Usai, S., Sansosti, E., Lanari, R., Tesauro, M., Fornaro, G., and Berardino, P., 2001. Modeling surface deformation observed with synthetic aperture radar interferometry at Campi Flegrei caldera. *Journal of Geophysical Research*, **106** (B9), 19355.

Massonnet, D., and Feigl, K. L., 1998. Radar interferometry and its application to changes in the earth's surface. *Reviews of Geophysics*, **36**, 331.

Massonnet, D., Rossi, M., Carmona, C., Adragna, F., Peltzer, G., Feigl, K., and Raboute, T., 1993. The displacement field of the Landers earthquake mapped by radar interferometry. *Nature*, **364**, 138.

Massonnet, D., Vadon, H., and Rossi, M., 1996. Reduction of the need for phase unwrapping in radar interferometry. *IEEE Transactions on Geoscience and Remote Sensing*, **34**, 489.

Matter, K. E., and Gray, A. L., 2002. Reducing ionospheric electron density errors in satellite radar interferometry applications. *Canadian Journal of Remote Sensing*, **28**, 583.

Michel, R., Avouac, J.-P., and Taboury, J., 1999. Measuring ground displacements from SAR amplitude images: application to the landers earthquake. *Geophysical Research Letters*, **26**, 875.

Motagh, M., Wang, R., Walter, T. R., Bürgmann, R., Fielding, E., Anderssohn, J., and Zschau, J., 2008. Coseismic slip model of the 2007 august pisco earthquake (Peru) as constrained by wide swath radar observations. *Geophysical Journal International*, **174**, 842.

Onn, F., and Zebker, H. A., 2006. Correction for interferometric synthetic aperture radar atmospheric phase artifacts using time series of zenith wet delay observations from a GPS network. *Journal of Geophysical Research*, **111**, B09102.

Pritchard, M. E., 2006. InSAR, a tool for measuring Earth's surface deformation. *Physics Today*, **59**(7), 68.

Raucoules, D., and de Michele, M., 2010. Assessing Ionospheric Influence on L-Band SAR Data: Implications on Coseismic Displacement Measurements of the 2008 Sichuan Earthquake. *IEEE Geoscience Remote Sensing Letters*, **7**, 286.

Rosen, P. A., Hensley, S., Zebker, H. A., Webb, F. H., and Fielding, E. J., 1996. Surface deformation and coherence measurements of

Kilauea volcano, Hawaii, from SIR-C radar interferometry. *Journal of Geophysical Research*, **101**(E10), 23109.
Schmidt, D. A., and Burgmann, R., 2003. Time-dependent land uplift and subsidence in the Santa Clara valley, California, from a large interferometric synthetic aperture radar data set. *Journal of Geophysical Research*, **108**(B9), 2416.
Simons, M., and Rosen, P. A., 2007. Interferometric synthetic aperture radar geodesy. In Herring, T. A. (ed.), *Treatise on Geophysics*. New York: Elsevier, Vol. 3, pp. 391–446.
Tobita, M., Murakami, M., Nakagawa, H., Yarai, H., and Fujiwara, S., 2001a. Two-dimensional field of three-dimensional components of deformations and velocities, and volume change around Usu Volcano associated with the 2000 eruption by matching of SAR images (in Japanese). *Journal of Geographical Survey Institute*, **95**, 37.
Tobita, M., Murakami, M., Nakagawa, H., Yarai, H., Fujiwara, S., and Rosen, P. A., 2001b. 3D surface deformation of the 2000 Usu Eruption measured by matching of SAR images. *Geophysical Research Letters*, **28**, 4291.
Werner, C. L., Wegmuller, U., Strozzi, T., and Wiesmann, A., 2003. Interferometric point target analysis for deformation mapping, paper presented at IGARSS'03, Toulouse: Geoscience Remote Sensing Society.
Zebker, H. A., Rosen, P. A., Goldstein, R. M., Gabriel, A., and Werner, C. L., 1994. On the derivation of coseismic displacement fields using differential radar interferometry: The Landers earthquake. *Journal of Geophysical Research*, **99**(B10), 19617–19634.
Zhou, X. B., Chang, N. B., and Li, S. S., 2009. Applications of SAR Interferometry in Earth and Environmental Science Research. *Sensors*, **9**, 1876.

Cross-references

Earthquakes and Crustal Deformation
Earthquake Rupture: Inverse Problem
Earthquakes, Source Theory
Geodesy, Ground Positioning and Leveling
GPS, Tectonic Geodesy
Inverse Theory, Global Optimization
Inverse Theory, Linear
Inverse Theory, Monte Carlo Method
Inverse Theory, Singular Value Decomposition
Remote Sensing and GIS Techniques for Tectonic Studies
Remote Sensing, Applications to Geophysics
Slow Earthquake

SATELLITE LASER RANGING

David Coulot[1], Florent Deleflie[2], Pascal Bonnefond[3], Pierre Exertier[3], Olivier Laurain[3], Bertrand de Saint-Jean[3]
[1]Laboratoire de Recherche en Géodésie, Institut Géographique National, GRGS Ecole Nationale des Sciences Géographique, Marne la Vallée Cedex 2, France
[2]Institut de Mécanique Céleste et de Calcul des Ephémérides, GRGS, Paris, France
[3]Observatoire de la Côte d'Azur, Géoazur, GRGS, Grasse, France

Definition

FTLRS. Acronym for French transportable laser ranging system (Nicolas et al., 2001).

ILRS. Acronym for International laser ranging service. International service established in 1998 (Pearlman et al., 2002). See entry *Geodesy, Networks and Reference Systems*.

Laser. Acronym for light amplification by stimulated emission of radiation.

Least-squares method. Optimization method. See (Tarantola, 2004).

LLR. Acronym for lunar laser ranging. See (Bender et al., 1973).

Introduction

The Satellite Laser Ranging (SLR) technique was born in 1964, with the first returns obtained from the satellite Beacon Explorer B. Since the launch of the first dedicated laser satellite Starlette in 1975 (Schutz et al., 1993; Barlier and Lefebvre, 2001), SLR has provided range measurements which have been the foundations of, not only precise orbit computations, but also major geodetic products, fundamental for many Geosciences applications. One of the advantages is that the onboard equipment is light, cheap, has an infinite lifetime, and does not consume any energy.

Technique

Principle and instrumentation

SLR is based on the measurement of the round-trip time of flight of laser pulses, between a ground station and a satellite orbiting round the Earth. As the measurements are carried out in the visible (or near visible) spectrum, the technique is dependent on weather conditions; it does not work under a cloudy sky. The satellite is equipped (if not entirely covered) with reflectors specially designed to reflect the laser in the incident direction. The ground station is composed of several instruments: a laser, a telescope for the emission and/or the reception of light, a detection system (for the start and the return of the laser), an event timer, a timing system for the chronometry, a frequency standard, a measurement calibration system (required because of the instrumental instabilities), a weather station (for the atmospheric correction, see Sect. Data analysis and orbitography), and a completely computerized system to pilot the telescope during the satellite tracking, to process the measurements, and to send the data obtained.

This complex ground instrumentation requires a permanent staff to (1) maintain the instruments and, consequently, to guarantee a stable quality of the measurements and to (2) track the satellites. This constant manpower requirement and the need for a cloudless sky to track satellites are drawbacks of the technique, even if some stations are now on the way toward automation. Moreover, these requirements make the SLR technique different (but not less effective, far from it) from the other satellite techniques: Doppler Orbitography and Radiopositioning Integrated by Satellite (DORIS), Global

Positioning System (GPS) – see entry *GPS, Data Acquisition and Analysis*, GLObal NAvigation Satellite System (GLONASS), and the future Galileo (Europe) and Compass (China) systems. Indeed, these systems (will) operate whatever the weather conditions and their complexity lies more in the spatial segment.

Data

The raw observations produced during a satellite pass over a station are not directly put at the disposal of the scientific community. Indeed, with nearly 10 laser pulses per second, a significant number of returns can be obtained. Roughly speaking, these latter are thus compressed to deduce the so-called "normal points," gathering all the information contained in full-rate points, over specific spans of time, depending on the satellite altitude. The ILRS recommended algorithm is available at http://ilrs.gsfc.nasa.gov/products_formats_procedures/normal_point/np_algo.html. Normal point files per satellite can be retrieved through the ILRS data centers at ftp://cddis.gsfc.nasa.gov/slr/ and http://www.dgfi.badw-muenchen.de/edc/edc.html.

Technological evolution

Since the early seventies, the SLR technology and, consequently, the measurement quality have constantly improved, starting at a 1.5 m precision level, with a significant step toward the millimeter accuracy reached in the nineties (Degnan, 1993). Presently, a typical laser configuration is (1) a wavelength of 532 nm (green), (2) a pulse width ranging from 35 to 200 ps, and (3) a fire rate of 5, 10, or 20 Hz. The overall accuracy of the measurements is assessed at the 5 mm level for a few stations – see (Exertier et al., 2006) who also provide a complete error budget of the technique.

The principal technological improvements under study are the following (Noll and Pearlman, 2007, 6-1): reduction of the pulse length to 10 ps to improve the single shot precision; kHz laser ranging to improve the normal point precision (by providing a larger number of averaged raw data per normal point) and the speed of target acquisition during satellite tracking, together with more precise event timers (precision of 3 ps or less); multi-wavelength ranging to test possible improvements of the atmospheric correction.

Satellite constellation

Passive laser satellites have been developed for geodesy purposes. Their low area to mass ratios minimize the amplitude of the surface forces (atmospheric drag, for instance). The SLR network first tracks dedicated satellites, which have a practically infinite lifetime. The two first SLR satellites were Starlette, developed by CNES (France) in 1975, with a 812 km perigee height, and LAGEOS, launched by NASA (United States) in 1976, with a 5,860 km perigee height. They were followed by similar satellites in the eighties and in the nineties: Ajisai (1986, Japan, 1,490 km), Etalon-1 and -2 satellites (1989, USSR, 19,120 km), LAGEOS-2 (1992, United States and Italy, 5,620 km), and Stella (1993, France, 800 km). The constellation of SLR geodetic satellites is now composed of orbits with a wide range of altitudes and inclinations to the equator, providing various dynamical configurations. Some new SLR-only missions are under investigation, such as the LAGEOS III mission, for purposes concerning fundamental physics (Ciufolini, 1994).

The network also tracks (or has tracked) non-specific SLR satellites equipped with laser retroreflectors: altimetric (TOPEX/Poséidon, Jason-1, Jason-2), remote sensing (ERS-1 and -2, ENVISAT), gravimetric (CHAMP, GRACE-A and -B, GOCE), and GNSS satellites (some GPS, GLONASS, and Galileo – GIOVE-A and -B – satellites).

Currently, 34 artificial satellites are being tracked by the network. Moreover, few stations, equipped with large telescopes have the capability to measure the range to the Moon (LLR technique). One station is located in Europe (Grasse, France, telescope with a 1.6 m aperture), and two others are located in North America: at Apache point, in southern New Mexico, where the Apollo project utilizes a 3.5 m telescope, and at the McDonald Observatory, University of Texas (75 cm). See http://ilrs.gsfc.nasa.gov/satellite_missions/list_of_satellites/.

Ground network

Starting with five stations in the early seventies, the network configuration evolved over the last 35 years and the number of stations has been stable for several years. The current network is composed of about 30 stations (see http://ilrs.gsfc.nasa.gov/stations), distributed among 20 countries or so. Most of these stations observe only during night, a few having as well the capability of observing in daylight (useful for Global Navigation Satellite System – GNSS – satellite tracking, see Sect. Data analysis and orbitography). The distribution of this network, as is the case for the Very Long Baseline Interferometry (VLBI) network (see entry *Very Long Baseline Interferometry*), is generally admitted as a weakness of the technique. Indeed, in the northern hemisphere, most stations are located at mid-latitudes; in the southern hemisphere, there are only few stations (5, typically). Moreover, the densification and the maintenance of an ideal network are dependent on political decisions and the economic context (because of the financial commitments required to both build and maintain a SLR station).

This situation should improve in the near future, thanks to projects such as the NASA SLR2000 project (completely automated, eye-safe, highly reliable, and low-cost SLR stations), recently relaunched (McGarry and Zagwodski, 2008) and now called NGSLR, or thanks to mobile systems (FTLRS, for instance).

Data analysis and orbitography

General principle

A typical data analysis follows this two-step sketch, which is more or less the same for GPS and DORIS data analyses:

1. *Orbit computation, per satellite*. The physical forces acting on the satellite motion account for gravitational (Earth's gravity field, including the time-varying part, luni-solar attraction) and non-gravitational perturbations (mainly atmospheric drag and radiation pressure). The equations of motion are numerically propagated to provide time series of a state vector of orbital elements. The theoretical range rate is then computed, on the basis of these time series, and on geodetic products such as the Terrestrial Reference Frame (TRF), Earth Orientation Parameters – EOP, and some station displacement models. These "theoretical observations" are finally compared, in a least-squares sense, with the real data to estimate some dynamical parameters (initial satellite state vector, drag coefficients, etc.). Moreover, empirical accelerations can be estimated to compensate for some part of the modeling deficiencies (Crétaux et al., 1994) and, consequently, to improve the post-fit residual level. The computations are iterated until a given stop condition is fulfilled.
2. *Geodetic parameter estimation*. The orbital arcs are simultaneously used to derive predicted range measurements, together with a priori values of the parameters to be estimated and required models. Updates of the geodetic parameters (station positions, EOP, etc.), and generally of some of the dynamical parameters estimated during step 1, are then estimated in a least-square sense. If necessary, during this step, the measurements can be gathered for several different satellites and over a long period of time.

Most of the models applied (station displacements, Earth's gravity field, atmospheric density, etc.) are not specific to the SLR data processing. Their list is provided in the International Earth Rotation and Reference Systems Service (IERS) conventions (McCarthy and Petit, 2004).

Specific corrections

As they are carried out in the visible spectrum, SLR laser beams are not affected by the ionosphere crossing, in comparison to the radio-electric signals. But the tropospheric propagation still limits their accuracy. Recently, an improved model of zenith delay has been developed (Mendes and Pavlis, 2004). This model, together with an improved mapping function (Mendes et al., 2001), is currently the standard for SLR data analyses. This correction takes typical values of several meters and its inaccuracy is still assessed at a few-mm level (Exertier et al., 2006). In order to reach the 1-mm accuracy in the near future, improvements are under study; for example, ray tracing (Hulley and Pavlis, 2007) and multi-wavelength systems (Hamal et al., 2008).

The temporal spread of returned laser pulses due to the reflection from multiple reflectors on the satellite is another significant error at the 1-mm level. To better handle these effects, (Otsubo and Appleby, 2003) have recently proposed system-dependent center-of-mass corrections for spherical satellites (LAGEOS, Ajisai, and Etalon), with variations up to 10 mm. Efforts are still in progress in order to better understand and take these effects into account in SLR analyses (Luck et al., 2008, Sect. 8).

Despite the systematic in situ calibrations, ranges can be affected by systematic errors, dependent on the instrumentation: time and range biases. Due to their strong correlation with station heights, much attention is paid to range biases and their temporal variations. They are indeed constantly monitored in the framework of the ILRS activities – cf. (Otsubo et al., 2009) for instance. Moreover, analysis efforts towards better estimations of these biases are in progress, see (Appleby et al., 2009) as an example. Typical bias values range from a few millimeter to centimetric values. These biases can be handled in different ways (applied and/or estimated) during any data processing.

Finally, in a relativistic context, the propagation of light between a ground station and a near-Earth satellite induces a supplementary time delay with respect to the purely geometric path (Ries et al., 1988). This correction takes few-mm values.

LAGEOS satellites

Due to their relatively high altitudes (about 6, 000 km), both LAGEOS satellites are essential targets and their orbits are permanently tracked and computed with great care. Since their launch, the quality of the computed orbits has evolved with the technique: from a 50 cm precision in 1976 (Exertier et al., 2006) to a few centimeters in the mid nineties (Marshall et al., 1995) to nearly 1 cm nowadays. Additionally, various non-gravitational effects have been evidenced in the LAGEOS-1 orbit: solar radiation pressure, Yarkovsky–Shah thermal effect, asymmetric reflectivity of the satellite surface, and asymmetric thermal emissivity of the Earth (Métris et al., 1999).

GNSS satellites

Since the first study (Degnan and Pavlis, 1994), and sometimes through dedicated campaigns such as IGEX-98 (Barlier et al., 2001), SLR has always contributed to the validation of GNSS orbits. This validation generally consists in the comparison between SLR range measurements to GNSS satellites and the predicted ranges computed with the GNSS-only orbits. According to (Urschl et al., 2007), the consistency (assessed over nearly four years) between SLR and GPS (resp. GLONASS) is at an accuracy level of −3 cm (resp. −3 mm) with a precision at the level of 2.5 cm (resp. 5 cm). Further studies are in progress to get an exhaustive understanding of the inter-technique differences. Furthermore, all satellites of the future GNSS

Galileo system with be equipped with Laser Ranging Arrays (LRA) for such orbit validations.

Applications to geodesy and geodynamics

Earth's gravity field

Before the launch of dedicated space missions (CHAMP, GRACE, and GOCE), the gravity field models were mainly built on the basis of SLR data acquired on the whole constellation, providing various dynamical configurations (semi-major axes, inclinations to the equator), and, consequently, various sensitivities to the mass distribution within the Earth's system. The GRIM5-S1 model (Biancale et al., 2000), one of the latest SLR-only based, was prepared in a joint German–French effort. The solution was obtained by solving simultaneously for the gravitational and ocean tide potentials and tracking station positions. Nowadays, long wavelengths of the gravity field are still deduced from SLR data, accurate enough to provide a value of the Earth's gravitational constant with 10 significant digits (corresponding to an orbital post-fit residual error of the order of a few centimeters), see (Tapley et al., 1993). Since the CHAMP mission was launched in 2000, followed by the GRACE mission in 2002, and GOCE in 2009, a new generation of nearly monthly Earth's gravity field models was initiated, benefiting both from the high sensitivity of the orbit (a few hundred km high) to regional variations and from accelerometric data, measuring the non-gravitational forces (mainly atmospheric drag). The GRACE-only models expanded in spherical harmonics provide coefficients up to degree and order 360. See (Reigber et al., 2003; Lemoine et al., 2007; Tapley et al., 2005) and entries *Gravity Field of the Earth* and *Gravity Field, Temporal Variations from Space Techniques*.

Mass variations within the Earth's dynamic system have a temporal spectrum ranging from hours to decades, and even longer, that can be evidenced by SLR data, because they are available over a very long period of time. Many of them are related to both long-term and short-term climate forcing, and are linked to mass redistribution in the atmosphere, ocean, and continental water (Chen et al., 2000; Cheng and Tapley, 2004). By the next years, SLR data, moreover, are likely to provide an unique opportunity to continue to monitor the Earth's gravity field, after the end-of-life of the CHAMP, GRACE, and GOCE missions.

Terrestrial reference frame and Earth's rotation

Since the first computation of a combined TRF in 1984 (called BTS84), SLR has always been a keystone of the BTS (BIH – Bureau International de l'Heure – Terrestrial System) and, after 1988, of the International Terrestrial Reference System (ITRS, see entry *Geodesy, Networks and Reference Systems*) realizations (McCarthy and Petit, 2004, Chap. 4). Indeed, their origin and scale definitions have always strongly involved this technique. Parallel to the SLR-positioning precision, starting with a decimetric precision (BTS84), the successive ITRF versions have turned 1 cm-precision level with the ITRF96. For the more recently published version (ITRF2005), the whole precision of positions (resp. velocities) is under 2 mm (resp. 1 mm/y). And, for the first time of the ITRF history, the ITRF2005 was computed with times series of (weekly for SLR) station positions and (daily for SLR) EOP. Regarding positioning, the weekly WRMS for the well performing SLR stations is 5 mm for both planimetry and height (Altamimi et al., 2007). For comparison, the weekly WRMS for GPS are 2 mm (resp. 5 mm) for planimetry (resp. for height). It must be noted that the most recent ITRS realization (ITRF2008) is now available (http://itrf.ign.fr/ITRF_solutions/2008).

Regarding Earth's rotation, SLR provides (with the LAGEOS data) the longest space-geodetic time series of EOP (Gross, 2009). Starting with a 10 milliarcsecond (mas) precision in the seventies, the discrepancy between the SLR and the IERS polar motion series was assessed at the level of 0.3 mas in the mid 2000 years (Gambis, 2004). (Coulot et al., 2010) have recently showed that significant precision improvements are still possible with the help of a rigorous referencing.

Global Earth's deformations

For many years, SLR has provided unique and inestimable information about the geocenter motion (Pavlis and Kuźmicz-Cieślak, 2009a). The geocenter motion is the 3-dimensional motion of the instantaneous Earth's center-of-mass with respect to any conventional reference (the ITRF secular origin, for instance). It is related to the responses of the Earth to the displacements of its fluid masses (atmosphere, ocean, etc.) and its temporal variations are dominated by annual signals of few mm amplitudes. (Collilieux et al., 2009) provide an exhaustive review about this geodynamical signal and, moreover, assess the effect of the SLR network on its determination.

The ITR2005 computation has been a good opportunity to study the residual position time series which contain the non-linear part of the station motions. Regarding SLR station positions, most of the time series evidence annual signals. As surface loading effects (see van Dam et al. (2002) and http://www.sbl.statkart.no/) are not currently modeled in the routine data processing, they are supposed to be evidenced by the position series. (Collilieux et al., 2010) show that an a posteriori correction for these effects with a geodynamical model leads to a reduction of the amplitudes of the annual signals for a majority of stations. Furthermore, (Pavlis et al., 2009) strongly recommend the modeling of such effects in the future SLR data processing to achieve a better accuracy.

Altimetry and sea level monitoring

Stable and accurate TRF, satellites orbits, and altimeter measurements are prerequisites for robust altimetric Mean Sea Level (MSL) computations over large time periods

(Leuliette et al., 2004; Beckley et al., 2007; Ablain et al., 2009). Indeed, these conditions are crucial to connect the data provided by successive missions. Since the advent of altimetry (launch of TOPEX/Poséidon in 1992), SLR has contributed to all these efforts, first regarding the TRF (see Sect. Terrestrial reference frame and Earth's rotation). SLR has also played a major role in the Precision Orbit Determination (POD) framework, additionally to DORIS and GPS data. For instance, the technique has recently contributed to the achievement of a 1-cm radial accuracy for the orbit of Jason-1 (Luthcke et al., 2003). In this context, the laser-based short-arc technique (Bonnefond et al., 1999) is also a major method to assess the radial accuracy of the altimetric satellites on an operational basis. Finally, thanks to thoroughly instrumented sites, it is possible to monitor both bias and drift of the radar altimeters, by taking advantage of direct overflight passes (see http://www.gavdos.tuc.gr/intro.jsp for an illustration of the calibration principle). The French campaigns led in Corsica with the FTLRS show that the quality of such a calibration mainly depends on the accurate positioning (mainly deduced from Starlette and Stella data) of the mobile system and on the stability of its range bias (Gourine et al., 2008).

Summary

Observing and understanding the system Earth require numerous and accurate geodetic measurements over both spatial and temporal wide spectra. Since the early seventies, the SLR technique has provided such measurements, inestimable for the progress of Space Geodesy and, consequently, of Earth's sciences. Nowadays, in the Global Geodetic Observing System (GGOS) context, and, more particularly, with the stringent requirements on the ITRF accuracy (1 mm and 0.1 mm/y) and the crucial need of taking advantage of the features of each space-geodetic technique, SLR still has a major role to play.

Moreover, to reach the ambitious goals fixed for the near future, this technique is currently renewed. From a technological point of view, the kHz ranging has recently showed the wide spectrum of its abilities, see the dedicated sections in (Luck et al., 2008; Schillak, 2009). Regarding the ground and spatial segments, the design of the ideal SLR network of the future is in progress (Pavlis and Kuźmicz-Cieślak, 2009b), as is the design of the next generation of SLR stations, and new satellite missions are already planned. Concerning the data analysis, new modelings should help to improve the accuracy of the products (Pavlis et al., 2009). More particularly, the unprecedented quality of the Earth's gravity field models reached with the GRACE (Tapley et al., 2004) and GOCE missions should allow the use of low-orbiting satellite (Starlette, Stella, and Ajisai) data for station position computations (Lejba and Schillak, 2009).

Finally, new fundamental objectives have recently been fixed for the SLR technique in the time (frequency) domain, though time transfer projects (Samain et al., 2009) and interplanetary navigation (Ni et al., 2002).

Bibliography

Ablain, M., Cazenave, A., Valladeau, G., and Guinehut, S., 2009. A new assessment of the error budget of global mean sea level rate estimated by satellite altimetry over 1993–2008. *Ocean Science*, **5**, 193–201.

Altamimi, Z., Collilieux, X., Legrand, J., Garayt, B., and Boucher, C., 2007. ITRF2005: a new release of the international terrestrial reference frame based on time series of station positions and earth orientation parameters. *Journal of Geophysical Research*, **112**, B09401.

Appleby, G., Wilkinson, M., Luceri, V., Gibbs, P., and Smith, V., 2009. Attempts to separate apparent observational range bias from true geodetic signals. In Schillak, S. (ed.), *Proceedings of the 16th International Workshop on Laser Ranging*, Poznán, Poland, October 13–17, 2008, Space Research Centre, Polish Academy of Sciences.

Barlier, F., Berger, C., Bonnefond, P., Exertier, P., Laurain, O., Mangin, J. F., and Torre, J. M., 2001. Laser-based validation of GLONASS orbits by short-arc technique. *Journal of Geodesy, Special issue IGEX-98*, **11**, 600–612.

Barlier, F., and Lefebvre, M., 2001. A new look at planet Earth: satellite geodesy and geosciences. In *The Century of Space Science*. The Netherlands: Kluwer, pp. 1623–1651.

Beckley, B. D., Lemoine, F. G., Luthcke, S. B., Ray, R. D., and Zelensky, N. P., 2007. A reassessment of global and regional mean sea level trends from TOPEX and Jason-1 altimetry based on revised reference frame and orbits. *Geophysical Research Letters*, **34**, L14608.

Bender, P. L., Currie, D. G., Dicke, R. H., Eckhardt, D. H., Faller, J. E., Kaula, W. M., Mulholland, J. D., Plotkin, H. H., Poultney, S. K., Silverberg, E. C., Wilkinson, D. T., Williams, J. G., and Alley, C. O., 1973. The lunar laser ranging experiment. *Science*, **182**, 229–238.

Biancale, R., Balmino, G., Lemoine, J. M., Marty, J. C., Moynot, B., Barlier, F., Exertier, P., Laurain, O., Gegout, P., Schwintzer, P., Reigber, C., Bode, A., Gruber, T., König, R., Massmann, F. H., Raimondo, J. C., Schmidt, R., and Zhu, S. Y., 2000. A new global earth's gravity field model from satellite orbit perturbations: GRIM5-S1. *Geophysical Research Letters*, **27**, 3611–3614.

Bonnefond, P., Exertier, P., and Barlier, F., 1999. Geographically correlated errors observed from a laser-based short-arc technique. *Journal of Geophysical Research*, **104**(C7), 15885–15893.

Chen, J. L., Wilson, C. R., Eanes, R. J., and Tapley, B. D., 2000. A new assessment of long-wavelength gravitational variations. *Journal of Geophysical Research*, **105**(B7), 16271–16277.

Cheng, M. K., and Tapley, B. D., 2004. Variations in the earth's oblateness during the past 28 years. *Journal of Geophysical Research*, **109**(B9), B09402.

Ciufolini, I., 1994. Gravitomagnetism and status of the LAGEOS III experiment. *Classical and Quantum Gravity*, **11**(6A), 73–81.

Collilieux, X., Altamimi, Z., Ray, J., van Dam, T., and Wu, X., 2009. Effect of the satellite laser ranging network distribution on geocenter. *Journal of Geophysical Research*, **114**, B04402.

Collilieux, X., Altamimi, Z., Coulot, D., van Dam, T., and Ray, J., 2010. Impact of loading effects on determination of the international terrestrial reference frame. *Advances in Space Research*, **45**, 144–154.

Coulot, D., Pollet, A., Collilieux, X., and Berio, P., 2010. Global optimization of core station networks for space geodesy: application to the referencing of the SLR EOP with respect to ITRF. *Journal of Geodesy*, **84**, 31–50.

Crétaux, J. F., Nouel, F., Valorge, C., and Janniere, P., 1994. Introduction of empirical parameters deduced from the Hill's equations for satellite orbit determination. *Manuscripta Geodaetica*, **19**, 135–156.

Degnan, J. J., 1993. Millimeter accuracy satellite laser ranging: a review. In: *Contributions of Space Geodesy to Geodynamics: Technology, Geodynamic Series*. Washington, D. C.: AGU, Vol. 25, pp. 133–162.

Degnan, J. J., and Pavlis, E. C., 1994. Laser ranging to GPS satellites with centimeter accuracy. *GPS World*, **5**(9), 62–70.

Exertier, P., Bonnefond, P., Deleflie, F., Barlier, F., Kasser, M., Biancale, R., and Ménard, Y., 2006. Contribution of laser ranging to earth's sciences. *Comptes Rendus Geoscience*, **338**, 958–967.

Gambis, D., 2004. Monitoring earth orientation using space geodetic techniques: state-of-the-art and prospective. *Journal of Geodesy*, **78**(4–5), 295–303.

Gourine, B., Kahlouche, S., Exertier, P., Berio, P., Coulot, D., and Bonnefond, P., 2008. Corsica SLR positioning campaigns (2002 and 2005) for satellite altimeter calibration missions. *Marine Geodesy*, **31**, 103–116.

Gross, R. S., 2009. Laser ranging contributions to earth rotation studies. In Schillak, S. (ed.), *Proceedings of the 16th International Workshop on Laser Ranging*, Poznán, Poland, October 13–17, 2008, Space Research Centre, Polish Academy of Sciences.

Hamal, K., Prochazka, I., Blazef, J., Fumin, Y., Jingfu, H., Zhongping, Z., Kunimori, H., Greene, B., Kirchner, G., Koidl, F., Riepfel, S., and Gurtner, W., 2008. Multi color satellite laser ranging at Czech technical university. In Luck et al. (eds.), *Proceedings of the 15th International Workshop on Laser Ranging*, Canberra, Australia, October 16–20, 2006, EOS Space Systems Pty. Ltd.

Hulley, G. C., and Pavlis, E. C., 2007. A ray-tracing technique for improving satellite laser ranging atmospheric delay corrections, including the effects of horizontal refractivity gradients. *Journal of Geophysical Research*, **112**, B06417.

Lejba, P., and Schillak, S., 2009. Determination of the SLR station coordinates and velocities on the basis of laser observations of low satellites. In Schillak, S. (ed.), *Proceedings of the 16th International Workshop on Laser Ranging*, Poznán, Poland, October 13–17, 2008, Space Research Centre, Polish Academy of Sciences.

Lemoine, J. M., Bruinsma, S., Loyer, S., Biancale, R., Marty, J. C., Perosanz, F., and Balmino, G., 2007. Temporal gravity field models inferred from GRACE data. *Advances in Space Research*, **39**(10), 1620–1629.

Leuliette, E. W., Nerem, R. S., and Mitchum, G. T., 2004. Calibration of TOPEX/Poséidon and Jason altimeter data to construct a continuous record of mean sea level change. *Marine Geodesy*, **27**(1), 79–94.

Luck, J., Moore, C., and Wilson, P. (eds.), 2008. *Proceedings of the 15th International Workshop on Laser Ranging*, Canberra, Australia, October 16–20, 2006, EOS Space Systems Pty. Ltd.

Luthcke, S. B., Zelensky, N. P., Rowlands, D. D., Lemoine, F. G., and Williams, T. A., 2003. The 1-centimeter orbit: Jason-1 precision orbit determination using GPS, SLR, DORIS and altimeter data. *Marine Geodesy*, **26**, 399–421.

Marshall, J. A., Klosko, S. M., and Ries, J. C., 1995. Dynamics of SLR tracked satellites. *Reviews of Geophysics, Supplement*, U.S. national report to IUGG 1991–1994, pp. 353–360.

McCarthy, D. D., and Petit, G. (eds.), 2004. IERS conventions (2003). IERS Technical Note 32.

McGarry, J., and Zagwodski, T., 2008. SLR2000: the path toward completion. In Luck et al. (eds.), *Proceedings of the 15th International Workshop on Laser Ranging*, Canberra, Australia, October 16–20, 2006, EOS Space Systems Pty. Ltd.

Mendes, V. B., Prates, G., Pavlis, E. C., Pavlis, D. E., and Langley, R. B., 2001. Improved mapping functions for atmospheric refraction correction in SLR. *Geophysical Research Letters*, **29**(10), 1414.

Mendes, V. B., and Pavlis, E. C., 2004. High-accuracy zenith delay prediction at optical wavelengths. *Geophysical Research Letters*, **31**, L14602.

Métris, G., Vokroulicky, D., Ries, J. C., and Eanes, R. J., 1999. LAGEOS Spin axis and non-gravitational excitations of its orbit. *Advances in Space Research*, **23**(4), 721–725.

Ni, W. T., Zhu, J., Wu, X. P., Chu, G. B., Yang, B., Gao, J., Guan, M., Tang, C. J., Chou, Y., Chang, C. H., Huang, T., Qu, Q. Y., Yi, Z. H., Li, G., Tao, J., Wu, A. M., Luo, J., Yeh, H. C., Zhou, Z. B., Xiong, Y., Bi, S. L., Xu, C., Wu, X. J., Tang, M. X., Bao, Y., Li, F., Huang, C., Yang, F. M., Ye, S. H., Zhang, S. L., Zhang, Y. Z., Nie, Y., Chen, G., Christensen-Dalsgaard, J., Dittus, H., Fujii, Y., Lämmerzahl, C., Mangin, J. F., Peters, A., Rüdiger, A., Samain, E., and Schiller, S., 2002. Mini-ASTROD: mission concept. *International Journal of Modern Physics D*, **11**, 1035–1348.

Nicolas, J., Pierron, F., Samain, E., and Barlier, F., 2001. Centimeter accuracy for the French transportable laser ranging station (FTLRS) through sub-system controls. *Surveys in Geophysics*, **22**, 449–464.

Noll, C., and Pearlman, M. (eds.), 2007. *International Laser Ranging Service 2005–2006 report*. NASA/TP-2007-214153.

Otsubo, T., and Appleby, G. M., 2003. System-dependent center-of-mass correction for spherical geodetic satellites. *Journal of Geophysical Research*, **108**(B4), 2201.

Otsubo, T., Kobayashi, M., Gotoh, T., and Kubo-oka, T., 2009. Multi-satellite daily bias report: how to read and handle it. In Schillak, S. (ed.), *Proceedings of the 16th International Workshop on Laser Ranging*, Poznán, Poland, October 13–17, 2008, Space Research Centre, Polish Academy of Sciences.

Pavlis, E. C., and Kuźmicz-Cieślak, M., 2009a. Geocenter motion: causes and modeling approaches. In Schillak, S. (ed.), *Proceedings of the 16th International Workshop on Laser Ranging*, Poznán, Poland, October 13–17, 2008, Space Research Centre, Polish Academy of Sciences.

Pavlis, E. C., and Kuźmicz-Cieślak, M., 2009b. SLR and the next generation global geodetic networks. In Schillak, S. (ed.), *Proceedings of the 16th International Workshop on Laser Ranging*, Poznán, Poland, October 13–17, 2008, Space Research Centre, Polish Academy of Sciences.

Pavlis, E. C., Kuźmicz-Cieślak, M., and Ninkey, P. M., 2009. Improved modeling approaches towards the mm SLR. In Schillak, S. (ed.), *Proceedings of the 16th International Workshop on Laser Ranging*, Poznán, Poland, October 13–17, 2008, Space Research Centre, Polish Academy of Sciences.

Pearlman, M. R., Degnan, J. J., and Bosworth, J. M., 2002. The international laser ranging service. *Advances in Space Research*, **30**(2), 135–143.

Reigber, C., Schwintzer, P., Neumayer, H., Barthelmes, F., König, R., Förste, C., Balmino, G., Biancale, R., Lemoine, J. M., Loyer, S., Bruinsma, S., Perosanz, F., and Fayard, T., 2003. The CHAMP-only earth gravity field model EIGEN-2. *Advances in Space Research*, **31**(8), 1833–1888.

Ries, J. C., Huang, C., and Watkins, M. M., 1988. The effect of general relativity on near-earth satellites in the solar system Barycentric and Geocentric reference frames. *Physical Review Letters*, **61**, 903–906.

Samain, E., Guillemot, P., Exertier, P., Albanese, D., Berio, P., Laurain, O., Para, F., Paris, J., Torre, J. M., Viot, H., Vrancken, P., Petitbon, I., and Leon, S., 2009. Time transfer by laser link – T2L2: first data. In Schillak, S. (ed.), *Proceedings of the 16th International Workshop on Laser Ranging*, Poznán, Poland, October 13–17, 2008, Space Research Centre, Polish Academy of Sciences.

Schillak, S. (ed.), 2009. *Proceedings of the 16th International Workshop on Laser Ranging*, Poznán, Poland, October 13–17, 2008, Space Research Centre, Polish Academy of Sciences.
Schutz, B. E., Cheng, M. K., Eanes, R. J., Shum, C. K., and Tapley, B. D., 1993. Geodynamic results from starlette orbit analysis. In Smith, D. E., and Turcotte, D. L. (eds.), *Contributions of Space Geodesy to Geodynamics: Earth Dynamics*. Geodynamics Series, Vol. 24, pp. 175–190.
Tapley, B. D., Schutz, B. E., Eanes, R. J., Ries, J. C., and Watkins, M. M., 1993. Lageos laser ranging contributions to geodynamics, geodesy, and orbital dynamics. In Smith, D. E., and Turcotte, D. L. (eds.), *Contributions of Space Geodesy to Geodynamics: Earth Dynamics*. Geodynamics Series, Vol. 24, pp. 147–174.
Tapley, B. D., Bettadpur, S., Watkins, M., and Reigber, C., 2004. The gravity recovery and climate experiment: Mission overview and early results. *Geophysical Research Letters*, **31**(9), L09607.
Tapley, B., Ries, J., Bettadpur, S., Chambers, D., Cheng, M., Condi, F., Gunter, B., Kang, Z., Nagel, P., Pastor, R., Pekker, T., Poole, S., and Wang, F., 2005. GGM02 – An improved earth gravity field model from GRACE. *Journal of Geodesy*, **79**(8), 467–478.
Tarantola, A., 2004. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Philadelphia: SIAM.
Urschl, C., Beutler, G., Gurtner, W., Hugentobler, U., and Schaer, S., 2007. Contribution of SLR tracking data to GNSS orbit determination. *Advances in Space Research*, **39**, 1515–1523.
van Dam, T., Plag, H. P., Francis, O., and Gegout, P., 2002. GGFC special bureau for loading: current status and plans. *IERS Technical Note*, **20**, 180–198.

Cross-references

Earth Rotation
Geodesy, Figure of the Earth
Geodesy, Networks and Reference Systems
Geodesy, Physical
GPS, Data Acquisition and Analysis
Gravity Field of the Earth
Gravity Field, Temporal Variations from Space Techniques
Gravity Field, Time Variations from Surface Measurements
Gravity, Global Models
Spherical Harmonic Analysis Applied to Potential Fields
Very Long Baseline Interferometry

SEAFLOOR SPREADING

Richard N. Hey
Hawaii Institute of Geophysics and Planetology, School of Ocean and Earth Science and Technology, University of Hawaii, Honolulu, HI, USA

Definition

Seafloor spreading is the mechanism by which new oceanic lithosphere is created at and moves away from divergent plate boundaries. The seafloor spreading hypothesis led to one of the most important paradigm shifts in the history of the Earth sciences.

Introduction

The revolutionary seafloor spreading hypothesis improved and subsumed the continental drift hypothesis, and rapidly culminated in what is now known as the plate tectonics scientific revolution. It was first hypothesized by Hess in 1960, in a widely circulated preprint and paper (Hess, 1962) he considered so speculative that he called it "an essay in geopoetry". This hypothesis was named "seafloor spreading" in another influential early paper (Dietz, 1961). It offered a simple explanation for many problems with the prevailing paradigm that the Earth was a mostly static, slowly contracting planet, with fixed continents and old ocean basins, and no large-scale horizontal displacements. This paradigm had previously been challenged, most notably by Wegener's continental drift hypothesis (Wegener, 1912), and by paleomagnetic measurements in the 1950s that were consistent with continental drift, but before the 1960s these ideas were not generally accepted.

The revolution

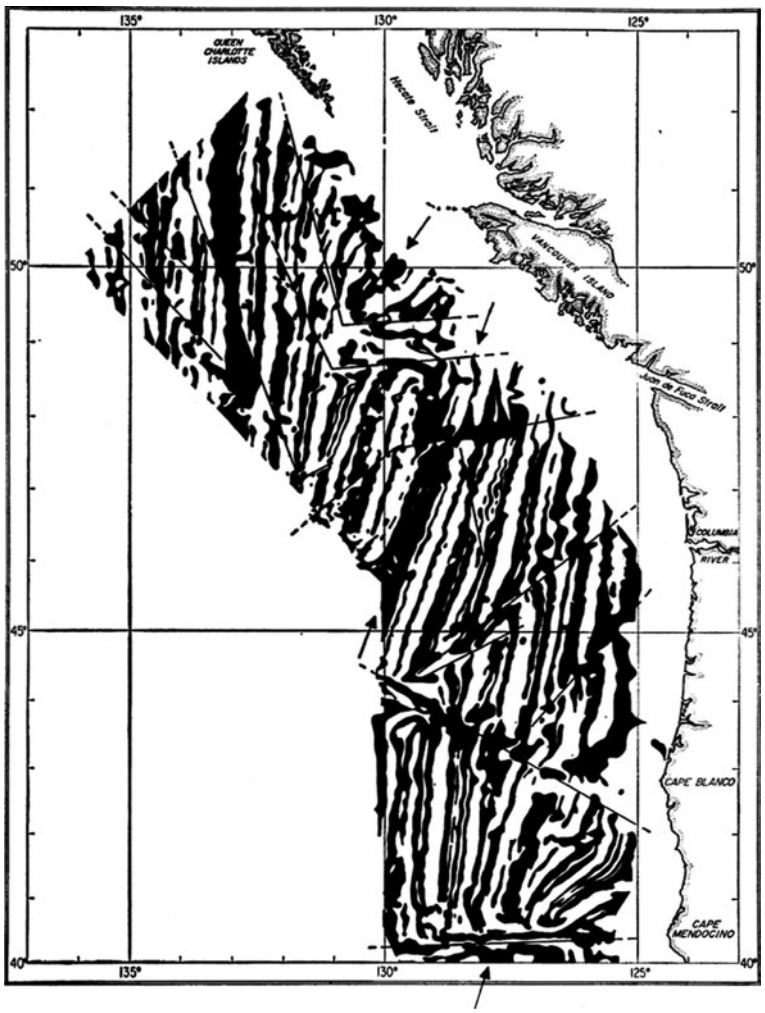
Hess realized that if mantle convection carried seafloor and continents away from seafloor spreading centers (mid-ocean ridges) toward trenches (subduction zones), with new seafloor formed at ridge axes and destroyed at trenches, this would explain the shallow bathymetry, earthquakes, high heat flow, lack of sediments, and extensional structures characterizing ridge axes, as well as the deep trenches, earthquakes, compressional structures, mountain ranges, and volcanoes characterizing subduction zones.

A key step in the confirmation of seafloor spreading was the recognition by Vine and Matthews (1963) (and independently by Morley, in a paper unfortunately turned down as too speculative by both *Nature* and the *Journal of Geophysical Research*, eventually published in Morley and Larochelle, 1964), which offered a simple explanation for the existence of puzzling magnetic "stripes" on the seafloor. The Vine–Matthews (or Vine–Matthews–Morley) hypothesis proposed that a combination of seafloor spreading and episodic reversals of the Earth's magnetic field (at the time another very speculative idea) would create alternating zones of normally and reversely magnetized crust, and thus linear positive and negative magnetic anomalies in a pattern symmetric about the spreading axis.

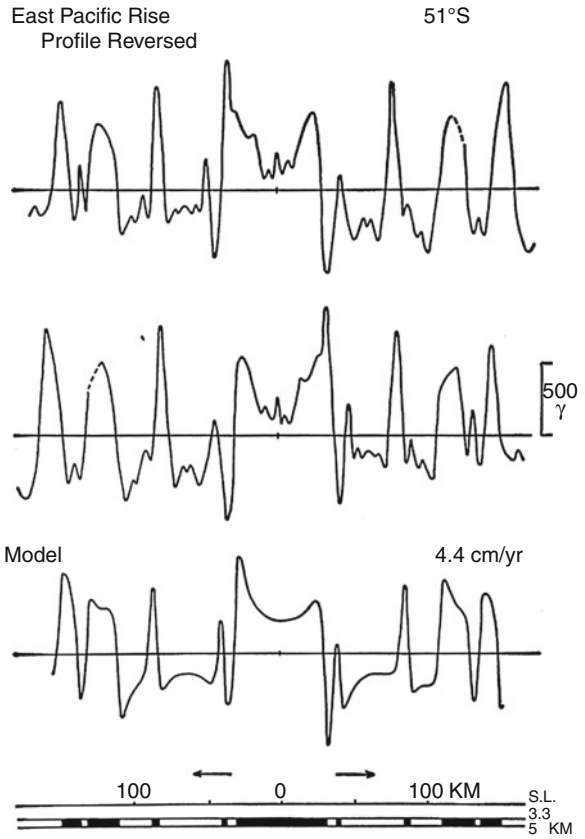
The next key step in the revolution occurred in 1965, when Wilson noted that deformation of the Earth's crust is concentrated in narrow mobile belts, and postulated that these features are all interconnected in a global network, the first qualitative model of plate tectonics (Wilson, 1965). The zones of extension and compression are connected by a new class of faults defined by Wilson as transform faults, which required relative plate motion and turned out to be the most important type of fault on Earth. He showed that seafloor spreading occurring on offset mid-ocean ridge axes would produce relative motion exactly opposite to the motion the sense of offset would predict without seafloor spreading, and that earthquakes should only occur between the offset seafloor spreading axes. These radical predictions, completely opposite to prevailing wisdom, were soon confirmed seismically. Furthermore, by correctly interpreting the San Andreas fault

as a transform fault between the Pacific and North American plates, he predicted the existence of previously unrecognized seafloor spreading offshore western North America. Vine and Wilson (1965) showed that the predicted magnetic symmetry existed in this area, where magnetic stripes had been discovered in the mid-1950s (Mason, 1958; Mason and Raff, 1961; Raff and Mason, 1961), over what they recognized as the Juan de Fuca Ridge (Figure 1), and furthermore, that the pattern of stripes corresponded perfectly with the pattern of magnetic field reversals (Cox et al., 1963; McDougall and Tarling, 1963), once the Jaramillo anomaly was discovered. Vine (1966) also demonstrated similar symmetry and correlation with the reversal timescale in another important data set, the Project Magnet aeromagnetic data collected over the Reykjanes Ridge south of Iceland. His color figures of these classic data sets and correlations (Vine, 1968) became iconic.

The smoking gun for many scientists was the discovery by Pitman and Heirtzler (1966) of near perfect symmetry in the Eltanin-19 profile collected over the Pacific–Antarctic Ridge. Essentially every tiny wiggle seen in the magnetic anomaly profile on the Pacific plate was mirrored on the Antarctic plate, and correlated perfectly with the magnetic reversal time scale (Figure 2). The symmetry in these data required a symmetric axial process, with new seafloor carried away on both plates, and thus provided compelling evidence for both the Vine–Matthews and seafloor spreading hypotheses. Vine convincingly summarized this evidence in influential symposia and publications (Vine, 1966, 1968), and, by the end of 1966, seafloor spreading was generally accepted by marine geophysicists, who quickly extrapolated the magnetic reversal time scale from ~10 Ma to ~80 Ma and worked out at least the basic recent evolutionary history of every ocean basin.



Seafloor Spreading, Figure 1 Raff and Mason (1961) magnetic stripes recognized as symmetric about seafloor spreading axes (*arrows*) by Vine and Wilson (1965). (From Vine, 1966. Reprinted with permission from AAAS.)



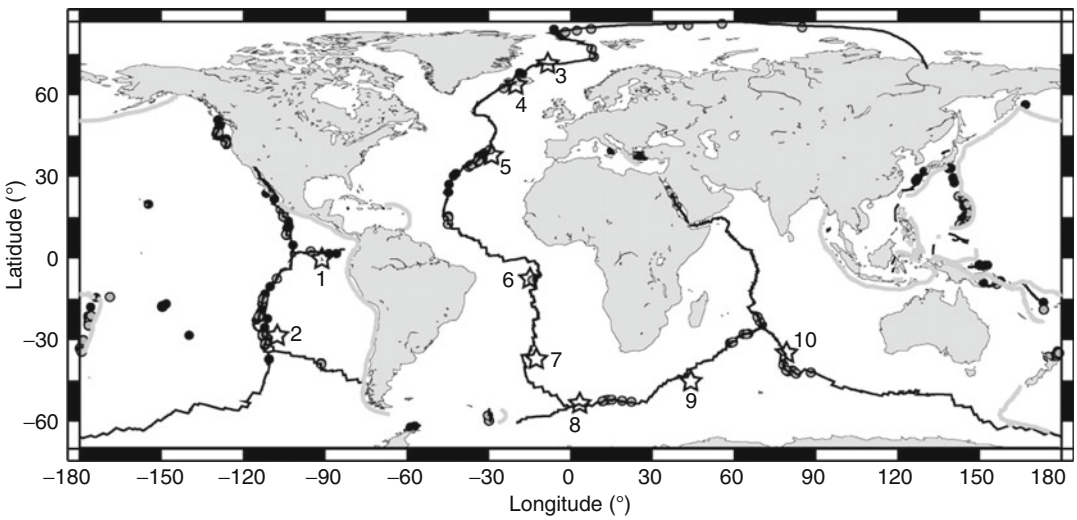
Seafloor Spreading, Figure 2 Eltanin-19 magnetic anomaly profile (*top*) from Pitman and Heirtzler (1966), shown compared with its mirror image (*middle*) and magnetic anomaly model calculated from the magnetic reversal timescale (*bottom*), assuming seafloor spreading at a constant 44 km/Myr. (From Vine, 1966. Reprinted with permission from AAAS.)

This scientific revolution culminated in plate tectonics the following year when Morgan (1968) and McKenzie and Parker (1967) made the key assumption that plates behave rigidly, and extended Wilson's transform fault geometry to the sphere. They showed that transforms are small circles about rotation poles describing relative plate motion, that seafloor spreading rates increase as the sine of the angular distance away from these poles, and that it was possible to use known patterns of seafloor spreading to quantitatively predict other plate motions and plate boundary evolution.

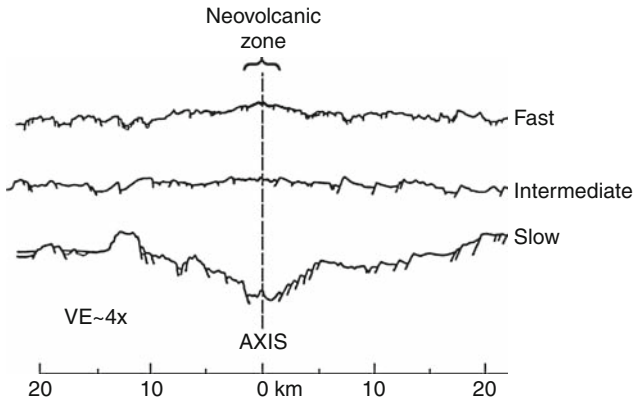
Ridge axis geometry, morphology, and crustal formation

Volcanism along the mid-ocean ridge system has formed the longest mountain chain on Earth (Figure 3). As the plates slowly (~0–150 mm/yr, or km/Myr, at present, DeMets et al., 2010, with rates remarkably similar whether using magnetic anomalies or geodetic measurements such as GPS) move apart by seafloor spreading, magma from the hot asthenosphere rises to fill the crack at mid-ocean ridge axes, elevated because of the higher temperatures. Both axial depth and morphology correlate with spreading rate (Macdonald, 1982). Fast-spreading ridge axes such as the East Pacific Rise generally have shallow depths and relatively smooth morphologies, with very small extensional axial summit troughs (except where these are completely buried by the latest eruption in the neovolcanic rift zone). Slow spreading ridges such as the Mid-Atlantic Ridge generally have much deeper, rougher, and higher-amplitude axial valleys (Figure 4), except near hotspots such as Iceland where magma supply is unusually large and even slow-spreading ridges have shallow axes with fast-spreading morphology.

The asthenosphere typically melts to become mid-ocean ridge basalt in a magma chamber under the ridge



Seafloor Spreading, Figure 3 Global distribution of seafloor spreading axes (*heavy black lines*) with known (*black dots*) or inferred (*gray dots*) hydrothermal fields. Stars are near-ridge hotspots. From NOAA/PMEL Vents Program website: http://www.pmel.noaa.gov/pubs/outstand/bake2544/images/fig01a.gif (Baker and German, 2004).



Seafloor Spreading, Figure 4 Ridge axis fault patterns and morphology as a function of spreading rate. (Modified from Macdonald, 1982.)

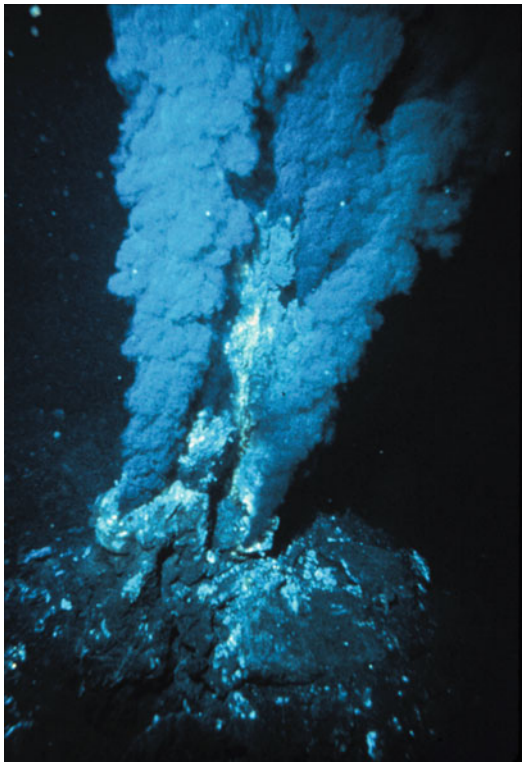
axis. Some basalt is erupted onto the seafloor to form pillow basalts and lava flows, which are progressively covered by sediments as the lithosphere ages and moves away from the axis. Below the basalts are the sheeted dikes, the paths through which the lava moved from magma chambers to the surface, and below the dikes are the layered gabbros which cooled and crystallized in place (intrusives) rather than erupting at axial volcanoes (extrusives). This typical sequence of mantle ultramafics – gabbro – sheeted dikes – extrusive basalts – sediments is observed on the seafloor in tectonic windows where existing lithosphere is rifted apart, and also where slices of old ocean lithosphere are thrust into ophiolite mountain belts as in Oman and Cyprus (Karson, 2002).

Hydrothermal vents

As the new seafloor cools it contracts and cracks. Water goes down these cracks, is heated, reacts with the surrounding rocks, and comes back up as buoyant hydrothermal vents (Baker and German, 2004). These occur at a range of chemistries and temperatures, up to ~400°C in black smokers, so-called because of the sulfide-rich plumes that precipitate suddenly when the superheated water is injected into the surrounding ~2°C seawater, forming sulfide chimneys (Figure 5). White and clear smokers occur at lower temperatures. The chemical reactions provide energy for an unusual kind of life that does not depend on photosynthesis but flourishes as chemosynthetic communities at many of these vents, rare deep-ocean oases in an otherwise extremely barren environment (Kelley et al., 2002).

Summary

Seafloor spreading was a critical step in the contentious scientific revolution from the previous static Earth paradigm to the now universally accepted plate tectonic paradigm. Today it refers to the processes creating new oceanic lithosphere where plates move apart. Seafloor spreading replaces the lithosphere destroyed by subduction, and exerts important influences on Earth's chemical and biological evolution.



Seafloor Spreading, Figure 5 Black smoker hydrothermal vents and sulfide chimneys. (From NOAA website: http://www.photolib.noaa.gov/htmls/nur04506.htm.)

Bibliography

Baker, E. T., and German, C. R., 2004. On the global distribution of hydrothermal vent fields. In German, C. R., et al. (eds.), *Mid-Ocean Ridges: Hydrothermal Interactions Between the Lithosphere and Oceans*. Washington, DC: American Geophysical Union. Geophysical Monograph, Vol. 148, pp. 245–266.

Cox, A., Doell, R. R., and Dalrymple, G. B., 1963. Geomagnetic polarity epochs and Pleistocene geochronometry. *Nature*, **198**, 1049–1051.

DeMets, C., Gordon, R. G., and Argus, D. F., 2010. Geologically current plate motions. *Geophysical Journal International*, **181**(1), 1–80, doi:10.1111/j.1365-246X.2009.04491.x.

Dietz, R. S., 1961. Continent and ocean basin evolution by spreading of the sea floor. *Nature*, **190**, 854–857.

Hess, H. H., 1962. History of ocean basins. In Engel, A. E. J., James, H. L., and Leonard, B. F. (eds.), *Petrologic Studies: A Volume to Honor A.F. Buddington*. New York: Geological Society of America, pp. 599–620.

Karson, J. A., 2002. Geologic structure of the uppermost oceanic crust created at fast- to intermediate-rate spreading centers. *Annual Review of Earth and Planetary Sciences*, **30**, 347–384.

Kelley, D. S., Baross, J. A., and Delaney, J. R., 2002. Volcanoes, fluids and life at mid-ocean ridge spreading centers. *Annual Review of Earth and Planetary Sciences*, **30**, 385–491.

Macdonald, K. C., 1982. Mid-ocean ridges: fine scale tectonic, volcanic and hydrothermal processes within the plate boundary zone. *Annual Review of Earth and Planetary Sciences*, **10**, 155–190.
Mason, R. G., 1958. A magnetic survey over the west coast of the United States between latitudes 32° and 36° N, longitudes 121° and 128° W. *Geophysical Journal of the Royal Astronomical Society*, **1**, 320–329.
Mason, R. G., and Raff, A. D., 1961. A magnetic survey off the west coast of North America 32° N to 42° N. *Bulletin of the Geological Society of America*, **72**, 1259–1265.
McDougall, I., and Tarling, D. H., 1963. Dating of polarity zones in the Hawaiian Islands. *Nature*, **200**, 54–56.
McKenzie, D. P., and Parker, R. L., 1967. The North Pacific: an example of tectonics on a sphere. *Nature*, **216**, 1276–1280.
Morgan, W. J., 1968. Rises, trenches, great faults and crustal blocks. *Journal of Geophysical Research*, **73**, 1959–1982.
Morley, L. W., and Larochelle, A., 1964. Paleomagnetism as a means of dating geological events. In *Geochronology in Canada*. Royal Society of Canada Special Publication, Vol. 8, pp. 39–50.
Pitman, W. C., III, and Heirtzler, J. R., 1966. Magnetic anomalies over the Pacific-Antarctic Ridge. *Science*, **154**, 1164–1171.
Raff, A. D., and Mason, R. G., 1961. Magnetic survey off the west coast of the United States between 40°N latitude and 52°N latitude. *Bulletin of the Geological Society of America*, **72**, 1267–1270.
Vine, F. J., 1966. Spreading of the ocean floor: new evidence. *Science*, **154**, 1405–1415.
Vine, F. J., 1968. Magnetic anomalies associated with mid-ocean ridges. In Phinney, R. A. (ed.), *The History of the Earth's Crust*. Princeton: Princeton University Press, pp. 73–89.
Vine, F. J., and Matthews, D. H., 1963. Magnetic anomalies over oceanic ridges. *Nature*, **199**, 947–949.
Vine, F. J., and Wilson, J. T., 1965. Magnetic anomalies over a young oceanic ridge off Vancouver Island. *Science*, **150**, 485–489.
Wegener, A., 1912. Die entstehung der kontinente. *Geologishe Rundschau*, **3**, 276–292.
Wilson, J. T., 1965. A new class of faults and their bearing on continental drift. *Nature*, **207**, 343–347.

Cross-references

Continental Drift
Continental Rifts
Lithosphere, Mechanical Properties
Lithosphere, Oceanic
Plate Driving Forces
Plates and Paleoreconstructions
Subduction Zones

SEDIMENTARY BASINS

Magdalena Scheck-Wenderoth
Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences
Sect. 4.4: Basin Analysis, Potsdam, Germany

Definition

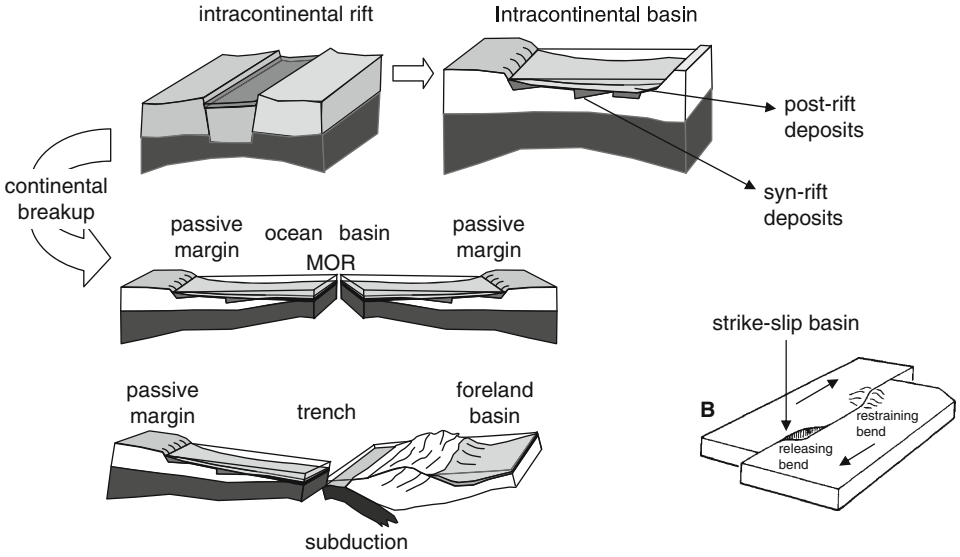
Sedimentary basins are regions of prolonged subsidence of the Earth's surface that provide the accommodation space for mineral and organic material (Allen and Allen, 2005). These deposits – the sedimentary rocks – are the record of the past geological history including tectonic events, climatic conditions, changes in sea level, and other environmental modifications. In addition, sedimentary basins are long-lived, low-temperature geo-reactors in which the accumulated material experiences a variety of transformations (Bjorlykke, 2010; Littke et al., 2008; Roure et al., 2009; Welte et al., 1997). As a result of these processes, basins contain our resources of fossil fuels, ground water, and inorganic commodities. Moreover, they are important reservoirs of heat and repositories for different socioeconomically relevant fluids such as CO_2.

Basin types

Basins can be classified in terms of their plate-tectonic setting. The plate-tectonic Wilson cycle (Figure 1) describes that the relative movements of plates on the sphere of the earth result in a closed loop of *continental rifting* and breakup, ocean basin development, closure of oceans at subduction zones and, in consequence of plate convergence, continental growth.

Basins thus can form as continental rifts that may evolve either into intracontinental basins or lead to continental breakup and ocean basin formation. New *oceanic lithosphere* (see entry *Lithosphere, Oceanic*) is produced along the Mid Ocean Ridges due to oceanic spreading. If oceanic lithosphere cools, it becomes denser and subsides in the oceanic basins that are bordered by passive continental margins. Of the latter "hot" or volcanic passive margins are distinguished from "cold" or nonvolcanic passive margins (White et al., 2003). If oceanic lithosphere cools beyond a specific threshold, it becomes too heavy to be sustained by the less dense asthenosphere and will finally descend back to the mantle at subduction zones. In these convergent settings, the descending plate is flexed downward and deep oceanic trenches develop above the down-going plate. Subduction may culminate in continent–continent collision if the oceanic lithosphere is entirely subducted and collisional fold and thrust belts form. Loading by collision-related fold and thrust belts also causes a downward flexure of the lithosphere and foreland basins to form. The Wilson Cycle may stop at any evolutionary step, because the causative forces cease to be effective. Accordingly, a continental rift not necessarily develops into an ocean basin, but may survive for hundreds of millions of years as an intracontinental basin (Heine et al., 2008; Littke et al., 2008). Finally, horizontal movements along strike-slip faults may also cause local extension and related pull-apart basins (Allen and Allen, 2005; Petrunin and Sobolev, 2008; Smit et al., 2010; Weber and group, 2009).

These different types of basins have a characteristic structure of the sediment fill as well as of the underlying crust and mantle lithosphere. This concerns the geometric configuration, the distribution of physical properties, and the resulting *isostatic* (see entry *Isostasy*) and thermal configuration. To assess the configuration of a basin, a wide spectrum of methods has to be integrated. Observations



Sedimentary Basins, Figure 1 Basins in their plate-tectonic setting.

obtained from field measurements, from *deep seismic imaging* (see entry *Deep Seismic Reflection and Refraction Profiling*), and from wells drilled into the basin fill, potential field data, and *heat flow* (see entry *Heat Flow Measurements, Continental*) data as well as new data sets from remote sensing need to be integrated with numerical models that simulate processes in basins at different scales.

Figure 2 shows exemplary crustal sections across the Norwegian passive margin and across the intracontinental Central European Basin System.

Basin-forming mechanisms

There are three main mechanisms leading to the formation of sedimentary basins (Allen and Allen, 2005)

1. Extensional stresses causing the lithosphere to thin
2. Flexure of the lithosphere in response to surface loading or compressive stresses
3. Viscous flow of the mantle causing positive or negative topography

All these mechanisms may act separately or in concert, and may affect the lithosphere over a range of spatial and temporal scales.

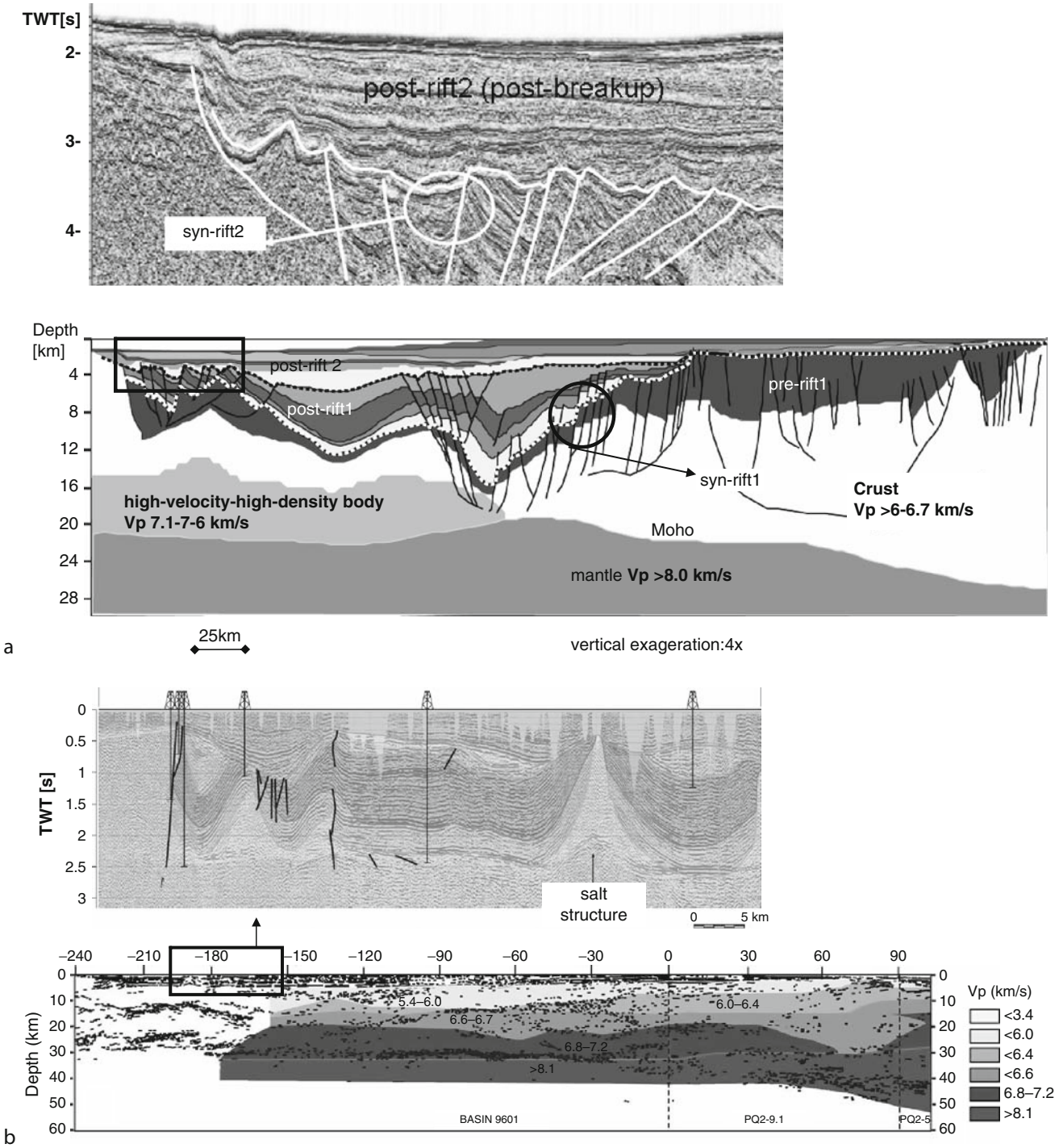
Conceptual models of basin formation

Conceptual models of basin formation (Figure 3) attempt to integrate geophysical and geological observations describing the geometry and the physical state of a basin into evolutionary schemes that explain basin evolution. These conceptual models also need to be consistent with data on the rheological behavior of rocks known from lab experiments as well as with data on the thermal and deformation history. All these conceptual models represent end-member scenarios and explain a large fraction of observed phenomena in specific basins.

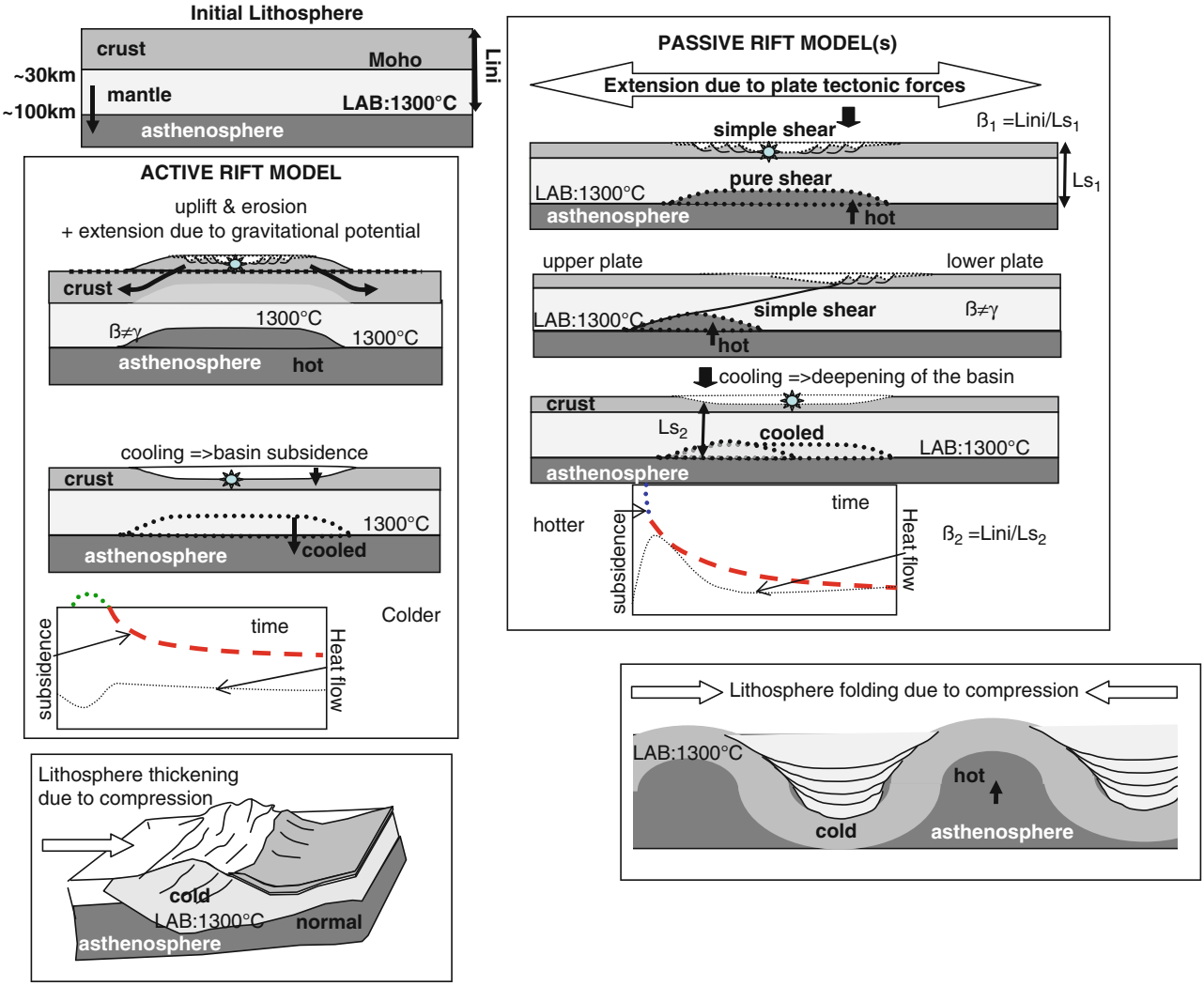
In the concept of *passive rifting* (Figure 3) the lithosphere is stretched in response to extensional tectonic stress. This may result in pure shear thinning (McKenzie, 1978), in simple shear thinning (Wernicke, 1981), or in a mixture of the two (Lister et al., 1986).

The *uniform stretching model* (McKenzie, 1978) predicts a crustal and lithosphere thickness proportional to the stretching factor ß where ß is the ratio between initial lithosphere thickness and the stretched lithosphere thickness. Lithosphere thinning results in a passive rise of the isotherm at the thermal lithosphere–asthenosphere boundary (Turcotte and Schubert, 2002). This, in turn, causes additional thermal subsidence after the stress ceases to be effective and cooling of the previously heated lithosphere causes a density increase. In this model, uniform and symmetric lithospheric thinning by ß takes place and is accommodated by brittle faulting in the upper crust as well as by ductile flow in the lower crust and upper mantle. The uniform stretching model predicts a 2-stage subsidence: (1) as a result of tectonic stretching – on a short timescale, (10–20 my), and (2) as a result of thermal subsidence – on a longer time scale (50–100 my). Structurally, the uniform stretching model predicts syn-rift extensional faults in the upper crust with syn-rift sediments overlain by onlapping post-rift sediments of the thermal subsidence phase. The thermal history predicted by the uniform stretching model includes an initial heat peak due to the rise of the isotherm at the LAB to a shallower position (and thus an increase of the thermal gradient within the lithosphere) followed by continuous cooling during thermal re-equilibration.

Depth-dependent stretching (Sclater and Christie, 1980) is a modification of the uniform stretching model



Sedimentary Basins, Figure 2 (**a**) *Lower panel*: Line drawing of a crustal scale integrated cross section with three generations of basins from the Norwegian passive margin modified after Faleide et al., 2008. *Upper panel*: Seismic close-up showing the breakup unconformity and pre-breakup extensional faulting during phase syn-rift 2. (**b**) *Lower panel*: Line drawing with superimposed seismic p-wave velocities from the intracontinental North German Basin modified after Bayer et al., 2002. *Upper panel*: Close-up with reflection seismic example illustrating polyphase subsidence with several sequences separated by unconformities and additional salt-tectonic deformation. (Modified after Mazur et al., 2005.)



Sedimentary Basins, Figure 3 Conceptual models of basin formation.

and accounts for different stretching magnitudes in the upper brittle part of the crust by a factor ß and the deeper lithosphere by a factor γ. In the *asymmetric stretching models* (Lister et al., 1986; Wernicke, 1981) crust- to lithosphere-scale listric faults accommodate large offsets in response to extensional stress and finally may lead to continental breakup with an "upper plate" margin developing from the hanging wall of the master detachment fault and a "lower plate" margin in the foot wall (Osmundsen and Ebbing, 2008; Sibuet et al., 2007; Zuber et al., 1986). In particular, nonvolcanic margins like the Galicia-Newfoundland (Sibuet et al., 2007) are commonly explained by this mechanism. Geodynamic modelling studies (Braun and van der Beek, 2004; Buck, 1991; Cloetingh et al., 1998; Huismans et al., 2005; Huismans and Beaumont, 2008; Kusznir and Ziegler, 1992; Lavier et al., 1999; Pascal and Cloetingh, 2002; Pérez-Gussinyé et al., 2006; Royden and Keen, 1980; Sclater and Christie, 1980; Simon et al., 2009; van Wijk, 2005; Wijns et al., 2005; Zuber et al., 1986) indicate that the nature of rheological coupling between the crust and the lithospheric mantle, local strain weakening and strengthening, and the rate of extension are the dominant factors controlling not only the subsidence evolution of basins but also if they develop in a symmetric or asymmetric style and in a narrow or wide rift mode.

In the concept of *active rifting*, basin initiation is caused by the impingement of a rising hot asthenospheric *mantle plume* (Allen and Allen, 2005; Burov and Cloetingh, 2009; Haxby et al., 1976; Turcotte and Schubert, 2002). The concept of active rifting is in fact a special case of basins initiated by *dynamic topography*. The emplacement of the *mantle plume* (see entry *Mantle Plumes*) causes uplift and induces extensional stresses at the flanks of the uplifted area due to gravitational potential. The active rise of the isotherm may lead to surface uplift above the

erosional level and the eroded material may be removed from the system. Density increase in the lithosphere due to subsequent cooling causes surface subsidence and the formation of a sedimentary basin in the previously uplifted area. As the process is transient, the system would re-equilibrate after the thermal anomaly and the related dynamic topography stop to exist. For *isostatic* reasons, net subsidence will only occur if either igneous underplating of the crust has taken place (light crust replaced by dense underplate) or if material is removed from the system by erosion during uplift. Systematic work assessing the contribution of the three processes to subsidence (extension due to gravity potential, crustal underplating, and erosion-related mass loss) is still missing. The active rift model predicts weak to missing syn-rift faulting, a syn-rift erosional unconformity, and a subsidence history corresponding predominantly to thermal subsidence. Actually, the active rift model predicts a subsidence history identical to a specific end-member case of (passive) depth-dependent stretching in which the mantle lithosphere is severely thinned whereas the crust is not. However, the active rift model predicts a colder heat flow scenario than the passive rift model, as the thermal gradient is not disturbed by an initial thinning of the lithosphere.

According to the concept of dynamic topography basins can also be initiated in response to downward flow of high density material in the convective mantle, a process causing *negative dynamic topography*. This mechanism could be related to detached subduction slabs sinking into the asthenospheric mantle and is discussed as an explanation for the subsidence of intracontinental basins that subside over several hundreds of millions of years at a very slow rate (Allen and Allen, 2005; Heine et al., 2008; Ritzmann and Faleide, 2009).

Two main conceptual models exist for *flexural basins* that form as response of the flexural rigidity of the lithosphere (Watts, 2001) to external loads: (1) downward flexure of the lithosphere due to vertical loading (Zhang and Bott, 2000) and (2) formation of crustal- to lithosphere-scale folds in response to horizontal compressive stress, a process also called *lithosphere buckling* (Burov and Cloetingh, 2009; Burov and Diament, 1992; Cloetingh and Burov, 2010). Both concepts predict a lack of syn-rift extensional faulting, a cold thermal scenario, and subsidence lasting only as long as the stress is effective (either vertical as load or horizontal as tectonic stress). Due to structural similarity, lithosphere buckling in response to far-field compressive tectonic stress (Burov and Cloetingh, 2009; Burov and Diament, 1992) can be mistaken as an expression of active rifting if the analyzed region is too small to capture the full wavelength of the buckling. Again, the different predicted subsidence and thermal histories can help to distinguish the two.

Vertical loading by sediments enhances subsidence in basins of all geodynamic settings (Watts, 2001). It has recently been proposed that mineral phase transitions may also play a role for subsidence history in different tectonic settings (Kaus et al., 2005).

The configuration of the sediment fill

The sediment fill can vary in thickness between a few to up to 20 km and is analyzed using stratigraphic methods (Allen and Allen, 2005; Catuneanu, 2006; Einsele et al., 1991; Vail et al., 1991) to assess the nature of sediments (sand, silt, shale, carbonates, evaporites), the depositional environment (continental or marine), and the climatic conditions prevailing during the time of deposition. Subsidence and deposition may alternate with nondeposition due to sea level fall or tectonic surface uplift, and lead to stratigraphic gaps separating different sequences (Figure 2). In addition, the rate of subsidence and sedimentation as well as of uplift and erosion may change with time. Stratigraphic gaps and changes in depositional conditions are expressed as unconformities separating sedimentary sequences the analysis of which is required to understand the history of a basin.

Progressive sedimentation and subsidence goes along with changes in physical properties of the deposited sediments. The two most prominent changes are (1) an increase in temperature and (2) a decrease in porosity with increasing burial depth. The initially high porosity of each layer decreases as the load of the covering deposits increases and fluids filling the initial pores are expelled. Also, chemical compaction may lead to a loss in porosity. With decreasing porosity, physical properties like density, seismic velocity, or thermal conductivity increase.

Stratigraphic analysis is performed in field studies if the fill of a basin is accessible in outcrops due to postdepositional uplift. Most of the sedimentary basins, however, extend by their very nature (areas of enduring subsidence) into the subsurface and to large depths. Accordingly, subsurface analytical methods are required to assess their configuration. The most important of these methods is *seismic imaging*. *Seismic reflections* (see entry *Deep Seismic Reflection and Refraction Profiling*) originate from the contrast in impedance at discontinuities in the subsurface that can represent stratigraphic boundaries, and unconformities or changes in attributes of a specific layer. The reflection seismic technique provides a picture of these interfaces in a two-way-travel time that is constrained by wells. Wells provide information on the absolute depth position of stratigraphic horizons and enable the conversion of seismic travel times to depth. Furthermore, wells yield data on the thermal gradient with depth as well as on changes in lithology and petrophysical properties. Properties such as composition, compaction, porosity, permeability, density, thermal and hydraulic conductivity, *radiogenic heat production* (see entry *Radiogenic Heat Production of Rocks*), and content and maturation of organic matter are measured using *Geophysical Well Logging* methods.

Apart from recording changes in relative sea level or climatic conditions, the configuration of the sediment fill may be influenced by deformation (Kley et al., 2008). Accordingly, sediments can be faulted, folded in response to tectonic forces, or deformed due to halokinetic movements (Hudec and Jackson, 2007; Scheck-Wenderoth

et al., 2008). This deformation can take place pre-, syn-, or postdepositionally, each of which is expressed by characteristic structural styles.

In particular, basin initiation in extensional plate-tectonic conditions is typically associated with normal faulting in the upper crust leading to a thinning of the lithosphere and to the accumulation of syn-rift deposits. After the tectonic stresses have declined, basins experience a phase of post-rift thermal subsidence and may even experience late basin inversion in response to compressive stress (Cooper and Williams, 1989; Mazur et al., 2005; Sandiford, 1999) (Figure 2).

Finally, there is a feedback between surface processes and deep mechanisms in the sense that *isostatic* loading by sediments will increase total subsidence and erosion of basin flanks will cause flexural rebound (Braun, 2005; Braun et al., 2008; Braun and van der Beek, 2004; Garcia-Castellanos, 2002; Kounov et al., 2007; Simoes et al., 2010; Van der Beek, 2007; Willenbring and von Blanckenburg, 2010).

The configuration of the crust beneath sedimentary basins

The different mechanisms responsible for the thinning of the lithosphere leave characteristic traces that can be detected with geophysical observations, which, however, are not always leading to unique interpretations. The most prominent attribute of the crust beneath sedimentary basins is that it is thinned compared to neighboring regions. Deep seismic data imaging of the preserved thickness of the crust helps to determine the amount of the stretching factor ß as the ratio between the initial crustal thickness as often preserved outside the basin margins and the crustal thickness beneath the basin. Crustal thinning can vary considerably across a basin with values of ß largest in the area of strongest thinning. The average initial thickness of the crust may reach up to 50 km in cratonic areas, around 30 km in Phanerozoic continental domains and less than 10 km below oceans.

As the crystalline crust is characterized by a higher density and higher velocities of seismic waves than the sediments, the interface between the two can be detected. Accordingly, the *interpretation of gravity anomalies* (see entry *Gravity Anomalies, Interpretation*) and the analysis of the observed variation of seismic velocities are key to evaluate the crustal structure. Refraction seismic techniques are used to detect variations in seismic p-wave velocities at depth whereas reflection seismic methods are useful to evaluate the changes in reflectivity of the deeper crust.

Seismic p-wave velocities (Vp) increase to values >6 km/s in the upper crystalline crust and can range up to 7.4 km/s (Christensen and Mooney, 1995) depending on the composition of the crust. There is a negative correlation between the Quartz content of crystalline rocks and seismic p-wave velocity as well as density. Silicic (Quartz-rich) rocks are characterized by smaller velocities and lower densities than mafic rocks (Table 1). Seismic shear wave velocities are also higher for crystalline than for sedimentary rocks, but are also dependent of additional parameters (Christensen and Mooney, 1995). Accordingly, the Vp/Vs-ratio is an additional parameter helping to interpret crustal structure (Afonso et al., 2010; Mjelde et al., 2003). As also the magnetic properties are different for mafic (Fe-rich) and silicic (Fe-poor) rocks, *the interpretation of magnetic anomalies* is an additional technique for the evaluation of the crustal structure beneath basins.

At the crust–mantle boundary (Moho), often a characteristic reflection and a sudden increase of seismic p-wave

Sedimentary Basins, Table 1 Overview on physical properties of rocks relevant to sedimentary basins: *seismic properties* (After Christensen and Mooney, 1995), *thermal properties* (After Artemieva, 2006; Cermak and Rybach, 1982; Fernandez et al., 2005; Förster and Förster, 2000; Ritter et al., 2004; Scheck-Wenderoth and Maystrenko, 2008)

	Dominant lithology	Thermal conductivity λ [W/mK]	Heat production S [mW/m^3]	Density [kg/m^3]	p-wave velocity [km/s]
	Sea water	0.563	0	1,030	1.48
S	Uncompacted fine-grained siliciclastics	1.2 (2.0)	1	1,950	2.05
E	Slightly compacted fine-grained siliciclastics	1.8 (2.2)	1	2,200	2.2
D	Medium compacted fine-grained siliciclastics	2.0 (2.2)	1	2,250	2.2
I	Compacted fine-grained siliciclastics	2.1 (2.3)	1	2,550	3.5
M	Strongly compacted fine-grained siliciclastics	2.1 (2.3)	1	2,630	4.3
E	Compacted siliciclastics with carbonates	2.1 (2.5)	0.9	2,640	5.5–6.0
N	Carbonates:chalk	1.7	0.6	2,000	3.4
T	Carbonates: limestones, dolomites	2.4	0.6	2,600	6.0
S	Rock salt	3.5 (5.4)	0.1	2,150	4.0–5.5
Crust	Granites and gneisses	2.7 (3.2)	0.8	2,790	6.0–6.7
Continent	Mafic granulites/gabbros/eclogites	2.6 (3.0)	0.3	3,120	6.8–7.0
	Gabbros	2.6 (3.0)	0.5	3,150	7.1–7.6
	Basalt	1.8 (2.1)	0.4	2,580	4.0–5.0
Crust	Sheeted dikes/gabbroic intrusions	2.3 (2.6)	0.2	2,890	5.0–6.7
Ocean	Gabbro	2.3 (3.0)	0.2	3,150	7.1–7.6
Mantle	Peridotite continent	3.95 (3.2)	0.03	3,330	8.0–8.3
	Peridotite ocean	3.95 (3.2)	0.03	3,180	7.8

velocity to values >8 km/s is observed in deep seismic data marking the transition to the lithospheric mantle (Bauer et al., 2000; DEKORPBASIN and Group, 1999; Fowler, 1996; Hirsch et al., 2009a; Levander et al., 2006; Meissner and Group, 1991; Thybo and Nielsen, 2009; Turcotte and Schubert, 2002; Weber and group, 2009). This is related to the change in composition from various crustal rocks to the mantle consisting mainly of peridotite.

As less dense sediments replace a denser crystalline crust, sedimentary basins should be characterized by a negative Bouguer anomaly with highest amplitude in the basin center after re-equilibration of the isotherm. Many basins, however, are characterized by a long-wavelength negative Bouger anomaly with positive anomalies superimposed. This observation, together with strong variations in the crustal velocity structure beneath sedimentary basins challenge the classical concepts of crustal thinning and related Moho uplift in that they commonly occur in concert with a flat Moho (Thybo and Nielsen, 2009).

Crustal bodies with seismic velocities >7 km, generally referred to as "high-velocity bodies," are observed in the lower crust beneath many intracontinental rift structures as, for example, the Baikal and the Kenya Rifts (Thybo and Nielsen, 2009), beneath intracontinental basins as, for example, in the Danish Basin (Nielsen and Thybo, 2006), and beneath extended passive margins as, for example, at the conjugate margins of the North and South Atlantic (Bauer et al., 2000; Contrucci et al., 2004; Faleide et al., 2008; Franke et al., 2007; Hirsch et al., 2009a; Mjelde et al., 2002; Mjelde et al., 2005; Sibuet et al., 2007). Gravity analysis indicates that these bodies are also characterized by higher densities than the average crust (Fernàndez et al., 2010, 2005; Franke et al., 2007; Hirsch et al., 2009a; Maystrenko and Scheck-Wenderoth, 2009; Osmundsen and Ebbing, 2008).

Differential thinning with depth (Huismans et al., 2005; Huismans and Beaumont, 2008; Kusznir and Ziegler, 1992; Lavier et al., 1999; Lavier and Steckler, 1997; Pascal and Cloetingh, 2002; Sclater and Christie, 1980; Steckler and Watts, 1978), emplacement of magmatic material in the crust during the thinning process or magmatic underplating (Thybo and Nielsen, 2009; Burg and Gerya, 2008) are some possible processes during basin formation and evolution that can be responsible for the actual configuration of the crust beneath a specific basin. The relative contributions of the processes vary strongly for different basins, and, accordingly, their relative importance in general terms is still under debate.

Moreover, structural inheritance may have a fundamental role in defining the crustal and lithosphere structure. Successive suturing of different plates in Earth history may result in a mosaic of crustal domains with contrasting physical properties, possibly also including lower crustal bodies predating the rifting process (Ebbing et al., 2009; Faleide et al., 2008; Plomerová and Babuska, 2010; Scheck-Wenderoth and Lamarche, 2005; van Wijk, 2005; Vauchez et al., 1998). In basins developing on such a substrate such older rheological discontinuities may be reactivated to localize deformation in areas of reduced strength, thus also facilitating discontinuous stretching with depth.

The configuration of the mantle lithosphere

Much less is known on the configuration of the mantle lithosphere below sedimentary basins due to the limited amount of direct observations. Though reflections and lateral variations in p-wave velocities in the lithospheric mantle have been observed in deep seismic data, the energy used in active seismic experiments is generally not sufficient to reach depth intervals below the Moho. Also, the gravity signal from the lithospheric mantle is difficult to isolate from the cumulative signal of the entire lithosphere. Recent development in the acquisition and evaluation of passive *seismological* data (Dalton and Faul, 2010; Fishwick, 2010; Geissler et al., 2010; Heintz and Kennett, 2005; Hieronymus and Goes, 2010; Plomerová and Babuska, 2010; Priestley et al., 2006; Ritzmann and Faleide, 2009; Zhang and Lay, 1996), electromagnetic and *magnetotelluric* methods (see entries *Magnetotelluric Interpretation*; *Magnetotelluric Data Processing*) (Jones et al., 2010), together with remote sensing gravity observation (Kaban et al., 2003; Schotman et al., 2009), as well as geochemical data (O'Reilly and Griffin, 2010; Trumbull et al., 2002; Wang, 2010) and thermal studies (Artemieva, 2006, 2009; Hasterok and Chapman, 2007; Hieronymus and Goes, 2010) indicates that the lithospheric mantle is less homogenous than previously thought. Apart from strong differences in thickness, also considerable lateral variations in surface *heat flow* (see entry *Heat Flow Measurements, Continental*) are observed, with a positive correlation between the two. The lithosphere-asthenosphere boundary (LAB) may be located deeper than 250 km in cratonic areas, lies at around 100 km beneath Phanerozoic continental domains and old oceans, and is close to the seafloor at Mid Ocean Ridges.

Agreement is established that the LAB is a fundamental boundary in plate-tectonic theory that separates the rigid plates from ductile convecting material below the plates (Artemieva, 2009; Eaton et al., 2009; Jones et al., 2010). Three broad definitions in terms of a mechanical boundary layer, a thermal boundary layer, and a chemical boundary layer are based on different types of data and partly in geometrical conflict. For the evolution of sedimentary basins, the depth of the thermal LAB is especially relevant as it determines the thermal and mechanical state of the lithosphere subjected to any of the basin-forming mechanisms. Systematic mapping of lithosphere thickness beneath sedimentary basins is, however, still lacking. The thermal LAB is interpreted as an isotherm of about 1,300°C and its depth corresponds to the depth where a continental conductive geotherm intersects the mantle adiabat. This is corroborated by heat flow inversion studies in *continental lithosphere* (see entry *Lithosphere, Continental*) (Artemieva, 2009), by combined thermal and gravity

modelling (Fernàndez et al., 2010, 2005; Hasterok and Chapman, 2007; Hyndman et al., 2009; Scheck-Wenderoth and Maystrenko, 2008) and by cooling models in *oceanic lithosphere* (see entry *Lithosphere, Oceanic*) (Crosby et al., 2006; McKenzie, 1978; Sclater, 2003; Stein and Stein, 1992).

Isostatically, thinning of the lithospheric mantle results in net surface uplift as heavier lithospheric mantle is replaced by less dense asthenospheric material (active rift model). Only if the interaction with surface processes takes place (erosion), net subsidence takes place due to subsequent cooling of the asthenospheric material.

Heat flow in sedimentary basins

The thermal field in sedimentary basins has been of primary interest in the exploration of fossil fuels (Welte et al., 1997) and gains increasing importance for the use of geothermal energy (Huenges, 2010). Moreover, together with composition, temperature is a main controlling factor for the rheological behavior of the lithosphere and accordingly its deformation. The hotter the lithosphere, the weaker is its rheology and the easier it is thinned to form a sedimentary basin. Accordingly, the main controlling factors for the thermal state of a basin are its plate-tectonic setting and its evolutionary stage.

In terms of observables, surface *heat flow* measurements and temperatures measured in wells characterize the present-day thermal state of a basin, whereas the maturity of organic matter (Welte et al., 1997) and thermochronological data (Andriessen, 1995; Braun, 2005; Braun and van der Beek, 2004; Kounov et al., 2007; Simoes et al., 2010; Van der Beek, 2007; Willenbring and von Blanckenburg, 2010) provide a record of the thermal history.

Different families of integrated process-oriented models attempt to reproduce these observables and indicate that several processes contribute to the heat flow in sedimentary basins. There is a first-order contrast in thermal properties between sedimentary and crystalline rocks, in that sediments are thermally less conductive and produce less radiogenic heat than crystalline crustal rocks (Table 1). The amount of heat entering the basin at the base of the sediment fill depends on the thickness and composition of the crystalline crust as well as on the depth of the thermal LAB. The mafic components of the crust and the lithospheric mantle are characterized by a high thermal conductivity but low *radiogenic heat production*. Accordingly, the shallower the LAB and the thicker and radiogenic the crystalline crust, the more heat arrives at the base of the sediments. Due to their higher porosities, the thermally low-conductive sediments act as a thermal blanket, causing heat storage in the basin (Cacace et al., 2010; Theissen and Rüpke, 2009; Van Wees et al., 2009). In addition, the sediments contribute a modest but, in the sum, considerable amount of radiogenic heat to the system. Also within the sediment fill, the thermal properties may vary (Table 1) with the thermal conductivity of salt being two times larger than that of clastic sediments. The upper part of sedimentary basins may store paleoclimatic signals of previous glaciations as present-day permafrost down to more than 700 m depth attest.

In response to the distribution of thermal parameters long-wavelength variations in temperatures in sedimentary basins (scale of hundreds of kilometers) are determined by the crustal structure and composition as well as by the thickness of the lithosphere (Cacace et al., 2010; Hasterok and Chapman, 2007; Hyndman et al., 2009; Scheck-Wenderoth and Maystrenko, 2008; Sclater, 2003). In contrast, the short-wavelength pattern of temperature distribution (scale of kilometers) is controlled by the internal configuration of the sediment fill.

Heat transfer by conduction is assumed to dominate the transport of heat in the lithosphere in contrast to the *convecting mantle* (see entry *Mantle Convection*). While geological and geochemical data proved that additional hydrothermal heat transport takes place along faults in the lithosphere, there is an ongoing debate on how far the thermal field in sedimentary basins is additionally influenced by convective heat transport due to moving pore fluids. The inherent vertical anisotropy due to the layered nature of sediments counteracts the development of free convection on a basin-wide scale (Bjorlykke, 2010) but local convection is indicated by geochemical data from hydrogeological studies and models of coupled fluid and heat transport (Magri et al., 2009).

Summarizing, the heat flow regime in a specific basin may vary spatially though some ranges can be given for different types of basins (Allen and Allen, 2005): Typically, the surface *heat flow* in intracontinental basins varies between 40 and 70 mW/m^2, can be up to 150 mW/m^2 in active rifts or close to volcanic arcs, and can reach values higher than 180 mW/m^2 at mid ocean ridges and oceanic rifts. In oceanic basins, surface *heat flow* decreases with increasing distance from the mid ocean ridges according to the cooling of the lithosphere with age (Parsons and Sclater, 1977; Stein and Stein, 1992). The surface *heat flow* at passive margins depends on the age of the adjacent oceanic lithosphere. At young margins as in the 55-my-old North Atlantic, the surface heat flow is still controlled by the cooling of the oceanic lithosphere that is considerably thinner than the lithosphere of the continental margin. This step in the thermal LAB is consistent with observed heat flow values increasing from 45 mW/m^2 at the continental side to 80 mW/m^2 at the oceanic side of the margin (Ritter et al., 2004; Scheck-Wenderoth and Maystrenko, 2008). At older passive margins as in the 130-my-old South Atlantic, an opposite trend is observed (Goutorbe and Bonneville, 2008; Hirsch et al., 2009b) with about 45 mW/m^2 at the oceanic side of the margin to up 65 mW/m^2 at the continental side. As the oceanic lithosphere had sufficient time to cool and thicken, the depth of the thermal LAB is continuous and similar between the continental and the oceanic part of the system. Instead, the radiogenic heat contribution of the *continental crust* (see entry *Earth's*

Structure, Continental Crust) thickening toward the continent becomes the controlling factor for heat input from the lithosphere.

Flexural basins are generally colder than extensional basins with average *heat flow* values around 40 mW/m^2 in oceanic trenches and foreland basins. The thermal signature of strike-slip basins is highly variable as it depends on the attributes of the lithosphere on which they develop and if hydrothermal heat transport takes place along the fault zones.

Summary

The *geodynamics* of sedimentary basins results from the interaction of a variety of processes acting on different spatial and temporal scales. Geophysical and geological observations indicate that intra-plate stress and heat input from the asthenospheric mantle are first-order controlling factors. On the *continental lithosphere* (see entry *Lithosphere, Continental*), structural inheritance and its impact on rheology is a third major player controlling the development of sedimentary basins. Intra-plate stress determines if an extensional or a compressional basin is formed. Temperature dominantly controls the rheological behavior of the lithosphere in addition to composition. These two parameters determine if the lithosphere is thinned uniformly, depth dependently, and in a symmetric or asymmetric mode. The more layered the lithosphere rheology is, the more discontinuous and asymmetric the stretching process will be. In contrast, a lithosphere deforms in a uniform stretching mode if the crust is strongly coupled to the mantle and no significant vertical or horizontal rheology contrasts are present. In both cases, the magnitude and rates of effective stress as well as the magnitude of the heat anomaly determine if stretching takes place slow enough to allow for cooling related strain hardening or fast enough to result in continental breakup.

These deep factors interact continuously with surface processes such as deposition or erosion of sediments influenced by climatic conditions. Deposition leads to *isostatic* loading and enhanced subsidence whereas erosion results in isostatic unloading and enhanced uplift.

Bibliography

Afonso, J. C., et al., 2010. On the Vp/Vs-Mg# correlation in mantle peridotites: implications for the identification of thermal and compositional anomalies in the upper mantle. *Earth and Planetary Science Letters*, **289**(3–4), 606–618.

Allen, P., and Allen, J., 2005. *Basin Analysis. Principles and Applications*. Oxford: Blackwell Science, p. 560.

Andriessen, P. A. M., 1995. Fission-track analysis: principles, methodology and implications for tectono-thermal histories of sedimentary basins, orogenic belts, and continental margins. *Geologie en Mijnbouw*, **74**, 1–12.

Artemieva, I. M., 2006. Global 1° × 1° thermal model TC1 for the continental lithosphere: implications for lithosphere secular evolution. *Tectonophysics*, **416**(1–4), 245–277.

Artemieva, I. M., 2009. The continental lithosphere: reconciling thermal, seismic, and petrologic data. *Lithos*, **109**, 23–46.

Bauer, K., et al., 2000. Deep structure of the Namibia continental margin as derived from integrated geophysical studies. *Journal of Geophysical Research*, **105**(B11), 25,829–25,853.

Bayer, U., Grad, M., Pharaoh, T. C., Thybo, H., Guterch, A., Banka, D., Lamarche, J., Lassen, A., Lewerenz, B., Scheck, M., and Marotta, A. M., 2002. The southern margin of the East European Craton: new results from seismic sounding and potential fields between the North Sea and Poland. *Tectonophysics*, **360**, 301–314.

Bjorlykke, K., 2010. *Heat Transport in Sedimentary Basins, Petroleum Geoscience*. Berlin: Springer, pp. 253–257.

Braun, J., 2005. Quantitative constraints on the rate of landform evolution derived from low-temperature thermochrohonology, low-temperature thermochronology: techniques, interpretations, and applications. *Reviews in Mineralogy and Geochemistry*, **58**, 351–374.

Braun, J., and van der Beek, P., 2004. Evolution of passive margin escarpments: what can we learn from low-temperature thermochronology? *Journal of Geophysical Research-Earth Surface*, **109**, F04009.

Braun, J., et al., 2008. DOUAR: a new three-dimensional creeping flow numerical model for the solution of geological problems. *Physics of the Earth and Planetary Interiors*, **171**(1–4), 76–91.

Buck, W. R., 1991. Modes of continental lithospheric extension. *Journal of Geophysical Research*, **96**(B12), 20.161–20.178.

Burg, J. P., and Gerya, T. V., 2008. Modelling intrusion of mafic and ultramafic magma into the continental crust: numerical methodology and results. *Bollettino Della Societa Geologica Italiana*, **127**(2), 199–203.

Burov, E., and Cloetingh, S., 2009. Controls of mantle plumes and lithospheric folding on modes of intraplate continental tectonics: differences and similarities. *Geophysical Journal International*, **178**(3), 1691–1722.

Burov, E. B., and Diament, M., 1992. Flexure of the continental lithosphere with multilayered rheology. *Geophysical Journal International*, **109**, 449–468.

Cacace, M., Kaiser, B. O., Lewerenz, B., and Scheck-Wenderoth, M., 2010. Geothermal energy in sedimentary basins: what we can learn from regional numerical models. *Chemie Der Erde-Geochemistry*, **70**, 33–46.

Catuneanu, O., 2006. *Principles of Sequence Stratigraphy*. New York: Elsevier, p. 386.

Cermak, V., and Rybach, L., 1982. Thermal properties: Thermal conductivity and specific heat of minerals and rocks. In Angenheister, G. (ed.), *Landolt-Börnstein, New Series, Geophysics and Space Research*. Berlin: Springer, pp. 305–343.

Christensen, N. I., and Mooney, W. D., 1995. Seismic velocity structure and composition of the continental crust: a global view. *Journal of Geophysical Research*, **100**(B7), 9761–9788.

Cloetingh, S., and Burov, E., 2010. Lithospheric folding and sedimentary basin evolution: a review and analysis of formation mechanisms. *Basin Research*, doi:10.1111/j.1365-2117.2010.00490.x.

Cloetingh, S., Boldreel, L. O., Larsen, B. T., Heinesen, M., and Mortensen, L., 1998. Tectonics of sedimentary basin formation: models and constraints. *Tectonophysics*, **300**(I–IV), 1–11.

Contrucci, I., et al., 2004. Deep structure of the West African continental margin (Congo, Zaire, Angola), between 5 degrees S and 8 degrees S, from reflection/refraction seismics and gravity data. *Geophysical Journal International*, **158**(2), 529–553.

Cooper, M. A., and Williams, G. D., 1989. *Inversion Tectonics. Geological Society Special Publication Classics*. London: The Geological Society of London.

Crosby, A. G., McKenzie, D., and Sclater, J. G., 2006. The relationship between depth, age and gravity in the oceans. *Geophysical Journal International*, **166**(2), 553–573.

Dalton, C. A., and Faul, U. H., 2010. The oceanic and cratonic upper mantle: clues from joint interpretation of global velocity and attenuation models. *Lithos*, **120**(1–2), 160–172.
DEKORP-BASIN and Group, R., 1999. Deep crustal structure of the Northeast German basin: new DEKORP-BASIN'96 deep-profiling results. *Geology*, **27**(1), 55–58.
Eaton, D. W., et al., 2009. The elusive lithosphere-asthenosphere boundary (LAB) beneath cratons. *Lithos*, **109**(1–2), 1–22.
Ebbing, J., Gernigon, L., Pascal, C., Olesen, O., and Osmundsen, P. T., 2009. A discussion of structural and thermal control of magnetic anomalies on the mid-Norwegian margin. *Geophysical Prospecting*, **57**(4), 665–681.
Einsele, G., Ricken, W., and Seilacher, A., 1991. Cycles and Events in Stratigraphy - Basic Concepts and Terms. In Einsele, G., Ricken, W., and Seilacher, A. (eds.), *Cycles and Events in Stratigraphy*. Berlin: Springer, pp. 1–19.
Faleide, J. I., et al., 2008. Structure and evolution of the continental margin off Norway and the Barents Sea. *Episodes*, **31**(1), 82–91.
Fernandez, M., et al., 2005. Lithospheric structure of the mid-Norwegian margin; comparison between the More and Voring margins. *Journal of the Geological Society of London*, **162**(6), 1005–1012.
Fernàndez, M., Afonso, J. C., and Ranalli, G., 2010. The deep lithospheric structure of the Namibian volcanic margin. *Tectonophysics*, **481**(1–4), 68–81.
Fishwick, S., 2010. Surface wave tomography: imaging of the lithosphere-asthenosphere boundary beneath central and southern Africa? *Lithos*, **120**(1–2), 63–73.
Förster, A., and Förster, H.-J., 2000. Crustal composition and mantle heat flow: implications from surface heat flow and radiogenic heat production in the Variscian Erzbgebirge. *Journal of Geophysical Research*, **105**(B12), 917–938.
Fowler, C. M. R., 1996. *The Solid Earth*. Cambridge: Cambridge University Press, p. 472.
Franke, D., Neben, S., Ladage, S., Schreckenberger, B., and Hinz, K., 2007. Margin segmentation and volcano-tectonic architecture along the volcanic margin off Argentina/Uruguay, South Atlantic. *Marine Geology*, **244**(1–4), 46–67.
Garcia-Castellanos, D., 2002. Interplay between lithospheric flexure and river transport in foreland basins. *Basin Research*, **14**(2), 89–104.
Geissler, W. H., Sodoudi, F., and Kind, R., 2010. Thickness of the central and eastern European lithosphere as seen by S receiver functions. *Geophysical Journal International*, **181**(2), 604–634.
Goutorbe, B. L. F., and Bonneville, A., 2008. The thermal regime of South African continental margins. *Earth and Planetary Science Letters*, **267**(1–2), 256–265.
Hasterok, D., and Chapman, D. S., 2007. Continental thermal isostasy: 1. Methods and sensitivity. *Journal of Geophysical Research*, **112**(B6), B06414.
Haxby, W. F., Turcotte, D. L., and Bird, J. M., 1976. Thermal and mechanical evolution of the Michigan Basin. *Tectonophysics*, **36**(1–3), 57–75.
Heine, C., Dietmar Müller, R., Steinberger, B., and Torsvik, T. H., 2008. Subsidence in intracontinental basins due to dynamic topography. *Physics of the Earth and Planetary Interiors*, **171**(1–4), 252–264.
Heintz, M., and Kennett, B. L. N., 2005. Continental scale shear wave splitting analysis: investigation of seismic anisotropy underneath the Australian continent. *Earth and Planetary Science Letters*, **236**(1–2), 106–119.
Hieronymus, C. F., and Goes, S., 2010. Complex cratonic seismic structure from thermal models of the lithosphere: effects of variations in deep radiogenic heating. *Geophysical Journal International*, **180**(3), 999–1012.
Hirsch, K. K., Bauer, K., and Scheck-Wenderoth, M., 2009a. Deep structure of the western South African passive margin – results of a combined approach of seismic, gravity and isostatic investigations. *Tectonophysics*, **470**(1–2), 57–70.
Hirsch, K. K., Scheck-Wenderoth, M., Maystrenko, Y., Sippel, J., and Anonymous, 2009. A lithospheric 3D temperature study from the South Atlantic. *Geophysical Research Abstracts*, **11**, EGU2009-8053.
Hudec, M. R., and Jackson, M. P. A., 2007. Terra infirma: understanding salt tectonics. *Earth Science Reviews*, **82**(1–2), 1–28.
Huenges, E., 2010. *Geothermal Energy Systems: Exploration, Development and Utilization*. Weinheim: Wiley-VCH, p. 464.
Huismans, R. S., and Beaumont, C., 2008. Complex rifted continental margins explained by dynamical models of depth-dependent lithospheric extension. *Geology*, **36**(2), 163–166.
Huismans, R. S., Buiter, S. J. H., and Beaumont, C., 2005. Effect of plastic-viscous layering and strain softening on mode selection during lithospheric extension. *Journal of Geophysical Research*, **110**, B02406.
Hyndman, R. D., Currie, C. A., Mazzotti, S., and Frederiksen, A., 2009. Temperature control of continental lithosphere elastic thickness, Te vs Vs. *Earth and Planetary Science Letters*, **277**(3–4), 539–548.
Jones, A. G., Plomerova, J., Korja, T., Sodoudi, F., and Spakman, W., 2010. Europe from the bottom up: a statistical examination of the central and northern European lithosphere-asthenosphere boundary from comparing seismological and electromagnetic observations. *Lithos*, **120**(1–2), 14–29.
Kaban, M. K., Schwintzer, P., Artemieva, I. M., and Mooney, W. D., 2003. Density of the continental roots; compositional and thermal contributions. *Earth and Planetary Science Letters*, **209**(1–2), 53–69.
Kaus, B. J. P., Connolly, J. A. D., Podladchikov, Y. Y., and Schmalholz, S. M., 2005. Effect of mineral phase transitions on sedimentary basin subsidence and uplift. *Earth and Planetary Science Letters*, **233**(1–2), 213–228.
Kley, J., Franzke, H.-J., Jähne, F., Krawczyk, C., Lohr, T., Reicherter, K., Scheck-Wenderoth, M., Sippel, J., Tanner, D., and van Gent, H., 2008. Strain and stress. In Littke, R. B., Baeyer, U., Gajewski, D., and Nelskamp, S. (eds.), *Dynamics of Complex Intracontinental Basins. The Central European Basin System*. Berlin: Springer, pp. 97–124.
Kounov, A., et al., 2007. Present denudation rates at selected sections of the South African escarpment and the elevated continental interior based on cosmogenic 3He and 21Ne. *South African Journal of Geology*, **110**(2/3), 235–248.
Kusznir, N. J., and Ziegler, P. A., 1992. The mechanics of continental extension and sedimentary basin formation: a simple-shear/pure-shear flexual cantilever model. *Tectonophysics*, **215**, 117–131.
Lavier, L. L., and Steckler, M. S., 1997. The effect of sedimentary cover on the flexural strength of continental lithosphere. *Nature*, **389**, 476–479.
Lavier, L. L., Buck, W. R., and Poliakov, A. N. B., 1999. Self-consistent rolling-hinge model for the evolution of large-offset low-angle normal faults. *Geology*, **27**(12), 1127–1130.
Levander, A., Niu, F., Lee, C.-T. A., and Cheng, X., 2006. Imag(in)ing the continental lithosphere. *Tectonophysics*, **416**(1–4), 167–185.
Lister, G. S., Etheridge, M. A., and Symonds, P. A., 1986. Datachement faulting and the evolution of passive margins. *Geology*, **14**, 246–250.
Littke, R., Bayer, U., Gajewski, D., and Nelskamp, S., 2008. *Dynamics of Complex Intracontinental Basins: The Central European Basin System*. Berlin: Springer.
Magri, F., et al., 2009. Salty groundwater flow in the shallow and deep aquifer systems of the Schleswig-Holstein area (North German Basin). *Tectonophysics*, **470**(1–2), 183–194.
Maystrenko, Y., and Scheck-Wenderoth, M., 2009. Density contrasts in the upper mantle and lower crust across the continent-ocean transition: constraints from 3-D gravity modelling at the

Norwegian margin. *Geophysical Journal International*, **179**(1), 536–548.
Mazur, S., Scheck-Wenderoth, M., and Krzywiec, P., 2005. Different modes of the late cretaceous-early tertiary inversion in the North German and Polish basins. *International Journal of Earth Sciences*, **94**(5–6), 782–798.
McKenzie, D., 1978. Some remarks on the development of sedimentary basins. *Earth and Planetary Science Letters*, **40**, 25–32.
Meissner, R., and Group, D. R., 1991. The DEKORP Surveys: Major results in tectonic and reflective styles. In Meissner, R., et al. (eds.), *Continental Lithosphere: Deep Seismic Reflections*. Washington, DC: American Geophysical Union, pp. 69–76.
Mjelde, R., et al., 2002. Lower crustal seismic velocity-anomalies; magmatic underplating or serpentinized peridotite? Evidence from the Voring Margin, NE Atlantic. *Marine Geophysical Researches*, **23**(2), 169–183.
Mjelde, R., et al., 2003. V (sub p)/V (sub s) ratio along the Voring Margin, NE Atlantic, derived from OBS data; implications on lithology and stress field. *Tectonophysics*, **369**(3–4), 175–197.
Mjelde, R., et al., 2005. Continent-ocean transition on the Vøring Plateau, NE Atlantic, derived from densely sampled ocean bottom seismometer data. *Journal of Geophysical Research*, **110**, B05101, doi:10.1029/2004JB003026.
Nielsen, L., and Thybo, H., 2006. Identification of crustal and upper mantle heterogeneity by modelling of controlled-source seismic data. *Tectonophysics*, **416**(1–4), 209–228.
O'Reilly, S. Y., and Griffin, W. L., 2010. The continental lithosphere-asthenosphere boundary: can we sample it? *Lithos*, **120**(1–2), 1–13.
Osmundsen, P. T., and Ebbing, J., 2008. Styles of extension offshore mid-Norway and implications for mechanisms of crustal thinning at passive margins. Tectonics, **27**(6), TC6016.
Parsons, B., and Sclater, J. G., 1977. An analysis of the variation of ocean floor bathymetry and heat flow with age. *Journal of Geophysical Research*, **82**(5), 803–827.
Pascal, C., and Cloetingh, S. A. P. L., 2002. Rifting in heterogeneous lithosphere: inferences from numerical modelling of the northern North Sea and the Oslo Graben. *Tectonics*, **21**(6), 1060–1071.
Pérez-Gussinyé, M., Morgan, J. P., Reston, T. J., and Ranero, C. R., 2006. The rift to drift transition at non-volcanic margins: insights from numerical modelling. *Earth and Planetary Science Letters*, **244**(1–2), 458–473.
Petrunin, A. G., and Sobolev, S. V., 2008. Three-dimensional numerical models of the evolution of pull-apart basins. *Physics of the Earth and Planetary Interiors*, **171**(1–4), 387–399.
Plomerová, J., and Babuska, V., 2010. Long memory of mantle lithosphere fabric – European LAB constrained from seismic anisotropy. *Lithos*, **120**(1–2), 131–143.
Priestley, K., McKenzie, D., and Debayle, E., 2006. The state of the upper mantle beneath southern Africa. *Tectonophysics*, **416** (1–4), 101–112.
Ritter, U., Zielinski, G. W., Weiss, H. M., Zielinski, R. L. B., and Sættem, J., 2004. Heat flow in the Vøring basin, mid-Norwegian shelf. *Petroleum Geoscience*, **10**, 353–365.
Ritzmann, O., and Faleide, J. I., 2009. The crust and mantle lithosphere in the Barents Sea/Kara Sea region. *Tectonophysics*, **470** (1–2), 89–104.
Roure, F., Cloetingh, S., Scheck-Wenderoth, M., and Ziegler, P.A., 2009. Achievements and challenges in sedimentary basin dynamics: a review. In Cloetingh, S., and Negendank, J., (eds.), *New Frontiers in Integrated Solid Earth Sciences.* International Year of Planet Earth. Springer, Dordrecht, pp. 145–233.
Royden, L., and Keen, C. E., 1980. Rifting process and thermal evolution of the continental margin of eastern Canada determined from subsidence curves. *Earth and Planetary Science Letters*, **51**, 343–361.
Sandiford, M., 1999. Mechanics of basin inversion. *Tectonophysics*, **305**, 109–120.
Scheck-Wenderoth, M., and Lamarche, J., 2005. Crustal memory and basin evolution in the Central European Basin System – new insights from a 3D structural model. *Tectonophysics*, **397** (1–2), 143–165.
Scheck-Wenderoth, M., and Maystrenko, Y., 2008. How warm are passive continental margins? A 3-D lithosphere-scale study from the Norwegian margin. *Geology*, **36**(5), 419–422.
Scheck-Wenderoth, M., Maystrenko, Y., Huebscher, C., Hansen, M., and Mazur, S., 2008. *Dynamics of Salt Basins. Dynamics of Complex Intracontinental Basins; the Central European Basin System*. Berlin: Springer, pp. 307–322.
Schotman, H. H. A., Vermeersen, L. L. A., Wu, P., Drury, M. R., and De Bresser, J. H. P., 2009. Constraints on shallow low-viscosity zones in Northern Europe from future GOCE gravity data. *Geophysical Journal International*, **178**(1), 65–84.
Sclater, J. G., 2003. *Heat Flow Under the Oceans, Plate Tectonics; an Insider's History of Modern Theory of the Earth*. Boulder: Westview Press.
Sclater, J. G., and Christie, P. A. F., 1980. Continental stretching: an explanation of the post-mid-cretaceous subsidence of the central North Sea basin. *Journal of Geophysical Research*, **85**(B7), 3711–3739.
Sibuet, J.-C., Srivastava, S., and Manatschal, G., 2007. Exhumed mantle-forming transitional crust in the Newfoundland-Iberia rift and associated magnetic anomalies. *Journal of Geophysical Research*, **112**(B6), B06105.
Simoes, M., Braun, J., and Bonnet, S., 2010. Continental-scale erosion and transport laws: a new approach to quantitatively investigate macroscale landscapes and associated sediment fluxes over the geological past. *Geochemistry Geophysics Geosystems*, **11**, Q09001.
Simon, K., Huismans, R. S., and Beaumont, C., 2009. Dynamical modelling of lithospheric extension and small-scale convection: implications for magmatism during the formation of volcanic rifted margins. *Geophysical Journal International*, **176** (Suppl.1), 327–350.
Smit, J., Brun, J. P., Cloetingh, S., and Ben-Avraham, Z., 2010. The rift-like structure and asymmetry of the Dead Sea Fault. *Earth and Planetary Science Letters*, **290**(1–2), 74–82.
Steckler, M. S., and Watts, A. B., 1978. Subsidence history and tectonic evolution of atlantic-type continental margins. *Earth and Planetary Science Letters*, **41**, 1–13.
Stein, C. A., and Stein, S., 1992. A model for the global variation in oceanic depth and heat flow with lithospheric age. *Nature (London)*, **359**(6391), 123–129.
Theissen, S., and Rüpke, L. H., 2009. Feedbacks of sedimentation on crustal heat flow: New insights from the Vøring Basin, Norwegian Sea. Basin Research, **22**(6), 976–990.
Thybo, H., and Nielsen, C. A., 2009. Magma-compensated crustal thinning in continental rift zones. *Nature*, **457**(7231), 873–876.
Torsvik, T. H., Rousse, S., Labails, C., and Smethurst, M. A., 2009. A new scheme for the opening of the South Atlantic ocean and the dissection of an Aptian salt basin. *Geophysical Journal International*, **177**(3), 1315–1333.
Trumbull, R. B., Sobolev, S. V., and Bauer, K., 2002. Petrophysical modeling of high seismic velocity crust at the Namibian volcanic margin. *Special Paper – Geological Society of America*, **362**, 221–230.
Turcotte, D. L., and Schubert, G., 2002. *Geodynamics*, 2nd edn. Cambridge: Cambridge University Press, p. 456.
Vail, P. R., Audemard, F., Bowman, S. A., Eisner, P. N., and Perez-Cruz, C., 1991. The Stratigraphic Signatures of Tectonics, Eustasy and Sedimentology - an Overview. In Einsele, G., Ricken, W., and Seilacher, A. (eds.), *Cycles and Events in Stratigraphy*. Berlin/Heidelberg: Springer, pp. 617–659.

Van der Beek, P., 2007. Thermochronological age – elevation profiles, denudation rates and relief development. *Geochimica Et Cosmochimica Acta*, **71**(15), A1055–A1055.
Van Wees, J. D., et al., 2009. Probabilistic tectonic heat flow modeling for basin maturation: assessment method and applications. *Marine and Petroleum Geology*, **26**(4), 536–551.
van Wijk, J. W., 2005. Role of weak zone orientation in continental lithosphere extension. *Geophysical Research Letters*, **32**, L02303.
Vauchez, A., Tommasi, A., and Barruol, G., 1998. Rheological heterogeneity, mechanical anisotropy and deformation of the continental lithosphere. *Tectonophysics*, **296**, 61–86.
Wang, Q., 2010. A review of water contents and ductile deformation mechanisms of olivine: implications for the lithosphere-asthenosphere boundary of continents. *Lithos*, **120**(1–2), 30–41.
Watts, A. B., 2001. *Isostasy and Flexure of the Lithosphere*. Cambridge: Cambridge University Press, p. 458.
Weber, M., and group, D.W., 2009. Anatomy of the Dead Sea Transform from lithospheric to microscopic scale. Reviews in Geophysics, **47**, RG2002, doi:10.1029/2008RG000264.
Welte, D. H., Horsfield, B., and Baker, D. R. (eds.), 1997. *Petroleum and Basin Evolution*. Berlin: Springer, p. 535.
Wernicke, B., 1981. Low-angle normal faults in the Basin and Range Province: nappe tectonics in an extending orogen. *Nature*, **291**, 645–648.
White, N., Thompson, M., and Barwise, T., 2003. Understanding the thermal evolution of deep-water continental margins. *Nature*, **426**(6964), 334–343.
Wijns, C., Weinberg, R., Gessner, K., and Moresi, L., 2005. Mode of crustal extension determined by rheological layering. *Earth and Planetary Science Letters*, **236**(1–2), 120–134.
Willenbring, J., and von Blanckenburg, F., 2010. Long-term stability of global erosion rates and weathering during late-Cenozoic cooling. *Nature*, **465**, 211–214.
Zhang, G.-B., and Bott, M. H. P., 2000. Modelling the evolution of asymmetrical basins bounded by high-angle reverse faults with application to foreland basins. *Tectonophysics*, **322**, 203–218.
Zhang, Y. S., and Lay, T., 1996. Global surface wave phase velocity variations. *Journal of Geophysical Research, Solid Earth and Planets*, **101**(B4), 8415–8436.
Zuber, M. T., Parmentier, E. M., and Fletcher, R. C., 1986. Extension of continental lithosphere: a model for two scales of basin and range deformation. *Journal of Geophysical Research*, **91**(B5), 4826–4838.

Cross-references

Continental Rifts
Deep Seismic Reflection and Refraction Profiling
Earth's Structure, Continental Crust
Geodynamics
Geophysical Well Logging
Gravity Anomalies, Interpretation
Heat Flow, Continental
Heat Flow, Seafloor: Methods and Observations
Isostasy
Lithosphere, Continental
Lithosphere, Mechanical Properties
Lithosphere, Oceanic
Lithosphere, Oceanic: Thermal Structure
Magnetic Anomalies, Interpretation
Magnetotelluric Interpretation
Mantle Convection
Mantle Plumes
Radiogenic Heat Production of Rocks
Seafloor Spreading
Seismic Imaging, Overview
Seismic Properties of Rocks
Seismology, Global Earthquake Model
Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat
Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity

SEISMIC ANISOTROPY

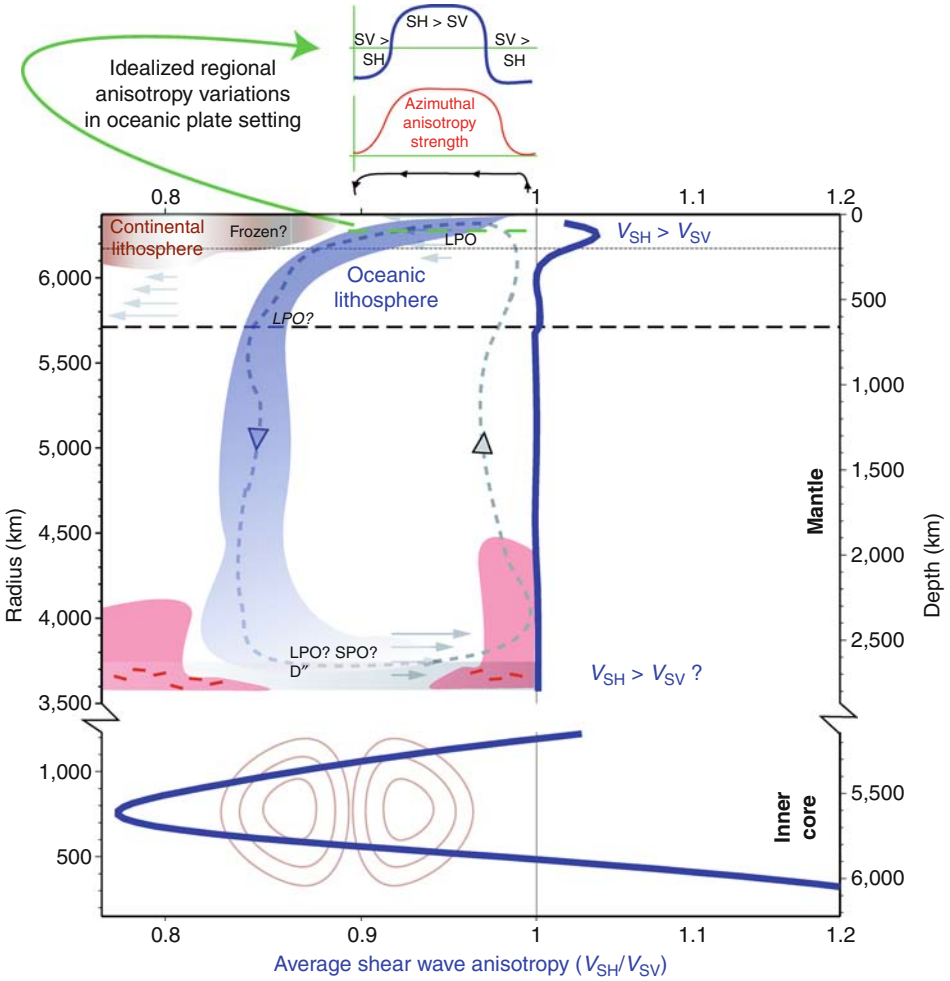
Thorsten W. Becker
Department of Earth Sciences, University of Southern California, Los Angeles, CA, USA

Definition

Seismic anisotropy refers to the directional dependence of seismic wave speeds and particle motion polarizations, as well as the splitting of normal modes, as caused by the elastic properties of rocks.

Introduction

Many of the minerals that make up Earth are intrinsically anisotropic. When rocks are progressively deformed over geologic timescales, the alignment of mineral grains (lattice-preferred orientation, LPO) can lead to bulk anisotropy of the rock. Bulk anisotropy can additionally be generated by an ordered assembly of individually isotropic materials of different wave speeds (shape-preferred orientation, SPO). Both types of anisotropy are found within the Earth; SPO anisotropy also highlights a fundamental ambiguity between isotropic heterogeneity and anisotropy. Seismic wave propagation through an anisotropic medium depends on the wavelengths over which a particular wave type averages, complicating the analysis of seismological data. Both LPO and SPO imply significantly different (up to ~10%) speeds for waves of different polarization or propagation directions, and velocity variations can be larger than those expected from compositional or thermal heterogeneity. Seismic anisotropy is therefore of fundamental importance for structural imaging studies. To get robust estimates of the quantities of interest for geodynamic interpretation, the trade-off between isotropic and anisotropic structure has to be considered. Seismic anisotropy provides a powerful link between seismic observations and the dynamic processes that shape the solid Earth, for example, convective flow in the case of LPO in the mantle (Figure 1, see *Mantle Convection*). However, anisotropic tomographic inversions are inherently more nonunique than isotropic imaging because a general anisotropic, linearly elastic medium has 21 independent components of the elasticity tensor, as opposed to 2 in the isotropic case. As a consequence of the increased number of parameters and the differences in how data sampling constrains isotropic and anisotropic structure, more data are needed for the same level of resolution in an anisotropic inversion. Typically, additional a priori constraints, such as from



Seismic Anisotropy, Figure 1 Cartoon of the possible distribution of whole Earth anisotropy (note scale break at CMB) with geodynamic interpretation (cf. Montagner, 2007); *dotted* and *dashed horizontal lines* indicate 200 and 660 km depths, respectively. The *heavy blue lines* in center show average radial anisotropy from Kustowski et al. (2008) for the mantle and from Beghein and Trampert (2003) for the inner core. Underneath oceanic plates, mantle flow is primarily horizontal, leading to LPO anisotropy alignment with $v_{SH} > v_{SV}$, while the radial mass transport associated with upwellings and downwellings may lead locally $v_{SV} > v_{SH}$. Beneath continental regions, both frozen in anisotropy from past tectonic deformation and asthenospheric anisotropy from present-day convection may contribute. The *gray, dashed, circular line* in the mantle indicates an idealized flow trajectory for a downwelling slab (*blue*) displacing a thermochemical "pile" (*red*) at the core mantle boundary (cf. Garnero and McNamara, 2008). This deep flow may affect CMB dynamics and lead to LPO and/or SPO anisotropy (Modified from Long and Becker, 2010).

petrology, are needed to narrow the parameter space. These complexities make the study of anisotropy in a geodynamic context inherently multidisciplinary, involving seismology, mineral physics, rock mechanics, and geodynamic modeling.

Basic mathematical description

Seismic anisotropy arises when the linear elasticity tensor C that connects stress, σ, and strain, ε, tensors as

$$\sigma_{ij} = \sum_{k=1}^{3} \sum_{l=1}^{3} C_{ijkl} \varepsilon_{kl} \tag{1}$$

does not simplify to the isotropic form

$$C_{ijkl} = \lambda \delta_{ij} \delta_{kl} + \mu \left(\delta_{ik} \delta_{jl} + \delta_{il} \delta_{jk} \right), \tag{2}$$

where λ and μ are the first and second (shear modulus) Lamé parameters, δ the Kronecker delta ($\delta_{ij} = 1$ for $i = j$, and 0 else), and the indices run over the three axes of the coordinate system, x_1, x_2, and x_3. In general, C has 81 components, out of which 21 are independent, and the most complex (triclinic) form of anisotropy requires the specification of all those components (e.g., Hearmon, 1961; Nye, 1985; Anderson, 1989). A typical simplification is to assume hexagonal anisotropy, which should capture most aspects of olivine LPO-related anisotropy in the upper

mantle (Montagner and Anderson, 1989; Browaeys and Chevrot, 2004; Becker et al., 2006).

If C is expressed in the 6 × 6 Voigt matrix, c, notation where c_{mn} relates to C_{ijkl} as $m = \delta_{ij}i + (1 - \delta_{ij})(9 - i - j)$, and $n = \delta_{kl}k + (1 - \delta_{kl})(9 - k - 1)$, then the five Love (1927) parameters for the transversely isotropic medium that results from hexagonal symmetry correspond to

$$\begin{aligned} A &= c_{11} = \rho v_{\mathrm{PH}}^2, \quad C = c_{33} = \rho v_{\mathrm{PV}}^2, \\ L &= c_{44} = \rho v_{\mathrm{SV}}^2, \quad N = c_{66} = \rho v_{\mathrm{SH}}^2, \quad \text{and} \\ F &= c_{11}. \end{aligned} \tag{3}$$

Here, ρ is density, and $v_{\mathrm{PH,\,PV}}$ the velocities for P waves propagating horizontally (H, in x_1–x_2 plane) and vertically (V, x_3-axis), respectively. For shear waves, $v_{\mathrm{SH,SV}}$ in Eq. 3 are the velocities for horizontally propagating waves that are horizontally or vertically polarized, respectively (see *Propagation of Elastic Waves: Fundamentals*). Transverse isotropy as a simplified description of material anisotropy is widely used and developed in exploration seismics (e.g., Thomsen, 1986). The top 220 km in the PREM 1D Earth model (Dziewoński and Anderson, 1981) are also transversely isotropic with vertical symmetry axis as in Eq. 3; such a medium is said to have bulk radial anisotropy. (Note that vertically propagating S waves in this case have the same velocity, v_{SV}, regardless of polarization direction.)

Different combinations of the Love parameters or c_{nm} are used in the literature (e.g., Babuška and Cara, 1991); for example, anisotropy in PREM is described by two measures of shear- and compressional-wave anisotropy strength

$$\xi = \left(\frac{v_{\mathrm{SH}}}{v_{\mathrm{SV}}}\right)^2 = \frac{N}{L} \quad \text{and} \quad \varphi = \left(\frac{v_{\mathrm{PV}}}{v_{\mathrm{PH}}}\right)^2 = \frac{C}{A}, \tag{4}$$

respectively, and the parameter $\eta = F/(A - 2L)$, which controls how velocities change between the vertical and horizontal directions. Another way to characterize the anisotropy of a transversely isotropic medium is due to Thomsen (1986), who defined

$$\begin{aligned} \varepsilon &= \frac{c_{11} - c_{33}}{2c_{33}} = \frac{A - C}{C} \quad \text{and} \\ \gamma &= \frac{c_{66} - c_{44}}{2c_{44}} = \frac{N - L}{L} \end{aligned} \tag{5}$$

as two different measures of the P and S wave anisotropy strength, respectively, and a combined parameter

$$\delta^* = \frac{1}{2c_{33}^2}\left[2\left(c_{13} + c_{44}\right)^2 - (c_{33} - c_{44})(c_{11} + c_{33} - 2c_{44})\right], \tag{6}$$

which, for weak anisotropy, simplifies to

$$\delta = \frac{(c_{13} + c_{44})^2 - (c_{33} - c_{44})^2}{2c_{33}(c_{33} - c_{44})}. \tag{7}$$

The δ parameter is important for near-vertical P wave propagation and identical to ε for "elliptical" anisotropy (Thomsen, 1986). Mainprice (2007) provides an account of other combinations of c_{mn} in use to characterize a transversely isotropic medium. Those differ, regrettably, quite substantially in different fields of study.

If the symmetry axis of the hexagonal anisotropy is in the horizontal plane, the anisotropy is termed azimuthal. This means that perpendicular fast and slow axes can be defined for horizontally propagating SV waves, where waves will propagate with $v_{\mathrm{SV1}} > v_{\mathrm{SV2}}$ along the fast and slow orientations, respectively. Any perturbations to phase velocity p, δp, due to general, but small anisotropy can be expressed as a series of isotropic, π-periodic, and π/2 periodic terms (e.g., Backus, 1965; Forsyth, 1975):

$$\frac{\delta p}{p} \approx A_0 + A_1 cos(2\Psi) + A_2 sin(2\Psi) + A_3 cos(4\Psi) + A_4 sin(4\Psi). \tag{8}$$

Here, Ψ is the azimuth of wave propagation, and Eq. 8 follows from the wave equation and the rank of the elasticity tensor (Smith and Dahlen, 1973). For mantle rocks, the 2Ψ terms are expected to be larger than the 4Ψ contributions for Rayleigh waves, which are predominantly sensitive to SV (Anderson, 1966; Montagner and Nataf, 1986). The 4Ψ terms are expected to be bigger than 2Ψ for Love waves, motivating the focus on Rayleigh waves for azimuthal anisotropy studies (see *Surface Waves*).

In general, the wave propagation effects of any elasticity tensor C can be analyzed by considering a plane wave $\mathbf{u} = \mathbf{a}\exp(-i\omega(t - \mathbf{s}\cdot\mathbf{x}))$ with ω angular frequency, and $\mathbf{u}$, $\mathbf{a}$, $\mathbf{s}$, and $\mathbf{x}$ the displacement, polarization, slowness, and location vectors, respectively (see *Propagation of Elastic Waves: Fundamentals*). $\mathbf{s}$ shall have the normalized direction $\check{\mathbf{s}}$ and length of $1/p$. Using the momentum equation $\ddot{u}_i = \partial_j \sigma_{ij}$, Eq. 1, the definition of the strain tensor, $\varepsilon_{ij} = \frac{1}{2}\left(\partial_i u_j + \partial_j u_i\right)$, C's symmetries, and defining $M_{ij} = \frac{1}{\rho} C_{ijkl}\,\hat{s}_j\,\hat{s}_l$, we can write

$$M\mathbf{a} = p^2\mathbf{a}, \tag{9}$$

which is an eigen problem for the symmetric matrix M. Equation 9 is called the Christoffel equation (e.g., Babuška and Cara, 1991). The eigen vector solutions correspond to one quasi-P and two quasi-S wave directions, and the associated eigen values are the density–velocity products ρv_{P}^2, ρv_{S2}^2, and ρv_{S1}^2. These quantities can be contoured for any elasticity tensor, for example as measured from single crystals, as a function of incidence angle and azimuth, to visualize the anisotropic properties of minerals or rocks in isolation (e.g., Mainprice, 2007). To generate more realistic synthetic body waves from three-dimensional (3D) variations in anisotropy,

semi-analytical reflectivity methods can be used if anisotropy is assumed to vary only with depth. However, for the general case of 3D variations of anisotropy on scales smaller than a Fresnel zone full, numerical wave propagation solutions are needed.

Seismological methods

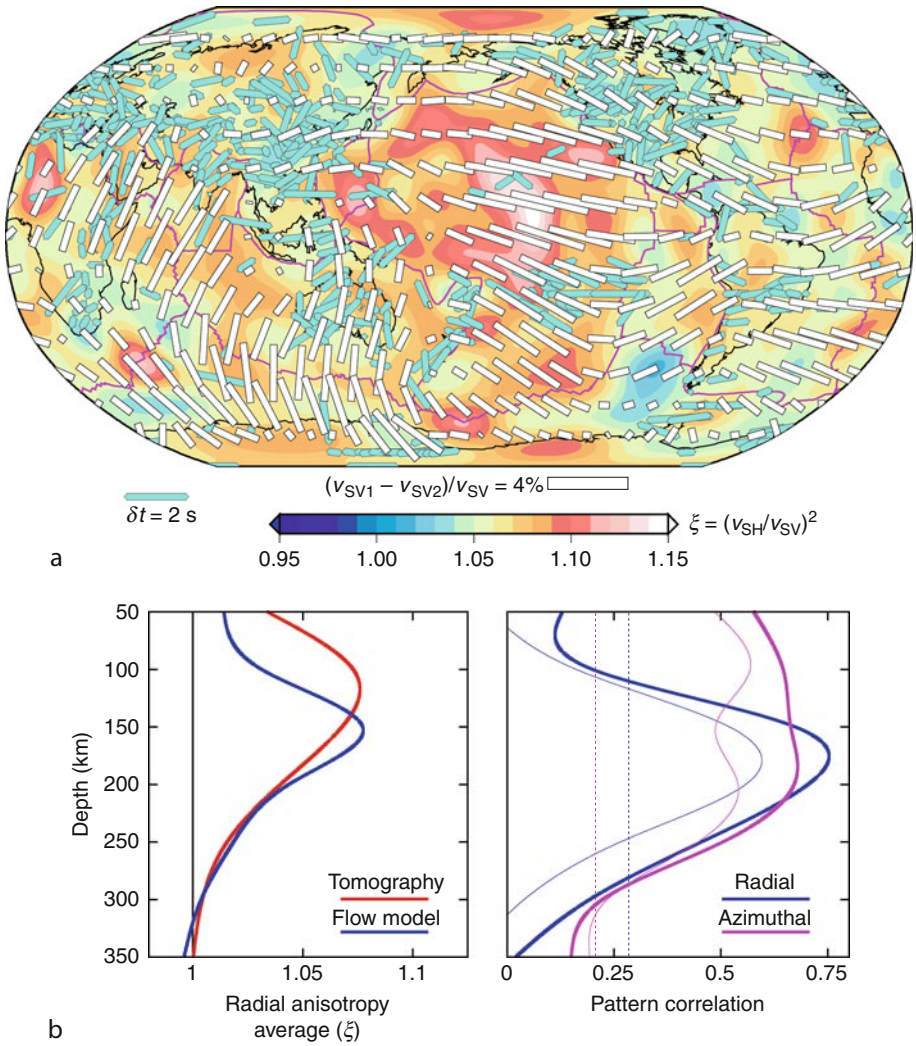
Seismic anisotropy can be detected in a number of ways that can be broadly classified into body- and surface-wave methods. The split of a shear wave into a fast and slow polarization direction as discussed for the solutions of Eq. 9 is akin to optical birefringence; it is exploited by the most popular method of measuring anisotropy, that utilizing *shear wave splitting* (Ando et al., 1983; Vinnik et al., 1984; Silver and Chan, 1991). For lithospheric and mantle applications, one typically considers near-vertical incidence SKS or SKKS core phases (see *Body Waves*), because the effects of any source-side anisotropy are removed by the S-to-P-to-S conversion upon traversal of the core. The most common splitting measurement consists of detecting the horizontal orientation of the fast (azimuth Ψ) pseudo-S wave from recorded particle motions, as well as determining the delay time δt between the arrival of the fast and slow *S* pulses (e.g., Savage, 1999; Long and Silver, 2009).

Shear wave splitting can be detected using a single earthquake measured at a single station, if wave propagation is out of a symmetry axis, and is a unique indicator for the presence of anisotropy along the ray path. However, only the highly idealized case of a single, transversely isotropic layer with horizontal symmetry axis can be directly interpreted in terms of Ψ and δt. Dipping symmetry axes, non-hexagonal anisotropy, or variations of anisotropy with depth will all cause a dependence of apparent splitting on back-azimuth (e.g., Schulte-Pelkum and Blackman, 2003). The nonlinear nature of the splitting measurement and layer splitting itself can lead to a bias of sensitivity toward the surface (∼one wavelength under the station), and not simple superposition (e.g., Saltzer et al., 2000). Such complexities make it imperative to strive for good back-azimuthal coverage, requiring the recording of several, suitable earthquakes, which is often a challenge given station-event geometry, or the duration of temporary deployments. If back-azimuth variations are detected, those can be used to make inferences about the variation of anisotropy with depth, which is undefined based on isolated measurements where anisotropy could, in principle, arise anywhere between the core mantle boundary (CMB) and the surface in the case of SKS splitting. If regional S arrivals are used, crossing ray paths can be used to infer 3D variations of anisotropy (e.g., Abt and Fischer, 2008). For teleseismic arrivals, the use of sensitivity kernels (e.g., Chevrot, 2006; Long et al., 2008) for the multichannel type of measurement of splitting holds great promise for resolving 3D anisotropy in regions for which close (closer than Fresnel zone width) station spacing is available. Broadly speaking, shear wave splitting is, however, a measurement with good lateral (∼50 km), but fairly poor depth resolution (Savage, 1999).

Another single, body-wave arrival method that follows from Eq. 9 is to use the orientation of the pseudo-*P* polarization, which may differ by more than 10° from along-ray, for *P* polarization anisotropy (Schulte-Pelkum et al., 2001), P_{pol}. A measurement of P_{pol} is sensitive to ∼half a wavelength underneath the station. If several, near-horizontal *P* paths with different azimuths are available, as in the case of the refracted P_n phase, which senses underneath the Moho, velocities can be plotted against azimuth to infer azimuthal anisotropy. This method was used for one of the earliest demonstrations of seismic anisotropy by Hess (1964), and a global comparison of P_n and SKS splitting can be found in Smith and Ekström (1999). The variations in delay times of teleseismic waves can also be used directly to infer anisotropy in the mantle lithosphere (e.g., Babuška et al., 1984; Bokelmann, 2002) by means of a tomographic inversion (see *Seismic Tomography*), but this method requires knowledge of the isotropic variations in wave speeds. This is a challenge for all tomographic approaches to anisotropy because there will always be a trade-off between isotropic and anisotropic heterogeneity in the absence of perfect data coverage (e.g., Tanimoto and Anderson, 1985). In terms of their depth-sensitivity, the body-wave methods can be ranked, from shallow to deep, as P_n, P_{pol}, SKS, and *P* delay times (Schulte-Pelkum and Blackman, 2003).

At crustal depths, anisotropy can additionally be detected by wide-angle refraction methods (e.g., Meissner et al., 2002, 2006). Orthogonal profiling, for example, may show a mismatch of derived seismic velocities, or a mismatch of refraction and reflection depths that can be interpreted in terms of anisotropic wave propagation. Receiver function methods (e.g., Park and Levin, 2002) (see *Seismic, Receiver Function Technique*) yield further evidence of crustal anisotropy from the existence of split *pS* conversions. Azimuthal variations in radial-transverse receiver function amplitudes are diagnostic of anisotropy versus tilted structure, and of the amount of anisotropy (e.g., Savage, 1998).

A wealth of information about anisotropy arises from the study of *surface waves*. The observation that Love waves, which mainly sense SH, travel faster than Rayleigh waves, which mainly sense SV due to their intrinsic polarities, implies the existence of a mean radial anisotropy in the upper mantle (Anderson, 1966; Dziewoński and Anderson, 1981). The existence of azimuthal anisotropy was documented for the Pacific by the study of Rayleigh waves (Forsyth, 1975), and Nataf et al. (1984) presented radially anisotropic, upper-mantle tomography. The current state of tomographic models for global azimuthal and radial anisotropy patterns is discussed by Montagner (2007). Surface wave propagation is dispersive, which allows the construction of 3D models of anisotropy (see *Earth's Structure, Global*; *Seismology, Global Earthquake Model*). The most easily measured phase-velocity period range for fundamental modes



Seismic Anisotropy, Figure 2 Global, uppermost mantle seismic anisotropy. (**a**) Seismological constraints: Radial (background, from Kustowski et al., 2008) and azimuthal anisotropy (white sticks indicating fast orientation, from Lebedev and van der Hilst, 2008) at 150 km, as well as SKS splitting (cyan sticks). SKS data are shown as a simple 5° average of the compilations by Fouch (2006) and Wüstefeld et al. (2009), but note that such averaging only provides a simplified view of azimuthal anisotropy (see text, and Schulte-Pelkum and Blackman, 2003, for example). (**b**) Radial anisotropy layer averages, on left, for the seismological model of Kustowski et al. (2008) and as predicted from the best-fitting geodynamic model of Becker et al. (2008). On right, pattern correlations up to spherical harmonic degree 8 between the same geodynamic model and radial (from Kustowski et al., 2008) and azimuthal (from Lebedev and van der Hilst, 2008) seismic tomography. *Heavy* and *thin lines* denote oceanic lithosphere only and global correlations, respectively. *Vertical, dashed lines* show 95% significance level (Modified from Long and Becker, 2010).

between ~50 and 150 s samples from the lithosphere down to ~300 km depth, and Figure 2a shows azimuthal and radial anisotropy at an asthenospheric depth of 150 km as well as a global compilation of SKS splitting results. At the shallow end, array methods (e.g., Deschamps et al., 2008) and in particular noise tomography (e.g., Lin et al., 2008) (see *Seismic, Ambient Noise Correlation*) facilitate extending the period range to shorter, crustal periods. Overtones can be used to constrain the deeper mantle, down to the 660-km phase transition (e.g., Trampert and van Heijst, 2002; Lebedev and van der Hilst, 2008). Lastly, the long-period surface wave equivalent of *free oscillations of the Earth* can provide constraints on the deep Earth, including inner core anisotropy (e.g., Woodhouse et al., 1986; Tromp, 2001).

Surface wave studies of anisotropy have fairly good depth sensitivity in that they are able to locate the origin of anisotropic signals in the upper mantle to within ~100 km depth. However, particularly compared to body wave measurements such as SKS splitting, the lateral resolution of surface waves is limited, for isotropic structure to perhaps ~500 and ~50 km for global and regional models, respectively, at present. Reasons for discrepancies between published tomographic models include the

different treatment of crustal corrections and phase-velocity measurements, as well as theoretical assumptions about wave propagation. Perhaps more important factors are the globally uneven ray path coverage or regularization choices.

A complete, 3D model of general seismic anisotropy would allow for more powerful petrological and geodynamic inferences than limited studies that focus only on a few aspects of anisotropy or wave propagation. Given the wide array of seismological observations, a desirable procedure to constrain the full 3D dependence of anisotropy is to compare different measures of anisotropy (e.g., Montagner et al., 2000; Wüstefeld et al., 2009) or to perform a joint inversion (e.g., Montagner and Nataf, 1988; Šíleny and Plomerová, 1996; Marone and Romanowicz, 2007). Sensitivity kernels that account for finite-frequency wave-propagation effects and the resulting complex 3D structure of a measurement's sensitivity to Earth structure (e.g., Chevrot, 2006; Long et al., 2008; Sieminski et al., 2009) can facilitate the relative weighting of different observations. Likewise, the incorporation of petrological constraints (e.g., Montagner and Anderson, 1989; Becker et al., 2006) can be used to simplify inversions further (Panning and Nolet, 2008; Chevrot and Monteiller, 2009).

Origin of anisotropy

The SPO type of anisotropy may be caused by any consistent alignment of entities with different isotropic wave speeds. Examples include lower crustal lamellae structures, cracks, or melt inclusions (e.g., Mainprice and Nicholas, 1989; Weiss et al., 1999; Meissner et al., 2006). Crack alignment will be only important for the shallowest crust where it may be indicative of crustal stress (e.g., Crampin and Chastin, 2003). Alignment of partial melt pockets may play a role both for shallow, extensional lithospheric domains, such as underneath mid-oceanic spreading centers or intracontinental rifts (e.g., Holtzman et al., 2003), and at the base of the mantle in the *Mantle D″ Layer* (e.g., Moore et al., 2004).

In between, the LPO type of anisotropy caused by the alignment of intrinsically anisotropic crystals is the most likely cause of anisotropy. The fundamental symmetry classes of material anisotropy of the constituent minerals (e.g., Nye, 1985; Anderson, 1989) determine the overall type of anisotropy in the Earth, and wave propagation depends critically on the type of anisotropy (e.g., Levin and Park, 1998; Schulte-Pelkum and Blackman, 2003). Several crustal rocks show LPO anisotropy; of particular interest are those rich in phyllosilicates (micas) in the upper-middle crust, and amphibole minerals in lower crust (e.g., Christensen and Mooney, 1995). In the upper mantle, the highly anisotropic olivine makes up ~60% of rocks (e.g., Mainprice, 2007). Laboratory experiments show that if multi-crystal olivine assemblages are deformed in the dislocation creep regime, crystals typically align such that the resulting fast propagation orientation rotates into the direction of shear, and many mantle xenoliths show corresponding LPO patterns (Mainprice, 2007; Karato et al., 2008).

This connection between rock deformation and seismic anisotropy allows an association of the patterns of azimuthal mantle anisotropy (e.g., Figure 2a) with *mantle convection* (e.g., McKenzie, 1979; Tanimoto and Anderson, 1984). A coarse approximation uses tectonic plate motion to imply deep flow direction, or, more realistically, flow can be calculated from global circulation models (Hager and O'Connell, 1981). The general association between mantle flow and anisotropy in terms of radial anisotropy is that flow in the upper boundary layer aligns olivine such that $v_{SH} > v_{SV}$ underneath oceanic plates due to a simple shear type of deformation (Figure 1). In regions of dominantly radial mass transport such as subduction zones and underneath spreading centers, $v_{SV} > v_{SH}$ (Chastel et al., 1993; Montagner, 2007). The radial and azimuthal anisotropy patterns shown in Figure 2a are broadly consistent with this expectation (Figure 2b), though there are also clear differences which are easier to constrain in regional studies (e.g., Gaherty et al., 1996). Complexities include variations azimuthal anisotropy orientations and amplitudes (e.g., Ekström and Dziewonski, 1998; Smith et al., 2004), and many of those patterns are accessible to geodynamic modeling, discussed below.

Given the importance of the details of the connection between seismology and geodynamics, several theoretical descriptions exist that predict microstructural LPO development given general deformation histories, as constrained by laboratory experiments (e.g., Kaminski and Ribe, 2001; Blackman, 2007). However, further laboratory constraints, for example on the reorientation of existing LPO fabrics under changing deformation regimes, are required to decide on the most appropriate treatment. Complex deformation histories are expected to lead to complex anisotropy. Yet, under monotonous deformation (e.g., by simple shear), olivine LPO is expected to saturate over finite strains of ~10. Amplitude variations compared to a single crystal may therefore be mainly due to orientation of the symmetry axis of the effective elastic tensor for an aggregate of crystals (cf. Karato et al., 2008).

Laboratory work over the last 10 years has further shown that the role of water content, deviatoric stress levels, and pressure can lead to significantly different LPO development from the typical, dry A-type fabrics that show the "fast axes along flow" alignment discussed above. For example, the high stress, high water content B-type fabric aligns the fast axes of olivine orthogonal to the direction of shear. Variations in water content have been used to explain some of the variability that is apparent in asthenospheric depth anisotropy, such as the decrease in azimuthal anisotropy strength across the Pacific from young to older seafloor, or the variability of orientations of SKS splitting in subduction zones (Mainprice, 2007; Karato et al., 2008).

LPO development under deformation of mantle rocks not only affects seismic properties, but also leads to thermal and mechanical anisotropy. The feedback of these effects into mantle convection and lithospheric deformation are potentially profound (e.g., Christensen, 1987; Chastel et al., 1993; Lev and Hager, 2008; Tommasi et al., 2009) and are currently an active area of research.

Observations of anisotropy and dynamic inferences

Whole earth anisotropy

Seismic anisotropy is found throughout the Earth, with the exception of the fluid outer core, though it is concentrated in certain depth regions (Figure 1). In the mantle, the best-constrained and strongest signal is found in the uppermost ~300 km where SH velocities are faster than SV by up to ~4% on average, as indicated by the Love–Rayleigh discrepancy. The exact shape of the average radial anisotropy profile is less certain, though most recent models agree that the largest anomalies are not found at the surface, but rather at ~100 km depth (Figure 2b). This peak may be associated with asthenospheric shear flow which is expected to lead to the largest strain-rates underneath the oceanic lithosphere, which is up to ~100 km thick when defined thermally (see *Mantle Convection*). Given that mantle anisotropy is most likely caused by LPO of olivine, the peak in seismic anisotropy in the uppermost mantle has been associated with the relatively high stress and low temperature depth region where dislocation dominates over diffusion creep (Karato, 1992; Gaherty and Jordan, 1995) (see *Mantle Viscosity*). Using composite rheologies, geodynamic models can be used to estimate the transition depths for the different creep laws, so delineating the region where LPO forms explicitly (e.g., McNamara et al., 2002; Podolefsky et al., 2004; Becker et al., 2008). Once rocks transition into the diffusion-creep dominated deformation regime, LPO is typically assumed to be destroyed quickly at high temperatures, or left preserved (frozen in) at low temperatures/small velocity gradients. The decrease in radial anisotropy toward the surface (Figure 2b) may therefore be associated with tectonically older, frozen in structure. On the scales accessible by surface wave studies, for example, anisotropy in old lithospheric domains may be less well aligned into the vertical, or into a coherent horizontal orientation, than in the asthenosphere, which is shaped by current mantle convection (e.g., Fouch and Rondenay, 2006).

At larger mantle depths, radial anisotropy becomes less well constrained (e.g., Visser et al., 2008). There is some indication that radial anomalies pick up around the transition zone (Figure 1), and several studies have argued for the existence of azimuthal anisotropy around 660 km (e.g., Trampert and van Heijst, 2002; Wookey et al., 2002). Most of the lower mantle is nearly isotropic until the D″ region close to the core mantle boundary where there is good evidence for the existence of anisotropy from regional studies (e.g., Moore et al., 2004), and indications for average radial anisotropy from global studies (Boschi and Dziewoński, 2000; Panning and Romanowicz, 2006). As for the upper mantle, one may invoke an LPO reactivation of dislocation creep, for example in cold, highly deformed subduction slabs (see Figure 1; McNamara et al., 2002). The other, at present perhaps equally likely, mechanism that has been invoked for D″ anisotropy is the alignment of melt tubules (SPO). Melt alignment may also play a role in the transition zone if the latter represents a melt-rich water filter (Bercovici and Karato, 2003). The D″ region is expected to be at least as dynamically complex as the upper thermal boundary layer, and both domains are affected by compositional anomalies. Those include the continental lithosphere, with its stiff, compositionally anomalous and presumably neutrally buoyant cratonic keels, and likely piles of dense material at the base of the mantle in regions displaced along the CMB from recent subduction (e.g., Garnero, 2004; Garnero and McNamara, 2008). We therefore expect significant lateral variations in the generation of anisotropy within D″ depending on the vertical flow setting (Figure 1, e.g., Moore et al., 2004). Close to the CMB, anisotropy may also vary with depth depending on if lower mantle material has transitioned to the post-perovskite phase (e.g., Wookey et al., 2005; Merkel et al., 2007).

There is also robust evidence for anisotropy within the Earth's core. Body waves that traverse the inner core and are aligned with the rotation axis arrive earlier than those that cross in the equatorial plane (Morelli et al., 1986). Evidence for anisotropy is also seen in the splitting of normal modes (Woodhouse et al., 1986), and more recent data and models for core anisotropy are discussed in Tromp (2001) and Souriau (2007). However, there are still debates on the exact nature of the anisotropy distribution with depth (cf. Ishii and Dziewoński, 2003). Figure 1 shows radial, shear-wave anisotropy for the inner core from Beghein and Trampert (2003). This particular model invoked a hexagonal close-packed phase of iron in the upper half of the inner core, and perhaps a transition into a different iron phase at depth, and predicts large amplitudes of radial anisotropy compared to the upper mantle. The origin of inner core anisotropy is also less clear than for the upper mantle (Mainprice, 2007). One hypothesis that has recently been discussed in some detail is freezing in of convective patterns during the cooling and evolution of the inner core (Jeanloz and Wenk, 1988; Buffett, 2009; Deguen and Cardin, 2009).

Structure and dynamics of the upper boundary layer

Seismic anisotropy at every depth range throughout the Earth holds valuable information on the dynamics of the planet. The connections can be made quantitative most easily for the shallower layers where seismological constraints abound, rock deformation is accessible via laboratory experiments, and geodynamic modeling is fairly well constrained. In the case of crack anisotropy in the shallow

crust, observations yield constraints on regional stress fields. Applications include industry work (vertical seismic profiling in boreholes), earthquake studies around faults, and volcano monitoring where cracking due to magma migration can be traced.

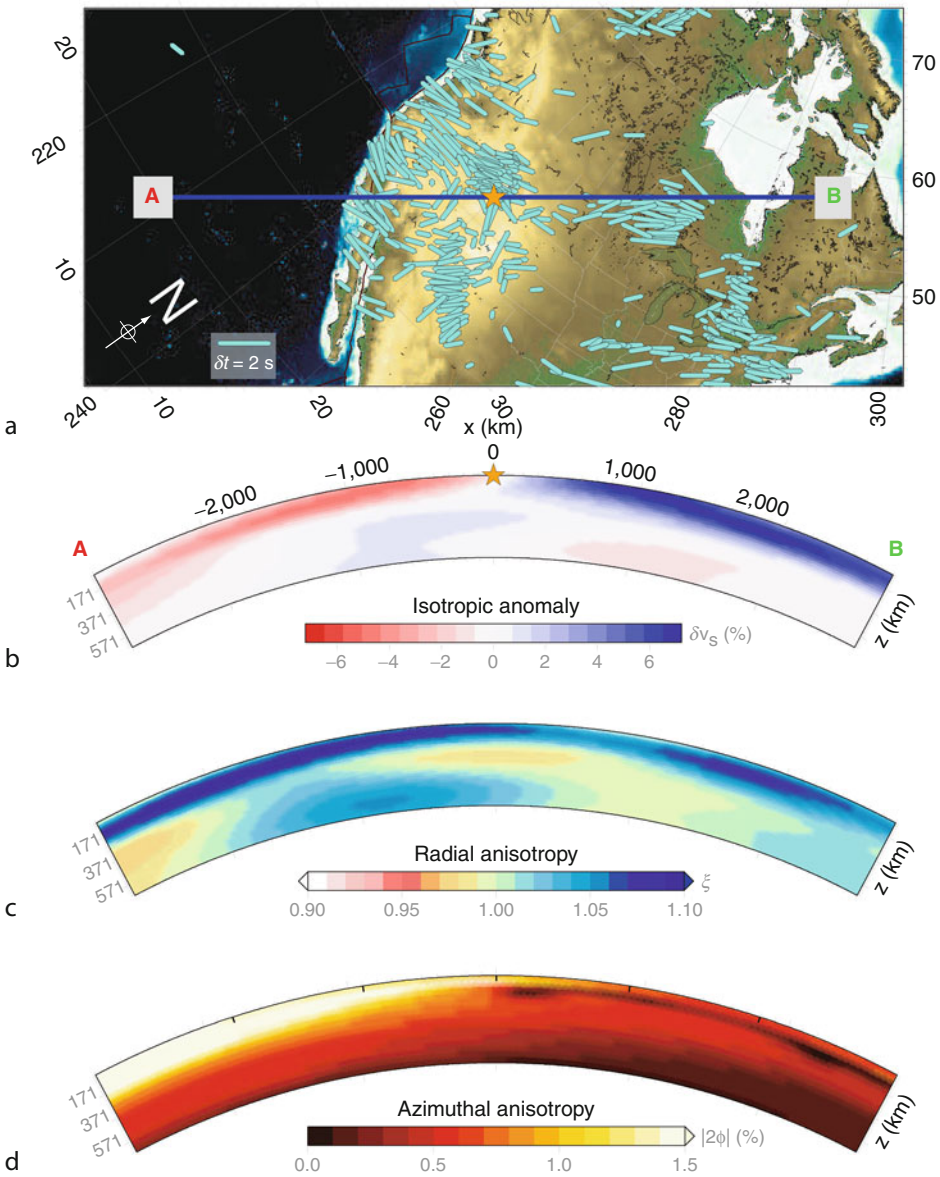
Within the upper convective boundary layer, the oceanic plate domains (see *Lithosphere, Oceanic*) should most closely resemble the simplified view of radial and azimuthal anisotropy due to LPO anisotropy formation in mantle flow as shown in Figure 1. Gaboret et al. (2003), Becker et al. (2003), and Behn et al. (2004) showed that mantle circulation from geodynamic models does indeed provide a valid explanation for azimuthal anisotropy patterns (Figure 2a), and that comparison of model predictions with anisotropy can yield constraints on mantle flow, such as the role of buoyant mantle upwellings as opposed to plate-induced shear. Becker et al. (2008) provided a quantitative model of radial anisotropy, and Figure 2b shows the fit of their preferred model to radial anisotropy averages in the upper mantle, as well as lateral patterns in azimuthal and radial anisotropy. Results are consistent with the expectation that the geodynamic models should describe recent (few 10 s of Myr) asthenospheric flow best. The correlations between geodynamics and the seismological models (Figure 2b) are comparable or better than the match between different seismological models. Such first-order agreement between global geodynamics and seismology motivates current modeling efforts, for example on constraining the amount of net rotations of the lithosphere or the degree of lateral viscosity variations (e.g., Becker, 2008; Conrad et al., 2007; Conrad and Behn, 2010; Kreemer, 2009).

Figure 2b shows that geodynamic models typically underpredict radial anisotropy in the shallower parts of the lithosphere, which is mainly due to continental domains. While anisotropy in younger continental lithosphere such as in the western USA appears to be well described by asthenospheric flow, older regions show more complex behavior such as a consistent orientation of seismic anisotropy over several hundred kilometers (e.g., Babuška and Plomerová, 2006). It has been suggested that anisotropy is concentrated in, and frozen into, the continental lithosphere, or, alternatively, that radial anisotropy is largest right underneath the mechanical boundary layer formed by stiff continents (e.g., Gaherty and Jordan, 1995; Gung et al., 2003; Fouch and Rondenay, 2006). Figure 3 shows a profile through North America; anisotropy, as inferred from these models, only partially conforms to the simplified expectations (cf. Panning and Romanowicz, 2006). The cross section of radial anisotropy shows the expected focusing of SH faster than SV in the Pacific plate, and some regionally enhanced $v_{SH} > v_{SV}$ within the eastern USA and the Canadian craton, but no enhanced anisotropy beneath what would be inferred to be the base of the continental lithosphere from the isotropic anomalies. Azimuthal anisotropy is also, expectedly, strong within the Pacific plate (compare Figure 2a), but there is an intriguing low-azimuthal anisotropy channel within the eastern North American continental lithosphere. If such features are due to complex tectonic deformation with small lateral shear-coherence, or due to the averaging properties of surface waves and incomplete ray illumination, remains to be determined. The study of continental anisotropy is an active area of research, and many questions such as to the vertical coherence of lithospheric deformation and the depth extent of fault zone localization will benefit from the information that seismic anisotropy can bring to the table. There are numerous other, regional tectonic settings where anisotropy can yield important constraints, and those cannot be comprehensively reviewed here. Important examples include continental transforms and collision zones, spreading centers, and subduction zones. Reviews of our current understanding of such settings can be found in Silver (1996); Savage (1999); Park and Levin (2002) and Long and Silver (2009).

Powerful dynamic insights notwithstanding, there are still large uncertainties in every step of the chain of modeling that has to be followed. Complexities arise from inferring mantle flow from geodynamics (e.g., role of chemical vs. thermal buoyancy, uncertainties about rheology), to predicting LPO textures (e.g., proper microphysical treatment), to inferring elasticity tensors (e.g., homogenization and temperature/pressure derivatives), to mapping those tensors in 3D to whatever seismological observable (preferred) or seismological model (more common) are used to benchmark the models (e.g., finite frequency wave propagation, sampling). The finding that overall patterns appear to be well explained (Figure 2), and that synthetic LPOs match those of xenolith samples provide some *a posterior* justification for the modeling rationale. Moreover, these agreements indicate that the bulk of the asthenospheric flow is indeed dominated by dry A-type fabrics. However, future refinements of seismological imaging, for example through array deployments such as EarthScope USArray and temporary seafloor studies, theoretical developments in seismology, and the improved geodynamic treatment of anisotropy will undoubtedly lead to adjustment of our understanding of whole Earth anisotropic structure.

Summary

Seismic anisotropy is ubiquitous throughout the Earth and provides constraints on dynamic processes, from the stress in the crust, the origin and evolution of the continental lithosphere, through convective flow in the upper mantle, to core evolution. The state of upper-mantle geodynamic modeling is such that important questions, such as about absolute plate motion reference frames, intraplate deformation, or the hydration state of the mantle can be addressed. Important issues about the resolution of different seismological datasets and degree of robustness of seismological images remain. Joint with the inherent uncertainties in geodynamic modeling and how to map flow into seismic anisotropy, this means that numerous



Seismic Anisotropy, Figure 3 Pacific and North American upper mantle anisotropy. (**a**) SKS splitting (as in Figure 2, but averaged by 0.5°) and location of cross-continental profile; (**b**) isotropic shear wave velocity relative to background (Voigt average, from Kustowski et al., 2008); (**c**) radial anisotropy ($\xi = (v_{SH}/v_{SV})^2$, from Kustowski et al., 2008); and (**d**) strength of azimuthal anisotropy ($|(v_{SV1} - v_{SV2})/v_{SV}|$, from Lebedev and van der Hilst, 2008). (Modified from Long and Becker, 2010.)

questions for the interpretation of anisotropy observable are open. This challenge mandates further theoretical and instrumentational efforts and that the study of anisotropy proceeds interdisciplinary and in a dynamics context. Answering those questions holds the promise of arriving at a new understanding of the workings of the mantle system.

Acknowledgments

Detailed comments by Donna Blackman, Mark Behn, and Sergei Lebedev and valuable suggestions from Lapo Boschi, Sebastien Chevrot, David Okaya, Mark Panning, Vera Schulte-Pelkum, and an anonymous reviewer helped improve this contribution.

Bibliography

Abt, D. L., and Fischer, K. M., 2008. Resolving three-dimensional anisotropic structure with shear-wave splitting tomography. *Geophysical Journal International*, **173**, 859–886.

Anderson, D. L., 1966. Recent evidence concerning the structure and composition of the Earth's mantle. In *Physics and Chemistry of the Earth*. Oxford: Pergmanon, Vol. 6, pp. 1–131.

Anderson, D. L., 1989. *Theory of the Earth*. Boston, MA: Blackwell Scientific Publications (Available online at http://caltechbook.library.caltech.edu/14/1/TheoryoftheEarth.pdf, accessed 01/2010).
Ando, M., Ishikawa, Y., and Yamasaki, F., 1983. Shear-wave polarization anisotropy in the mantle beneath Honshu, Japan. *Journal of Geophysical Research*, **88**, 5850–5864.
Babuška, V., and Cara, M., 1991. *Seismic Anisotropy in the Earth*. Dordrecht: Kluwer Academic Publishers.
Babuška, V., and Plomerová, J., 2006. European mantle lithosphere assembled from rigid microplates with inherited seismic anisotropy. *Physics of the Earth and Planetary Interiors*, **158**, 264–280.
Babuška, V., Plomerová, J., and Šíleny, J., 1984. Spatial variations of Presiduals and deep structure of the European lithospere. *Geophysical Journal of the Royal Astronomical Society*, **79**, 363–383.
Backus, G. E., 1965. Possible forms of seismic anisotropy of the uppermost mantle under oceans. *Journal of Geophysical Research*, **70**, 3429–3439.
Becker, T. W., 2008. Azimuthal seismic anisotropy constrains net rotation of the lithosphere. *Geophysical Research Letters*, **35**, L05303, doi:10.1029/2007GL032928, correction: doi:10.1029/2008GL033946.
Becker, T. W., Kellogg, J. B., Ekström, G., and O'Connell, R. J., 2003. Comparison of azimuthal seismic anisotropy from surface waves and finite-strain from global mantle-circulation models. *Geophysical Journal International*, **155**, 696–714.
Becker, T. W., Chevrot, S., Schulte-Pelkum, V., and Blackman, D. K., 2006. Statistical properties of seismic anisotropy predicted by upper mantle geodynamic models. *Journal of Geophysical Research*, **111**, B08309, doi:10.1029/2005JB004095.
Becker, T. W., Kustowski, B., and Ekström, G., 2008. Radial seismic anisotropy as a constraint for upper mantle rheology. *Earth and Planetary Science Letters*, **267**, 213–237.
Beghein, C., and Trampert, J., 2003. Robust normal mode constraints on inner-core anisotropy from model space search. *Science*, **299**, 552–555.
Behn, M. D., Conrad, C. P., and Silver, P. G., 2004. Detection of upper mantle flow associated with the African Superplume. *Earth and Planetary Science Letters*, **224**, 259–274.
Bercovici, D., and Karato, S-i., 2003. Whole-mantle convection and the transition-zone water filter. *Nature*, **425**, 39–44.
Blackman, D., 2007. Use of mineral physics, with geodynamic modelling and seismology, to investigate flow in the Earth's mantle. *Reports on Progress in Physics*, **70**, 659–689.
Bokelmann, G. H. R., 2002. Convection-driven motion of the north American craton: evidence from *P*-wave anisotropy. *Geophysical Journal International*, **148**, 278–287.
Boschi, L., and Dziewoński, A. M., 2000. Whole Earth tomography from delay times of P, PcP, PKP phases: lateral heterogeneities in the outer core, or radial anisotropy in the mantle? *Journal of Geophysical Research*, **105**, 25,567–25,594.
Browaeys, J., and Chevrot, S., 2004. Decomposition of the elastic tensor and geophysical applications. *Geophysical Journal International*, **159**, 667–678.
Buffett, B., 2009. Onset and orientation of convection in the inner core. *Geophysical Journal International*, **179**, 711–719.
Chastel, Y. B., Dawson, P. R., Wenk, H.-R., and Bennett, K., 1993. Anisotropic convection with implications for the upper mantle. *Journal of Geophysical Research*, **98**, 17,757–17,771.
Chevrot, S., 2006. Finite-frequency vectorial tomography: a new method for high-resolution imaging of upper mantle anisotropy. *Geophysical Journal International*, **165**, 641–657.
Chevrot, S., and Monteiller, V., 2009. Principles of vectorial tomography – the effects of model parametrization and regularization in tomographic imaging of seismic anisotropy. *Geophysical Journal International*, **179**, 1726–1736.
Christensen, N. I., and Mooney, W. D., 1995. Seismic velocity structure and composition of the continental crust: a global review. *Journal of Geophysical Research*, **100**, 9761–9788.
Christensen, U. R., 1987. Some geodynamical effects of anisotropic viscosity. *Geophysical Journal of the Royal Astronomical Society*, **91**, 711–736.
Conrad, C. P., and Behn, M., 2010. Constraints on lithosphere net rotation and asthenospheric viscosity from global mantle flow models and seismic anisotropy. *Geochemistry, Geophysics, Geosystems*, **11**, Q05W05, doi:10.1029/2009GC002970.
Conrad, C. P., Behn, M. D., and Silver, P. G., 2007. Global mantle flow and the development of seismic anisotropy: differences between the oceanic and continental upper mantle. *Journal of Geophysical Research*, **112**, B07317, doi:10.1029/2006JB004608.
Crampin, S., and Chastin, S., 2003. A review of shear wave splitting in the crack-critical crust. *Geophysical Journal International*, **155**, 221–240.
Deguen, R., and Cardin, P., 2009. Tectonic history of the Earth's inner core preserved in its seismic structure. *Nature Geoscience*, **2**, 419–422.
Deschamps, F., Lebedev, S., Meier, T., and Trampert, J., 2008. Azimuthal anisotropy of Rayleigh-wave phase velocities in the east-central United States. *Geophysical Journal International*, **173**, 827–843.
Dziewoński, A. M., and Anderson, D. L., 1981. Preliminary reference Earth model. *Physics of the Earth and Planetary Interiors*, **25**, 297–356.
Ekström, G., and Dziewonski, A. M., 1998. The unique anisotropy of the Pacific upper mantle. *Nature*, **394**, 168–172.
Forsyth, D. W., 1975. The early structural evolution and anisotropy of the oceanic upper mantle. *Geophysical Journal of the Royal Astronomical Society*, **43**, 103–162.
Fouch, M., 2006. Upper mantle anisotropy database. Online, accessed in 06/2006, http://geophysics.asu.edu/anisotropy/upper/.
Fouch, M. J., and Rondenay, S., 2006. Seismic anisotropy beneath stable continental interiors. *Physics of the Earth and Planetary Interiors*, **158**, 292–320.
Gaboret, C., Forte, A. M., and Montagner, J.-P., 2003. The unique dynamics of the Pacific Hemisphere mantle and its signature on seismic anisotropy. *Earth and Planetary Science Letters*, **208**, 219–233.
Gaherty, J. B., and Jordan, T. H., 1995. Lehmann discontinuity as the base of an anisotropic layer beneath continents. *Science*, **268**, 1468–1471.
Gaherty, J. B., Jordan, T. H., and Gee, L. S., 1996. Seismic structure of the upper mantle in a central Pacific corridor. *Journal of Geophysical Research*, **101**, 22,291–22,310.
Garnero, E. J., 2004. A new paradigm for Earth's core-mantle boundary. *Science*, **304**, 835–836.
Garnero, E. J., and McNamara, A. K., 2008. Structure and dynamics of the Earth's lower mantle. *Science*, **320**, 626–628.
Gung, Y., Panning, M., and Romanowicz, B., 2003. Global anisotropy and the thickness of continents. *Nature*, **422**, 707–711.
Hager, B. H., and O'Connell, R. J., 1981. A simple global model of plate dynamics and mantle convection. *Journal of Geophysical Research*, **86**, 4843–4867.
Hearmon, R. F. S., 1961. *An Introduction to Applied Anisotropic Elasticity*. London: Oxford University Press.
Hess, H. H., 1964. Seismic anisotropy of the uppermost mantle under oceans. *Nature*, **203**, 629–631.
Holtzman, B. K., Kohlstedt, D. L., Zimmerman, M. E., Heidelbach, F., Hiraga, T., and Hustoft, J., 2003. Melt segregation and strain partitioning: implications for seismic anisotropy and mantle flow. *Science*, **301**, 1227–1230.

Ishii, M., and Dziewoński, A. M., 2003. Distinct seismic anisotropy at the centre of the earth. *Physics of the Earth and Planetary Interiors*, **140**, 203–217.

Jeanloz, R., and Wenk, H. R., 1988. Convection and anisotropy of the inner core. *Geophysical Research Letters*, **15**, 72–75.

Kaminski, É., and Ribe, N. M., 2001. A kinematic model for recrystallization and texture development in olivine polycrystals. *Earth and Planetary Science Letters*, **189**, 253–267.

Karato, S-i., 1992. On the Lehmann discontinuity. *Geophysical Research Letters*, **51**, 2255–2258.

Karato, S-i., Jung, H., Katayama, I., and Skemer, P., 2008. Geodynamic significance of seismic anisotropy of the upper mantle: new insights from laboratory studies. *Annual Review of Earth Planetary Sciences*, **36**, 59–95.

Kreemer, C., 2009. Absolute plate motions constrained by shear wave splitting orientations with implications for hot spot motions and mantle flow. *Journal of Geophysical Research,* **114**, B10405, doi:10.1029/2009JB006416.

Kustowski, B., Ekström, G., and Dziewoński, A. M., 2008. Anisotropic shear-wave velocity structure of the Earth's mantle: a global model. *Journal of Geophysical Research,* **113**, B06306, doi:10.1029/2007JB005169.

Lebedev, S., and van der Hilst, R. D., 2008. Global upper-mantle tomography with the automated multimode inversion of surface and S-wave forms. *Geophysical Journal International*, **173**, 505–518.

Lev, E., and Hager, B. H., 2008. Rayleigh Taylor instabilities with anisotropic lithospheric viscosity. *Geophysical Journal International*, **173**, 806–814.

Levin, V., and Park, J., 1998. *P – SH* conversions in layered media with hexagonally symmetric anisotropy: a cookbook. *Pure and Applied Geophysics*, **151**, 669–697.

Lin, F., Moschetti, M. P., and Ritzwoller, M. H., 2008. Surface wave tomography of the western United States from ambient seismic noise: Rayleigh and Love wave phase velocity maps. *Geophysical Journal International*, **173**, 281–298.

Long, M. D., and Becker, T. W., 2010. Mantle dynamics and seismic anisotropy. *Earth and Planetary Science Letters,* **297**, 341–354.

Long, M. D., and Silver, P. G., 2009. Shear wave splitting and mantle anisotropy: measurements, interpretations, and new directions. *Surveys in Geophysics*, **30**, 407–461.

Long, M. D., de Hoop, M. V., and van der Hilst, R. D., 2008. Wave equation shear wave splitting tomography. *Geophysical Journal International*, **172**, 311–330.

Love, A. E. H., 1927. *A Treatise on the Mathematical Theory of Elasticity.* Cambridge: Cambridge University Press. reprinted in 1944 by Dover Publications, New York.

Mainprice, D., 2007. Seismic anisotropy of the deep Earth from a mineral and rock physics perspective. In Schubert, G., and Bercovic, D. (eds.), *Treatise on Geophysics*. Oxford: Elsevier, Vol. 2, pp. 437–492.

Mainprice, D., and Nicholas, A., 1989. Development of shape and lattice preferred orientations: application to the seismic anisotropy of the lower crust. *Journal of Structural Geology*, **11**, 175–189.

Marone, F., and Romanowicz, F., 2007. The depth distribution of azimuthal anisotropy in the continental upper mantle. *Nature*, **447**, 198–201.

McKenzie, D. P., 1979. Finite deformation during fluid flow. *Geophysical Journal of the Royal AstronomicalSociety*, **58**, 689–715.

McNamara, A. K., van Keken, P. E., and Karato, S-i., 2002. Development of anisotropic structure in the Earth's lower mantle by solid-state convection. *Nature*, **416**, 310–314.

Meissner, R., Mooney, W. D., and Artemieva, I., 2002. Seismic anisotropy and mantle creep in young orogens. *Geophysical Journal International*, **149**, 1–14.

Meissner, R., Rabbel, W., and Kern, H., 2006. Seismic lamination and anisotropy of the lower continental crust. *Tectonophysics*, **416**, 81–99.

Merkel, S., McNamara, A. K., Kubo, A., Speziale, S., Miyagi, L., Meng, Y., Duffy, T. S., and Wenk, H.-R., 2007. Deformation of (Mg, Fe)SiO3 post-perovskite and D″ anisotropy. *Science*, **316**(5832), 1729–32.

Montagner, J.-P., 2007. Upper mantle structure: Global isotropic and anisotropic elastic tomography. In Schubert, G., and Bercovici, D. (eds.), *Treatise on Geophysics*. Amsterdam: Elsevier, Vol. 1, pp. 559–589.

Montagner, J. P., and Anderson, D. L., 1989. Petrological constraints on seismic anisotropy. *Physics of the Earth and Planetary Interiors*, **54**, 82–105.

Montagner, J.-P., and Nataf, H.-C., 1986. A simple method for inverting the azimuthal anisotropy of surface waves. *Journal of Geophysical Research*, **91**, 511–520.

Montagner, J. P., and Nataf, H. C., 1988. Vectorial tomography-I. Theory. *Geophysical Journal*, **94**, 295–307.

Montagner, J.-P., Griot-Pommera, D.-A., and Laveé, J., 2000. How to relate body wave and surface wave anisotropy? *Journal of Geophysical Research*, **105**, 19,015–19,027.

Moore, M. M., Garnero, E. J., Lay, T., and Williams, Q., 2004. Shear wave splitting and waveform complexity for lowermost mantle structures with low-velocity lamellae and transverse isotropy. *Journal of Geophysical Research*, **103**, B02319, doi:10.1029/2003JB002546.

Morelli, A., Dziewonski, A. M., and Woodhouse, J. H., 1986. Anisotropy of the inner core inferred from PKIKP travel times. *Geophysical Research Letters*, **13**, 1545–1548.

Nataf, H.-C., Nakanishi, I., and Anderson, D. L., 1984. Anisotropy and shear velocity heterogeneity in the upper mantle. *Geophysical Research Letters*, **11**, 109–112.

Nye, J. F., 1985. *Physical Properties of Crystals*. London: Oxford University Press.

Panning, M., and Romanowicz, B., 2006. A three-dimensional radially anisotropic model of shear velocity in the whole mantle. *Geophysical Journal International*, **167**, 361–379.

Panning, M. P., and Nolet, G., 2008. Surface wave tomography for azimuthal anisotropy in a strongly reduced parameter space. *Geophysical Journal International*, **174**, 629–648.

Park, J., and Levin, V., 2002. Seismic anisotropy: tracing plate dynamics in the mantle. *Science*, **296**, 485–489.

Podolefsky, N. S., Zhong, S., and McNamara, A. K., 2004. The anisotropic and rheological structure of the oceanic upper mantle from a simple model of plate shear. *Geophysical Journal International*, **158**, 287–296.

Saltzer, R. L., Gaherty, J. B., and Jordan, T. H., 2000. How are vertical shear wave splitting measurements affected by variations in the orientation of azimuthal anisotropy with depth? *Geophysical Journal International*, **141**, 374–390.

Savage, M., 1998. Lower crustal anisotropy or dipping boundaries? Effects on receiver functions and a case study in New Zealand. *Journal of Geophysical Research*, **103**, 15,069–15,087.

Savage, M. K., 1999. Seismic anisotropy and mantle deformation: what have we learned from shear wave splitting? *Reviews of Geophysics*, **37**, 65–106.

Schulte-Pelkum, V., and Blackman, D. K., 2003. A synthesis of seismic *P* and *S* anisotropy. *Geophysical Journal International*, **154**, 166–178.

Schulte-Pelkum, V., Masters, G., and Shearer, P. M., 2001. Upper mantle anistropy from long-period *P* polarization. *Journal of Geophysical Research*, **106**, 21,917–21,934.

Sieminski, A., Trampert, J., and Tromp, J., 2009. Principal component analysis of anisotropic finite-frequency sensitivity kernels. *Geophysical Journal International*, **179**, 1186–1198.

Šíleny, J., and Plomerová, J., 1996. Inversion of shear-wave splitting parameters to retrieve three-dimensional orientation of anisotropy in continental lithosphere. *Physics of the Earth and Planetary Interiors*, **95**, 277–292.
Silver, P. G., 1996. Seismic anisotropy beneath the continents: probing the depths of geology. *Annual Review of Earth Planetary Sciences*, **24**, 385–432.
Silver, P. G., and Chan, W. W., 1991. Shear wave splitting and subcontinental mantle deformation. *Journal of Geophysical Research*, **96**, 16,429–16,454.
Smith, D. B., Ritzwoller, M. H., and Shapiro, N. M., 2004. Stratification of anisotropy in the Pacific upper mantle. *Journal of Geophysical Research*, **109**, B11309, doi:10.1029/2004JB003200.
Smith, G. P., and Ekström, G., 1999. A global study of P_n anisotropy beneath continents. *Journal of Geophysical Research*, **104**, 963–980.
Smith, M. L., and Dahlen, F. A., 1973. The azimuthal dependence of Love and Rayleigh wave propagation in a slightly anisotropic medium. *Journal of Geophysical Research*, **78**, 3321–3333.
Souriau, A., 2007. The Earth's core. In Romanowicz, B., and Dziewonski, A. (eds.), *Treatise on Geophysics*. Amsterdam: Elsevier, Vol. 1, pp. 655–693.
Tanimoto, T., and Anderson, D. L., 1984. Mapping convection in the mantle. *Geophysical Research Letters*, **11**, 287–290.
Tanimoto, T., and Anderson, D. L., 1985. Lateral heterogeneity and azimuthal anisotropy of the upper mantle: Love and Rayleigh waves 100–250 s. *Journal of Geophysical Research*, **90**, 1842–1858.
Thomsen, L., 1986. Weak elastic anisotropy. *Geophysics*, **51**, 1954–1966.
Tommasi, A., Knoll, M., Vauchez, A., Sgnorelli, J., Thoraval, C., and Loge, R., 2009. Structural reactivation in plate tectonics controlled by olivine crystals anisotropy. *Nature Geosciences*, **2**, 423–427.
Trampert, J., and van Heijst, H. J., 2002. Global azimuthal anisotropy in the transition zone. *Science*, **296**, 1297–1299.
Tromp, J., 2001. Inner core anisotropy and rotation. *Annual Review of Earth and Planetary Sciences*, **29**, 47–69.
Vinnik, L., Kosarev, G. L., and Makeyeva, L. I., 1984. Anisotropy of the lithosphere from the observations of *SKS* and *SKKS* phases. *Proceedings of Academic Sciences USSR*, **278**, 1335–1339.
Visser, K., Trampert, J., Lebedev, S., and Kennett, B. L. N., 2008. Probability of radial anisotropy in the deep mantle. *Earth and Planetary Science Letters*, **270**, 241–250.
Weiss, T., Siegesmund, S., Rabbel, W., Bohlen, T., and Pohl, M., 1999. Seismic velocities and anisotropy of the lower continental crust: a review. *Pure and Applied Geophysics*, **156**, 97–122.
Woodhouse, J. H., Giardini, D., and Li, X.-D., 1986. Evidence for inner core anisotropy from free oscillations. *Geophysical Research Letters*, **13**, 1549–1552.
Wookey, J., Kendall, J.-M., and Barruol, G., 2002. Mid-mantle deformation inferred from seismic anisotropy. *Nature*, **415**, 777–780.
Wookey, J., Stackhouse, S., Kendall, J.-M., Brodholt, J., and Price, G. D., 2005. Efficacy of the post-perovskite phase as an explanation for lowermost-mantle seismic properties. *Nature*, **438**, 1004–1007.
Wüstefeld, A., Bokelmann, G. H. R., Barruol, G., and Montagner, J. P., 2009. Identifying global seismic anisotropy patterns by correlating shear-wave splitting and surface-wave data. *Physics of the Earth and Planetary Interiors*, **176**, 198–212.

Cross-references

Body Waves
Earth's Structure, Global
Free Oscillations of the Earth
Inverse Theory, Global Optimization
Inverse Theory, Linear
Inverse Theory, Monte Carlo Method
Inverse Theory, Singular Value Decomposition
Lithosphere, Continental
Lithosphere, Oceanic
Mantle Convection
Mantle D″ Layer
Mantle Viscosity
Propagation of Elastic Waves: Fundamentals
Seismic Tomography
Seismic, Ambient Noise Correlation
Seismic, Receiver Function Technique
Seismology, Global Earthquake Model
Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting
Surface Waves

SEISMIC DATA ACQUISITION AND PROCESSING

Kabir Roy Chowdhury
Department of Earth Sciences, Utrecht University, Utrecht, The Netherlands

Definition

Seismic data acquisition

Generation of (artificial) seismic signals on land (on surface, or, buried) or in water, reception of the signals after they travel through the interior of the earth, and their (digital) recording for later analysis.

Seismic data processing

Analysis of recorded seismic signals to filter (reduce/eliminate) unwanted components (noise) and create an image of the subsurface to enable geological interpretation, and eventually to obtain an estimate of the distribution of material properties in the subsurface (inversion).

Introduction

Reflection seismics is akin to the "echo-in-the-well" experiment, it involves calculating the depth of the geological boundary from the two-way travel-time (TWT) of the seismic signal and its speed.

Seismic data acquisition and processing aims mainly to obtain an image of the sedimentary basins in interior of the earth, using waves generated by "artificial" earthquakes. These images can then be used to identify locations favorable for accumulation of hydrocarbons (oil and gas), which may then be drilled to determine the ground truth – and eventually to exploit the resources. Since the first known reflection seismic experiment in 1921 near Oklahoma City, USA (Figure 1.3, Sheriff and Geldart 1995), reflection seismics has established itself as the most accurate technique to image the sedimentary basins for the exploration of hydrocarbons.

The phrase "seismic" instead of "seismological" in the following stresses the "man-made" nature of the waves used. Both seismics and seismology use the basic theory

Seismic Data Acquisition and Processing, Table 1 Imaging the earth using natural/artificial earthquakes

Keyword	Seismics	Seismology
Wave source	Explosions, vibrations	Natural earthquakes
Energy penetration	Shallow	Deep
Max imaging depth	Base of crust	Whole earth
Location of source	Precisely known	Estimated post facto
Time of occurrence	Precisely known	Estimated post facto
Energy involved	Small-medium	Can be huge
Wave-propagation	Mostly vertical	Mostly horizontal
Frequencies mostly excited/used	1–100 Hz	0.01–1 Hz
Receivers	Geophones	Seismometers
Wave-field sampling	Dense	Sparse (getting better)
Data volume	Terabytes	Gigabytes
Accuracy	Large	Small–medium
Main application	Oil and gas	Earth-structure
Other applications	Civil engg., crustal	Civil engg.
Investment	$$$$	$$

of wave-propagation through the earth, for which Aki and Richards (2002) is a good resource. Table 1 summarizes the important differences between the two approaches though; let us briefly look at two.

Frequency vs. period: Due to the spectral range of the signals involved, seismology traditionally uses period (s) to describe the waves, whereas in seismics, frequency (Hz) is used. Waves provide information about the medium through which they propagate at the scale of their wavelength, use of higher frequencies in seismics (shorter wavelengths) leads therefore to a greater resolution (of the structure) compared to seismology.

Wave-propagation: seismology – again historically – mostly uses refracted energy, whereas exploration seismics is often synonymous with reflection seismics, although refraction seismic exploration predates the latter.

This essay will be mainly concerned with acquisition and processing of reflection seismic data. Note, however, that seismics is being increasingly applied to both shallower depths (high-resolution seismics) and crustal-scale studies down to Moho and beyond (deep seismics), see *Deep Seismic Reflection and Refraction Profiling* for details of the latter. Seismic data acquisition and processing is a broad subject, the treatment here will have to make choices based upon space constraints, etc. Some subtopics, e.g., "*Seismic, Migration*" are, however, dealt with in separate essays.

The reader is assumed to be familiar with the basic theory of elasticity and wave-propagation, and the related concepts of reflection, refraction, and scattering. The concept of rays will be frequently used – especially in illustrations – for convenience; real seismic signals are of course associated with wave-fronts. Similarly, the figures will depict a 2-D (section) of the 3-D earth.

There are many good resources available even for the narrower field of Reflection Seismic Data Acquisition and (Signal) Processing, e.g., Vermeer (2002), Yilmaz (2001), Menke (1989), Liner (2004), and Sheriff and Geldart (1995); the last one also contains some historical background and material over refraction seismics. Recently, some resources have also been made available for downloading on the internet, e.g., Claerbout (1985a, b).

In this article, all-capitals will be used to denote acronyms for jargons, of which there are quite a few (e.g., TWT above); phrases within double quotes will refer to an article elsewhere in this volume.

Seismic data acquisition

Before seismic signals could be processed, an artificial wave-field has to be generated using suitable *sources* at appropriate locations, measured by *receivers* at other locations after getting reflected back from within the earth, and stored using *recorders*. Design of a seismic survey (*geometry*) needs inputs from regional geology, exploration objectives, and logistical considerations.

At first confined to land, seismic surveys are now-a-days carried out mostly in marine environments in round-the-clock operations using large vessels and a lot of instrumentation; single-channel seismics has faded away in favor of multi-channel acquisition, allowing much more information to be obtained (see *Single and Multichannel Seismics*). Table 2 gives an overview of the equipments used under different field environments.

Seismic sources

One needs a signal that is high in energy (amplitude) to ensure a good depth penetration, and short in duration to ensure accurate determination and *differentiation* of the travel-times – a Dirac-Delta spike would be ideal, which, however, is a-causal. In practice, a sharp, compact, and *repeatable* signal is preferred. This quasi-idealized wave-form, possessing finite temporal duration and frequency band-width (both with respect to the ambient noise), is called a *wavelet*. The source wavelet changes form as it travels through the earth due to several physical processes to be briefly discussed below.

Repeatability of the source wavelet – that is, that of its *amplitude* and *phase* content – is an important prerequisite for the later processing steps. Explosives were the initial choice for source on land, providing large energy (good depth of penetration) but having non-repeatable signal shape and negative environmental impact. Development of large, truck-mounted electromechanical vibrators has led since 1960s to their increasing use in land-seismics, with both above disadvantages of impulsive sources reduced significantly.

In marine environment, compressed air is used – with explosion (*air gun*) or implosion (*water gun*) – to create the acoustic waves. The sources are towed by a ship together with the receivers (*single vessel seismic*), or, by a separate ship (*two-ship seismics*).

Seismic Data Acquisition and Processing, Table 2 Sources and receivers used in seismic surveys

Environment	Sources	Receivers
Land	Explosives/ vibrators/impact	Geophones
Marine	Air/water -guns	Hydrophones
Water bottom	Explosives/guns	Geo/hydro-phones
Onshore-offshore	Explosives/guns	Geo/hydro-phones

There have also been experiments with shear-wave sources – both impact-type and vibratory. These – either alone, or together with compressive sources – can provide extra information about the subsurface medium. For investigating shallower structures in engineering, environmental and archeological applications, small impact-based sources, e.g., weight-drops, rifles, etc., and even portable vibrators are being frequently used, and provide the required higher resolution.

Seismic receivers

All land-seismic (and seismological) receivers (*geophones, seismometers*) are electromechanical devices, that transform the relative motion of the medium in which they are embedded, into electrical voltages. Fidelity of this transformation, both in *amplitude* and *phase*, is important to ensure maximum information retention for later retrieval – a flat amplitude response, with no phase distortion within the band of frequencies that are of interest, would be ideal. The output of a geophone can be made to be proportional to the displacement/velocity/acceleration associated with the causative motion. Dennison (1953) provides an overview of the physico-mathematical underpinnings of geophone design; see also *Seismic Instrumentation*.

Originally, geophones were designed to move – and hence record information – only in the vertical direction. Later, the importance of recording and analyzing the entire three-dimensional elastic wave-field came to be realized. Multi-component receivers, enabling recording/analysis of both horizontal components, or, all three spatial components of the ground movement are being increasingly used even in large-scale surveys.

For use in water, small piezoelectric elements – hydrophones – are employed to record pressure variations – modern deployments typically consist of thousands of such elements being towed near the water surface by *streamers*, several kilometers long, which are liquid-filled plastic tubes, fitted with fins (for buoyancy), gps receivers (for location information) and fiber-optic cables (to transfer the data) to the ship.

Finally, three-component receivers may be deployed together with hydrophones at the water bottom (4C), to record the wave-field across it, see *Ocean Bottom Seismics*.

Note, that both sources and receivers may be deployed in *groups*, using specific patterns, which affect the generation and sampling of the wave-field due to their direction-dependent radiation/reception characteristics.

Seismic recorder

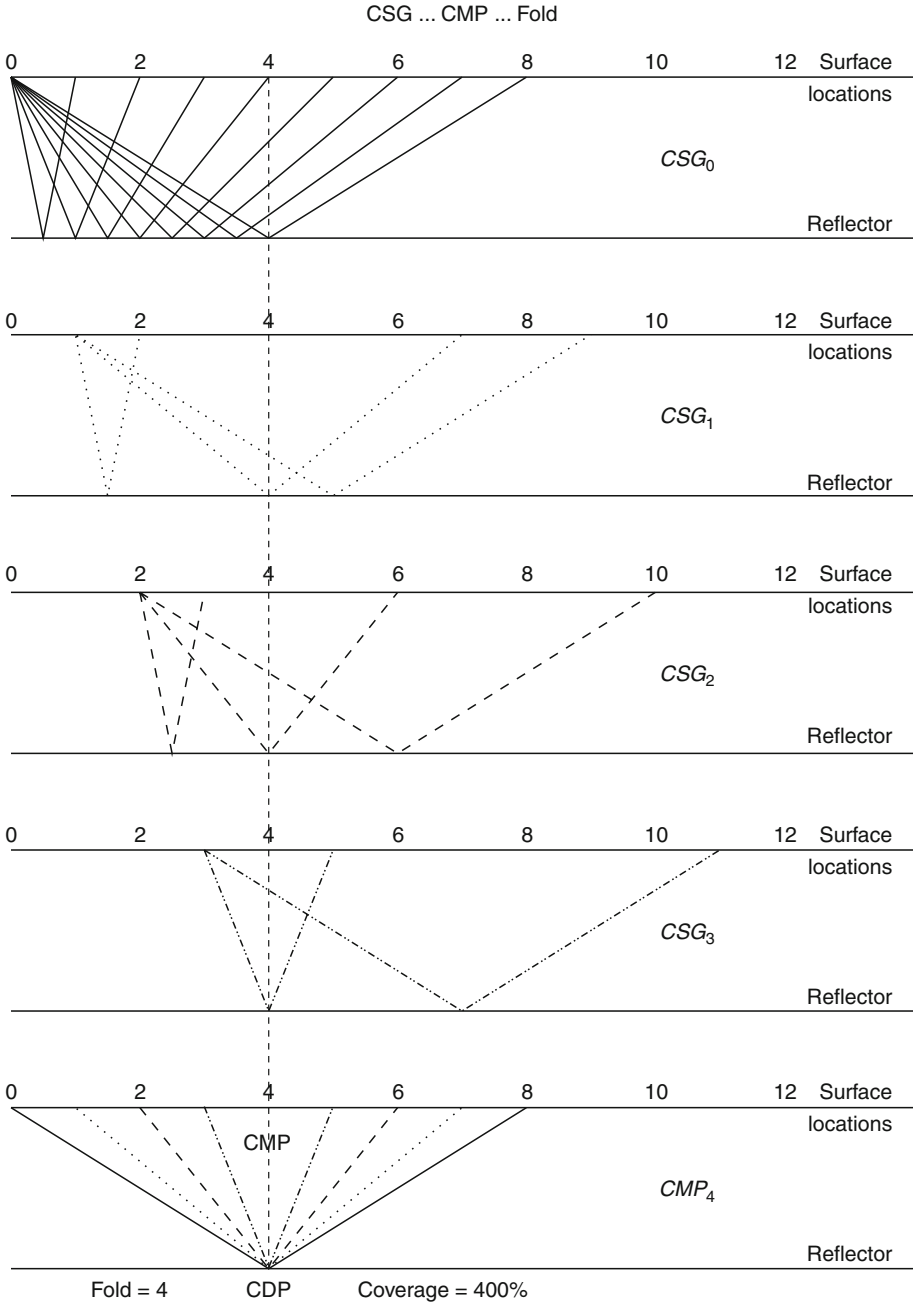
The time-varying electrical signals output by the receivers represent *arrivals* back-scattered from different depths, all juxtaposed in time and embedded within the ever present background noise, and require storage for later processing. In the beginning, photographic films and later magnetic tapes were used for this purpose. The digital revolution starting in the 1960s, itself partly driven by the needs of seismic data acquisition and processing, caused a complete shift to in-situ digitization and digital storage. Similarly, the wires connecting the receivers to the recorder have been mostly replaced by fiber-optic cables or wireless. Preserving the frequency, amplitude, and phase of the signal and the desired *dynamic range* are important considerations in designing the digitizing (*sampling*) unit. As each *receiver (group)* corresponds to a different *channel* in the recorder, digitization in such systems (typically consisting of thousands of channels) must preserve the time-base, to enable comparison of the arrival times between different *traces*. Also, the actual time of the *shot*, t_0, must be transferred from the source to the recorder and recorded; it is often used to start the recording process itself, as seismics is only interested in travel-times, i.e., arrival-times with respect to t_0. The digitized data – uniformly sampled time-series – from each individual experiment (*shot*), consisting of multiple *traces* (output of receivers), is called a *seismic record*.

Acquisition geometry

Assuming a *layer-cake* model (sedimentary beds parallel to the surface), early surveys deployed a number of sources and receivers along a straight line on the earth-surface to obtain a vertical cross-section of the geology below this line (2-D).

Figure 1 shows schematically the approach in such a survey, and is – in spite of its simplifications – useful in understanding several basic ideas. All the five panels show the earth-surface at the top, and a reflecting boundary (*target*), parallel to it, at some depth. Numerals on the surface represent surveyed equi-distant flag-positions to denote locations. The top panel shows the first measurement, with the source at "0" and eight receivers at locations "1" through "8". Assuming a homogeneous and isotropic medium, the paths predicted by Snell's law for a part of the source energy to first travel downward to the target and then reflect upward from it to reach the receivers are indicated by the oblique lines.

The signals output from the receivers are digitized in the recorder to yield a *seismic record*, i.e., a collection of *seismic traces*. Such an ordered collection of seismic traces is called a *gather*. Having a common *source*, the record resulting from our first measurement is a *common source gather* (CSG_0), the suffix denoting source position.



Seismic Data Acquisition and Processing, Figure 1 Schematics of seismic data acquisition by common-mid-point (CMP) profiling. Panels CSG_0 through CSG_3 represent common source gathers (CSG); CMP_4 is a common mid point gather for one common depth point (CDP). See text for details.

Under the twin-idealizations of no background noise, and a spike-like source signal, each channel in the recorder (*seismic trace*) will consist of one single blip corresponding to the arrival time of the signal; in reality, the arrivals will have random background oscillations due to noise, and one *wavelet* corresponding to the single reflection arrival. Assuming constant speed of propagation v and depth to the target H, it is trivial to show (e.g., Sheriff and Geldart, 1995) that the travel-times to the receivers can be written as $t_x^2 = (x^2 + 4H^2)/v^2 = t_0^2 + x^2/v^2$, t_x being the arrival time recorded by a receiver at a source-receiver *offset* of x. The travel-time curve for such a situation is thus a hyperbola – this simple relationship underlies much of seismic processing. t^2 plotted against x^2 thus yields a straight line, the slope being v^{-2}, i.e., square of the

slowness of the medium. Note that in seismics, the *velocity*, which sensu stricto is a vector, is almost always used to denote the local wave *speed* (a scalar), which is a property of the medium (rocks) . . . we shall follow this usage.

Our aim is to find H, the depth to the target ($= t_0/2v$). We have thus to estimate t_0 from the rest of the reflection hyperbola. Note that the reflection points on the target for the different receivers are different.

Hence, in what has become almost universal practice, the measurement is repeated after shifting the whole set-up laterally along the measurement line, keeping all the relative distances the same. In panel 2 of Figure 1, the source and the receivers have been shifted right by one unit; only a few ray-paths are shown for this gather (CSG_1). Similarly, gathers CSG_2 and CSG_3 are also measured and recorded. During these measurements, the same receiver locations recorded signals from different sources so that a post-measurement re-arrangement of the traces could also yield *common receiver gathers* (CRG); in our case, we would obtain $CRG_1 - CRG_{11}$. These are useful for certain processing situations.

The lowest panel of this figure shows a special kind of re-sorting, collecting the traces from the four *shots* with one *common reflection point* (CRP). Four traces corresponding to source-receiver combinations of 0/8, 1/7, 2/6 and 3/5 were selected respectively from the four gathers. For our simple geometry, the four ray-paths shown share two things – a *common mid point* (CMP_4) between their respective source and receiver locations and the *common depth point* (CDP) at the target depth, the latter being the same as CRP. Such a gather is called a CMP-gather, and indexed by the position of the CMP. The travel-time plot of the reflection arrivals in a CMP-gather is also a hyperbola.

The four ray-paths shown for the gather CMP_4 all have the same reflection point, and thus contain information about the same subsurface geology. The arrival times of the reflection signal in the four traces are of course different, as the travel paths are different. If this difference is corrected for, then adding the four traces should increase the coherent signal (information regarding the CRP) with respect to the *random* noise. The improvement of S/N by adding N traces is given by

$$\frac{\sum(\text{N traces with identical signal})}{\sum(\text{N traces with random signal})} \approx \frac{N}{\sqrt{N}} = \sqrt{N}.$$

The improvement of the signal-to-noise (S/N) ratio is thus roughly proportional to the square-root of the number of traces added. This number (4 in our case) depends upon the survey geometry, and is called the *fold* of the survey. Starting from fold 1 for $CMP_{0.5}$ (not shown), it gradually builds up to its nominal value (4 in this case), and again drops-off at the other end of the survey.

Acquisition configuration can be specified by expressions describing the position of the source relative to the receivers, viz., *end-on, split-spread, broad side*, etc. Depending upon the geology and the noise regime, these configurations, as also varying *fold*, leave subtle but important *footprints* on the data.

In reality, the geology is of course not as in Figure 1, presence of structure (*dips, faults, folds*, etc.) is what makes hydrocarbon accumulation possible in the first place. Processing of 2-D data can remedy this situation – though only partially. Availability of more equipment and data processing power led therefore to development of 3-D acquisition, with receivers laid out on the surface in a 2-D pattern, and sources also positioned in a different 2-D pattern, thus causing a better illumination of the subsurface by the seismic waves. Here too the basic concept of adding *fold* number of traces in a CMP-gather holds sway – point-shaped CMPs and CDPs being replaced by finite *bins*, their sizes depending upon the survey design and objectives (see Vermeer, 2002 for further insight into acquisition design).

In areas with structural complexity, the simplifying assumptions of CMP-processing break down, and the availability of computer power may make it possible – nay desirable – to process each trace of the recorded CSG separately, to try to obtain a better image (see also *Seismic Imaging, Overview* and *Seismic, Migration*).

Restricting the deployment to the surface of the earth implies – as we shall see later – a bias for horizontal structures; this was eventually removed by carrying out measurements inside bore-holes called VSP; see *Vertical Seismic Profiling* for details. Finally, better recording instrumentation coupled with the need to detect changes in the hydrocarbon reservoirs resulting from exploitation has given rise to *time lapse seismic* (4D), whereby repeat imaging of the same area, carried out after several years of production, is used to validate/improve production models for reservoirs.

Seismic data processing

Introduction

Reflection seismic data, acquired in the field, has to be taken through several processing steps, before it can be interpreted in terms of the subsurface structure. The source signal, on its way down, and back up to the receivers is modified by many factors; the aim of processing is to *undo* (i.e., correct for) as many/much of these effects as possible, leaving only the effects due to the causative structure of interest (geology) to be interpreted.

Seismic data is a spatio-temporal sampling of the back-scattered seismic wave-field, an ordered collection of traces, and can be considered to be a 2-D or 3-D data matrix along with some auxiliary information regarding location, etc. The traces themselves are an ordered collection of uniformly sampled amplitude values (*time-series*), with relevant information contained in their respective headers in (internationally) agreed formats. All processing steps aim to improve the spatio-temporal S/N ratio of the data by reducing the noise and/or by sharpening the wave-form (to improve the resolution).

Signal versus noise

Before proceeding further, it is useful to reflect on the terms *signal and noise*. That it is a matter of perspective is clear from this relative definition: *signal is useful noise and noise is useless signal*. In other words, someone's noise is someone else's signal, and vice-versa. For example, the *ground-roll*, hated in reflection seismics, is useful in surface-wave seismology and shallow-seismics. Amazingly, using noise for seismic imaging has now become a field of active research (see the section "*Seismic Noise*" for references).

In reflection seismics, signal is synonymous with *primary reflection. Primaries*, as these are often referred to, represent seismic energy reflected only once during its travel from source to receiver. Everything else, present in the traces, is taken to be *noise*. This includes multiply reflected energy (*multiples*), diffractions (caused by *sharp* structures in the subsurface, e.g., faults, pinch-outs), refracted arrivals, surface waves (*ground-roll*). Non-geological noise sources include nature (wind, waves, animals) and man (traffic, industry, etc.). From processing point of view, noise could be coherent (ground-roll, water-pump, multiples), or, incoherent, each needing a different strategy. See *Seismic Noise* for details.

Kinematics of the seismic signal (primaries)

Starting with some simple (but inaccurate) assumptions, e.g., horizontal layering, constant speed, etc., useful structural information can be extracted – a large data volume contributing to the robustness of the processing algorithms (also see *Seismic, Migration*). In this section, we focus on the travel-times of the waves (visualized as *rays*), see also additional information in "*Seismic, Ray Theory*."

NMO

The travel-time for a primary reflection from a horizontal reflector, shown earlier to be hyperbolic, can be rewritten as:

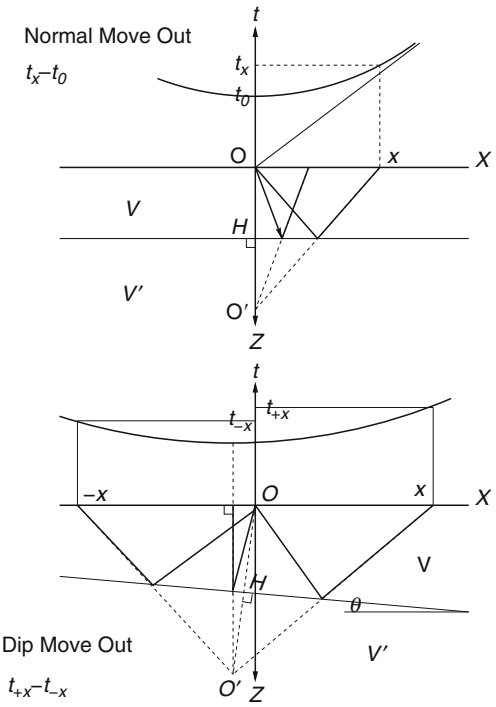
$$t_x - t_0 = \frac{\sqrt{x^2 + 4H^2} - 2H}{v}. \tag{1}$$

The quantity on the left is the difference (see Figure 2) between the oblique reflection-time at source-receiver offset (distance) x and the vertical TWT, and leads to the relation:

$$\Delta t_x = \frac{2H}{v}\left(\sqrt{1 + \frac{x^2}{4H^2}} - 1\right) = t_0\left(\sqrt{1 + \frac{x^2}{4H^2}} - 1\right). \tag{2}$$

Expanding the expression under square-root, and recognizing that in most seismic measurements offset $\ll$ target depth, we obtain the approximate relation (Equation 3), which could be improved by retaining additional higher order terms.

$$\Delta t_x \approx \frac{x^2}{4vH} = \frac{x^2}{2v^2 t_0} \tag{3}$$



Seismic Data Acquisition and Processing, Figure 2 NMO of a primary reflector; horizontal (above), dipping (below). See text for details.

Δt_x ($= t_x - t_0$) is called the *normal move-out* (NMO) associated with the reflection travel-time. NMO can be used to align the primary reflection in all traces at t_0 (TWT) by removing the effect of source-receiver distance (offset), i.e., by *flattening the reflector*. NMO, an important concept in seismics, is used both to first identify *primaries*, and later to align them for imaging the reflector. Note that to use 3, we need to know x (source-receiver offset), v (speed), and H (target depth); in practice, x is known and iteration is used to obtain optimal values for v and H.

Dipping bed

For a dipping reflector (Figure 2), travel-time for the primary reflection is still hyperbolic, given by

$$v^2 t_\theta^2 = x^2 + 4H^2 + 4Hx\,\sin\theta. \tag{4}$$

The minimum of the hyperbola is now shifted *updip*; the quantity $t_{+x} - t_{-x}$ is a measure of the asymmetry, and can be used to estimate the dip.

Many reflectors: layer-cake

Dix (1955) considered the case of many reflectors parallel to the surface – a good starting model for sedimentary sequences – and showed, that here too, the travel-time curve can be approximated at short offsets by a hyperbola:

Seismic Data Acquisition and Processing, Table 3 Jargons associated with the term velocity in seismics

Jargon	Brief description
v_{int}	Speed in a geological interval (assumed constant)
v_{av}	Average speed between two points along a ray path
v_{app}	Apparent speed measured by receivers in field (= dx/dt)
v_{nmo}	Speed used for NMO correction (strictly, only for one layer)
v_{rms}	Dix' root-mean-square NMO velocity for layer-cake situation
v_{stk}	*Best* velocity to stack CMP-gathers
v_{mig}	*Best* velocity to migrate the seismic data

$$t_x^2 \approx t_0^2 + \frac{x^2}{v_{rms}^2}, \quad \text{with} \quad v_{rms} = \sqrt{\frac{\sum v_i^2 \Delta t_i}{\sum \Delta t_i}}. \tag{5}$$

The homogeneous velocity v (= v_{nmo}) is now replaced by v_{rms} (root-mean-square velocity), which depends upon the velocities of the layers v_i and the vertical transit times t_i through them. v_{rms} plays a role similar to v_{nmo} in flattening the primaries in the multi-layer case. Individual layer-velocities may then be computed from the Dix' equation:

$$v_n = \sqrt{\frac{v_{rms,n}^2 t_n - v_{rms,n-1}^2 t_{n-1}}{t_n - t_{n-1}}}. \tag{6}$$

Velocities in seismics

In seismics, different terms are used to denote "velocity" depending upon the context. Table 3 lists a few, along with brief explanations, some of these will be elaborated later.

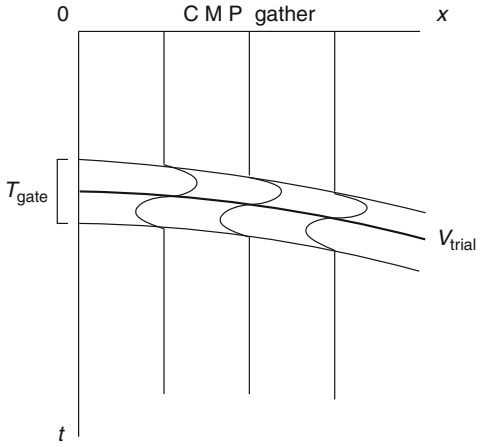
NMO stretch

After NMO correction, a time-interval Δt_x, say corresponding to a period of the wavelet recorded on a trace at an offset x, becomes Δt_0; the wavelet is thus distorted (*stretched*). The expression 100 $(\Delta t_0 - \Delta t_x)/\Delta t_x$ is a percentage measure of this stretch – 0% implying no distortion. In practice, a threshold percentage is specified to exclude parts of data – relatively largeoffsets and small arrival times – from being NMO corrected (and taking part in further processing).

Semblance: a measure of signal alignment

To apply optimal NMO correction, a quantitative measure of alignment of amplitudes, across several traces, is useful. Such a measure of similarity between n (amplitude) values, called *semblance*, is defined by

$$S = \frac{\left(\sum_n val\right)^2}{n \sum_n val^2}, \quad \text{and}$$
$$S_{gate} = \frac{\sum_{gate} \left(\sum_n val\right)^2}{\sum_{gate} \left(\sum_n val^2\right)}. \tag{7}$$



Seismic Data Acquisition and Processing, Figure 3 Schematical drawing showing calculation of multi-channel semblance. *Curved bold line* represents the move-out curve for a trial velocity, the *two surrounding lines* represent the boundaries of the time-gate; see text for details.

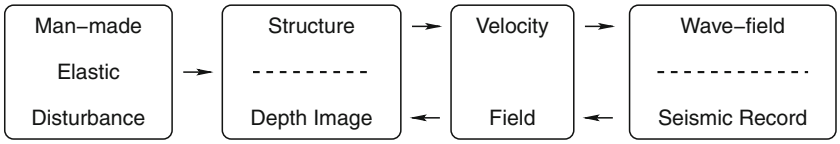
Note that semblance is a dimensionless number between 1 (perfect match), and is 0 (perfect mismatch). The second form uses a time-gate along the traces, generally having the width of the dominant period of the signal, for increased robustness. Semblance is used extensively in modern *reflection velocity analysis*, to evaluate the goodness of alignments of primary reflections along move-out curves computed for a range of trial velocities (Figure 3).

Velocity: processing point of view

Wave-speed (called velocity in seismics) in the medium is the missing link needed to convert the travel-time information to depths required for structural interpretation – and eventual drilling. Note that *velocity* is needed to find the structure (geology), but structure is needed to find the velocity. This catch-22 situation is solved iteratively – shown schematically in Figure 4.

Velocity is a macroscopic (wavelength-scale average) property of the rock depending upon the density and elastic properties of the minerals making up the lithology (see *Seismic Properties of Rocks*). In rocks of interest in seismics (*sandstone, shale, limestone*), velocity is not a good indicator of lithology, with considerable overlap in values, with some exceptions, e.g., salt, anhydrite (relatively higher velocity). Presence of *porosity and pore-fluids (water, oil, gas)* is the most important factor for this overlap, and is in turn caused by the *burial history* of the rocks. Wave-propagation in fluid-filled porous media is described by Biot-Gassman theory, see Lee (2008) for references and recent developments.

Propagation velocity (the missing link) can be estimated by direct measurements (see, e.g., Sheriff and Geldart, 1995 for details), which have shortcomings though (see Table 4). The velocity used for processing



Seismic Data Acquisition and Processing, Figure 4 Iteratively solving for both structure and velocity in seismics.

Seismic Data Acquisition and Processing, Table 4 Direct determination of seismic velocities and their shortcomings

Method	Shortcoming
Uphole-time	Useful only for the weathering layer
Check-shots, well-shoot	Limited depth-range, destructive
VSP	Available late in exploration, expensive
Sonic log	Available late, noisy (high-frequency)
Lab measurements	Limited availability

seismic reflection data is usually determined iteratively from the data itself, and will be described later.

Amplitude changes along the propagation path

Several factors cause the amplitude of the seismic waves to change as they travel from source to receiver. These can be corrected for, so as not to mask the weaker changes (signals) of interest.

Geometrical spreading

Conservation of energy requires a continuous reduction of amplitude, as a seismic wave-front spreads through a medium – hence the term *geometrical spreading*. The loss depends upon the mode of spreading and the distance traveled (r). For primaries (body-waves), amplitude ($\propto \sqrt{\text{energy}}$) decreases $\propto r^{-1}$, whereas for ground-roll (surface-wave), the decrease is $\propto r^{-1/2}$, the latter shows why ground-rolls, with their (relatively) large amplitudes, are a big problem in seismics.

Absorption

The propagating wave continuously loses energy due to absorption too, which is a physical property of the medium, and can be described by several equivalent parameters, e.g., *absorption coefficient, damping factor*, etc., the most common being the *quality factor* of the material $Q = 2\pi/$[fractional energy lost per cycle]. It is a dimensionless quantity, with a value 0 implying perfect absorption and ∞ implying perfect elasticity. Absorption, with Q considered to be frequency-independent within the band-width of interest in seismics, causes relatively greater attenuation of higher frequencies – leading to a change in the wave-form during propagation. See *Seismic, Viscoelastic Attenuation* for more details.

Energy partitioning at interfaces

Boundaries of geological heterogeneities (layering, faults, etc.) also cause changes in the amplitude of the wavelet; such changes are, indeed, of prime interest in seismics. As in optics, the interaction between the wave-fronts and the geological structure depends upon their relative dimensions, i.e., their radii of curvature-with specular reflections and point-scattering building the two end-members, both of which are encountered in seismics. Another concept from optics, *diffraction*, is useful to understand the complexity of the interaction between the wave-front and the medium. See *Seismic Diffraction*; *Seismic Waves, Scattering*, and *Energy Partitioning of Seismic Waves* for additional details.

Waveforms: convolution, deconvolution

Factors modifying the source signal along the path of the seismic wave may be divided as: near-source (*ns*), i.e., weathering layer, earth (*e*), i.e., the target geology, near-receiver (*nr*), receiver (*r*), and recorder (*rec*), with the output trace (o) as the final result. Each of these, denoted in Equation 8 below by the expression in parentheses, affects (*filters*) the source wavelet (s). In a series of papers/reports (Robinson, 2005; Treitel, 2005), the MIT geophysical analysis group (GAG) laid the foundation of the *digital revolution* in seismic data processing, by examining the nature of these filters and developing methods to undo their effects. These resulted in major advances in *time-series analysis and digital filtering* (Robinson and Treitel, 1964), and a critical evaluation of the (statistical) nature of earth's reflectivity (*target geology*).

Convolutional model of the seismic trace

As the source- and recorded-signals are both time-series (uniformly sampled, ordered collection of amplitudes), it is useful to represent all the other elements mentioned in the above-paragraph also as such. For a column of vertically layered reflectivity, such a time-series would correspond to values equal to *RC*s placed at times converted from depths using velocities. Now, making the crucial assumption that all these filter elements are linear systems, the recorded trace can be expressed as:

$$o(t) = s(t) * ns(t) * e(t) * nr(t) * r(t) * rec(t) + n(t). \quad (8)$$

In Equation 8, $*$ (*star*) is the convolution operator, well-known in the theory of linear systems; $n(t)$ represents some additive noise which does not follow this model, hopefully, it is mostly removed early in the processing.

The time-series that transform $s(t)$ into $o(t)$ can also be interpreted as the *impulse response* of the corresponding elements, e.g., $r(t)$ is the response of the receiver to a sudden spike signal. Using *Fourier Transforms* to change the time-series into their *spectra*, and remembering that convolution in time-domain corresponds to multiplication in frequency domain, one obtains:

$$O(\omega) = S(\omega) \cdot NS(\omega) \cdot E(\omega) \cdot NR(\omega) \cdot R(\omega) \cdot REC(\omega), \tag{9}$$

where the noise term has been neglected (see Sheriff and Geldart, 1995 for introduction to linear operators and Fourier theory). Equation 9 clearly shows the *filtering* effect of the different elements, each one modifying the spectrum of the incoming signal by modifying/removing a part of its frequencies. Our aim, in seismic data processing, is to extract $e(t)$, the geological structure, from the recorded signal $o(t)$.

Deconvolution as inverse filtering

Undoing the act of the filterings implied in Equations 8 and 9 is called *deconvolution* (decon), or, *inverse filtering*. Equation 9 can be rewritten as $O(\omega) = \mathrm{E}(\omega).\,\mathrm{REST}(\omega)$, where $\mathrm{REST}(\omega)$ groups together all the elements on the right besides the geology. Then, $E(\omega)$, or $e(t)$, can be *estimated* from

$$E(\omega) \cong O(\omega)/REST(\omega), \quad \text{or,} \qquad e(t) \cong o(t) * rest(t)^{-1}. \tag{10}$$

The approximation sign, for both forms of Equation 10 – in frequency domain (first), or, in time-domain (second) – is necessary, even in the *noise-free* case. Spectral division needs precautions to avoid zero-division in parts of the spectrum, where frequencies have been weakened/removed. Fortunately, addition of *noise* helps, since signals of interest in seismics exist – by definition – only above the ambient noise level. See Liner (2004), Sheriff and Geldart (1995) and Yilmaz (2001) for the stabilizing role of *spectral whitening* in decon.

Wavelet processing

Wavelets: Let's take a closer look at *seismic wavelet*, introduced in the section about seismic sources, as a signal of finite frequency band-width and temporal duration. Using standard concepts from time-series analysis (Sheriff and Geldart, 1995; Yilmaz, 2001), simple examples of wavelets are:

$$a : (3, -2, 1), \quad b : (2, 3, -1) \quad \text{and} \quad c : (-1, 2, 3),$$

the numbers representing uniformly sampled amplitudes starting from $t = 0$. Remembering that squares of the amplitudes in a wave(let) are measures of energy, we see that these three wavelets, while looking very different, have the same total energy. Depending upon the energy build-up, wavelet a is called *minimum delay* (energy is front loaded), b is *mixed delay*, and c is *maximum delay*; physical (causal) wavelets are minimum delay, although in the example, a is not strictly causal, due to the instantaneous build-up of energy at $t = 0$. In frequency domain, the expressions minimum/mixed/maximum-phase are used instead.

Wavelet estimation: *Auto-correlation* of the wavelets a-c are all symmetrical about $t = 0$, i.e., have no phase information, e.g., $\phi_{bb} = (-2, 3, 14, 3, -2)$; these are Fourier Transforms of the respective power-spectra. In seismics, an estimate of the power spectrum is often available from the data. The question then arises whether an estimate of the wavelet may be obtained from it – an outline follows. Using Z-transform notation, one can write the wavelet, say c, and its auto-correlation as polynomials:

$$C(Z) = -1 + 2Z + 3Z^2, \text{ and,}$$
$$\Phi_{cc}(Z) = -3Z^{-2} + 4Z^{-1} + 14 + 4Z - 3Z^2,$$

Z being the *unit-delay* operator, its powers denoting time-shifts with respect to $t = 0$. According to the fundamental theorem of algebra, a polynomial of degree n in Z must have n roots, i.e., it can be expressed as a product of n factors of the form: $(Z - Z_1)(Z - Z_2) \ldots (Z - Z_n)$, each factor representing a basic wavelet (*doublet*). Half the doublets of an auto-correlation polynomial are minimum delay, their product represents the Z-transform of the unique minimum delay causative wavelet. See Yilmaz (2001) and Sheriff and Geldart (1995) for details, assumptions, critical remarks, and alternate approaches (e.g., *homomorphic deconvolution*) to deconvolution of time-series.

Wavelet manipulation: Much of seismic processing is involved with manipulating the wavelet (deconvolution in a general sense). While very powerful, it contains potential for pitfalls, if applied without a proper understanding of the suitability of the particular technique, as each decon step also causes artifacts.

Spiking decon aims to sharpen the shape of the signal, to improve temporal resolution – and interpretation. Ideally, it involves convolving the wavelet with its inverse operator, to yield a spike, i.e., perfect resolution.

Zero-phasing converts the signal to one with zero-phase; the result is a symmetrical signal (a-causal), and is primarily useful for interpretation if the peak can be made to coincide with the reflecting boundary.

Any-phasing is used in merging seismic datasets of different vintages and with differing source wavelets.

General shaping groups methods to convert the signal to any desired shape optimally – using some statistical criteria.

Depending upon whether a model is available for decon, the methods could also be divided in *deterministic*, i.e., model-based and *statistical*.

Deterministic deconvolution

Vibroseis processing. Vibrators (see the section on sources) use a repeatable source signal, called *sweep*. It is a time-limited (typically, 10–20 s long) signal with the frequency continuously varying between given start- and end-values, and comes in many flavors, e.g., up-, down-, linear-, non-linear-sweeps. Neglecting other terms, one could write from Equations 8 and 9: $o(t) = s(t)^* e(t)$. The recorded signal is thus the convolution of *earth reflectivity* with the sweep signal. We could remove its effect (deconvolve) by *cross-correlating* the observed signal with the sweep (which we know precisely), a process, which is equivalent to convolving with its time-reversed version, and get

$$s(-t) * o(t) = s(-t) * s(t) * e(t) \approx \delta(t) * e(t) \qquad (11)$$

Due to the sweep signal being time-limited, its auto-correlation is not a Delta-spike (ideal), but is a symmetrical (zero-phase) signal called *Klauder wavelet*. The result is thus not quite the desired earth reflectivity (although it has the correct phase) and needs further processing for improvement (see Yilmaz, 2001; Liner, 2004; Sheriff and Geldart, 1995).

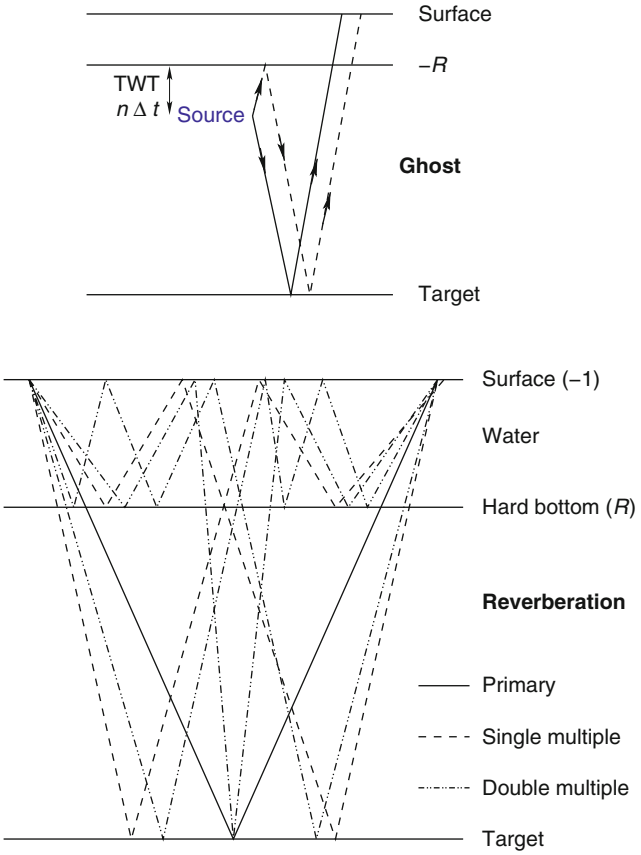
De-ghosting. The effect of large *RC*s in the shallow sub-surface has been mentioned earlier. Figure 5 shows one such situation; here the source is placed below the weathering layer, for better energy transmission toward the deeper target (ray going directly downward). A part of the wave-energy, also travels upward, and gets reflected down from the base of the weathering layer (ray going first up, and then down). In certain cases, the $RC_{\text{weathering}}$ could be quite large and negative. The energy reflected downward follows with a short delay behind the direct wave, and is called a *ghost*; the observed record is thus corrupted by that caused by a delayed ghost. Removing the latter from the recorded trace is called *deghosting,* and is an example of model-based decon. Assuming the TWT between the source and the base of the weathering to be n samples ($= n\Delta t$), one can write:

$$o(t) = s(t) - Rs(t - n\Delta t), \text{ or, using Z} - \text{transforms,}$$
$$O(Z) = S(Z) - RS(Z)Z^n = S(Z)(1 - RZ^n).$$

$(1 - RZ^n)$ is, clearly, the Z-transform of the ghost-operator. Hence, $S(Z) = R(Z)\ (1 - RZ^n)^{-1}$, or, $s(t) = o(t) + s(t - n)$. The last form above implies recursive filtering in the time-domain to achieve deghosting. Alternately, expanding $(1 - RZ^n)^{-1}$, the inverse-filter operator in time-domain can be written as

$$g(t)^{-1} = (1, 0, 0, \ldots + R, 0, 0, \ldots + R^2, 0, 0, \ldots)$$

De-reverberation. The lower part of Figure 5 shows another situation, where strong reflectivity associated with the water bottom causes long trains of high-amplitude *reverberation* of signals in the water layer. The ray-paths



Seismic Data Acquisition and Processing,
Figure 5 Deterministic deconvolution applied to ghost (above) and reverberation (below). The near vertical ray-paths are shown obliquely for better visualization, see text for details.

shown schematically are: one primary reflection from the target, two multiples reflected once in the water layer, and three multiples reflected twice; there could be many more, posing a serious problem in marine seismics. Depending upon the depth of water, certain frequencies will, as a result, experience severe distortion (enhancement or suppression). In the simplified case of a water-column with a TWT equal to the sampling interval, and remembering that the negative reflectivity causes phase-change, the total operator (signal + reverberation) can be written as:

$$w(t) = (1, -2R, +3R \ldots) \rightarrow W(Z)$$
$$= 1 - 2RZ + 3R^2Z^2 - \ldots = (1 + RZ)^{-2}.$$

It follows that the deconvolution in this case can be achieved by the operator

$$W(Z)^{-1} = (1 + RZ)^2, \quad \text{or,} \quad w(t)^{-1} = (1, 2R, R^2).$$

This elegant operator is called the Backus filter (see Backus, 1959).

Statistical deconvolution

In the absence of a deterministic model, one could attempt to change the signal wavelet to any desired shape, by designing filters that are optimal in a statistical sense. Based upon work in information theory by Norbert Wiener and others, the applications in seismics were pioneered by the MIT-GAG group, e.g., Robinson (1967). Schematically, the basic approach is:

assume filter → compute output → compute error →
minimize to get normal equations → solve for filter coefficients

If errors are assumed to be Gaussian, and l_2 norms are used, the operators obtained are called Wiener filters. Such *optimum filters* are used widely, e.g., in:

- Zero-lag spiking – to increase resolution
- Zero-phasing – to ease interpretation
- Prediction filtering – to remove multiples which are predictable, the remnant being the prediction error, corresponding to the deeper signal

Wiener optimum filter. The normal equations for the filter coefficients f are given by the matrix equation shown in Equation 12 in its compact form

$$\phi_{\text{input,input}} * f = \phi_{\text{input,output}}, \tag{12}$$

which relates the auto-correlation of the recorded (input) wavelet to its cross-correlation with the desired (output) wavelet. For the derivation of Equation 12, and a detailed treatment of statistical deconvolution, see, e.g., Yilmaz (2001) or Sheriff and Geldart (1995) – an example is shown below to illustrate the approach.

Spiking filter. If the wavelets are all n-sample long, the Matrix Equation 12 can be expanded as

$$\begin{pmatrix} \phi_{i,i}(0) & \phi_{i,i}(1) & \dots & \phi_{i,i}(n-1) \\ \phi_{i,i}(1) & \phi_{i,i}(0) & \dots & \phi_{i,i}(n-2) \\ \phi_{i,i}(n-1) & \phi_{i,i}(n-2) & \dots & \phi_{i,i}(0) \end{pmatrix} \begin{pmatrix} f_0 \\ f_1 \\ f_{n-1} \end{pmatrix} = \begin{pmatrix} \phi_{i,d}(0) \\ \phi_{i,d}(1) \\ \phi_{i,d}(n-1) \end{pmatrix} \tag{13}$$

The auto-correlation matrix $\phi_{i,i}$ in Equation 12, with the same element in each diagonal descending from left to right, is a Toeplitz matrix; f is a column vector with the filter coefficients to be determined and $\phi_{i,o}$ is a column vector with elements from the cross-correlation matrix. Equations with Toeplitz matrices can be efficiently solved by a procedure called *Levinson recursion*.

Wiener filter: a simple example. Given the input wavelet $i_t = (1, -1/2)$, let us find the optimum two-element Wiener-operator to transform it to the desired wavelet $d_t = (1, 0)$, i.e., a zero-delay unit-spike. We get $\phi_{i,i} = (5/4, -1/2)$, and, $\phi_{i,d} = (1, 0)$. Equation 13 then becomes

$$\begin{pmatrix} 5/4 & -1/2 \\ -1/2 & 5/4 \end{pmatrix} \begin{pmatrix} f_0 \\ f_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \text{yielding} \quad f_{\text{Wiener}} = \left(\frac{20}{21}, \frac{20}{21}\right). \tag{14}$$

Applying this filter to the input, we obtain the output (20/21, −2/21, −4/21), which compared to the desired output, gives a squared error of 1/21. The ideal filter for this decon is the *inverse filter* for the input wavelet. Writing $I(Z) = 1 - Z/2$ for the Z-transform of the input, the Z-transform of the inverse filter (which will convert the input to an ideal unit-spike) is $= (1 - Z/2)^{-1} = 1 + Z/2 + Z^2/4 + \ldots$, which is an infinitely long operator! For an honest comparison of its performance with that of the Wiener filter, we apply its first two terms to the input, getting the filtered version as (1, 0, −1/4); although looking better at the first glance, its squared error is 1/16, i.e., larger than that of the Wiener filter! It can be shown that the Wiener filter is the *best* two-element filter for this problem.

Suppose the input wavelet is (−1/2, 1), i.e., not minimum delay, which we want to transform to a zero-delay spike. Normal equations now give the Wiener filter as (−10/21, −4/21), with the output (5/21, −8/21, −4/21) and the squared error as 6/21. Inverse filter is now (−2, −4, −8, ...), which is extremely unstable! Its first two filter elements give the output (1, 0, −4) with 16 as error! Wiener filter performs here worse than in the first case, because, it was trying to convert a maximum delay wavelet to a minimum-delay spike, but it still does better than the (finite) inverse filter. In this case, if a maximum delay spike (0, 1) was desired, Wiener filter coefficients would be (16/21, −2/21), giving a filtered output of (−8/21, 17/21, −2/21) with a squared error 4/21, which is better than that for a zero-lag spike output. Table 5 summarizes the results.

The processing flow: putting it all together

Most of the processing *modules* (*filters*) operate on the data (time-series) sequentially, the entire process resembling a *flow*, though there are a few stand-alone modules

Seismic Data Acquisition and Processing, Table 5 Performance of Wiener and inverse filters

Input wavelet	Desired wavelet	Two-point Wiener		Two-point inverse	
		Filter	Error	Filter	Error
(1, −0.5)	(1, 0)	(20/21, 8/21)	1/21	(1, 0.5)	1/16
(−1/2, 1)	(1, 0)	(−10/21, −4/21)	6/21	(−2, −4)	16
(−1/2, 1)	(0, 1)	(16/21, −2/21)	4/21		

too. The operations could be on individual traces (*single-channel*), or on a gather of traces (*multi-channel*). Schematically, a seismic processing flow looks like:

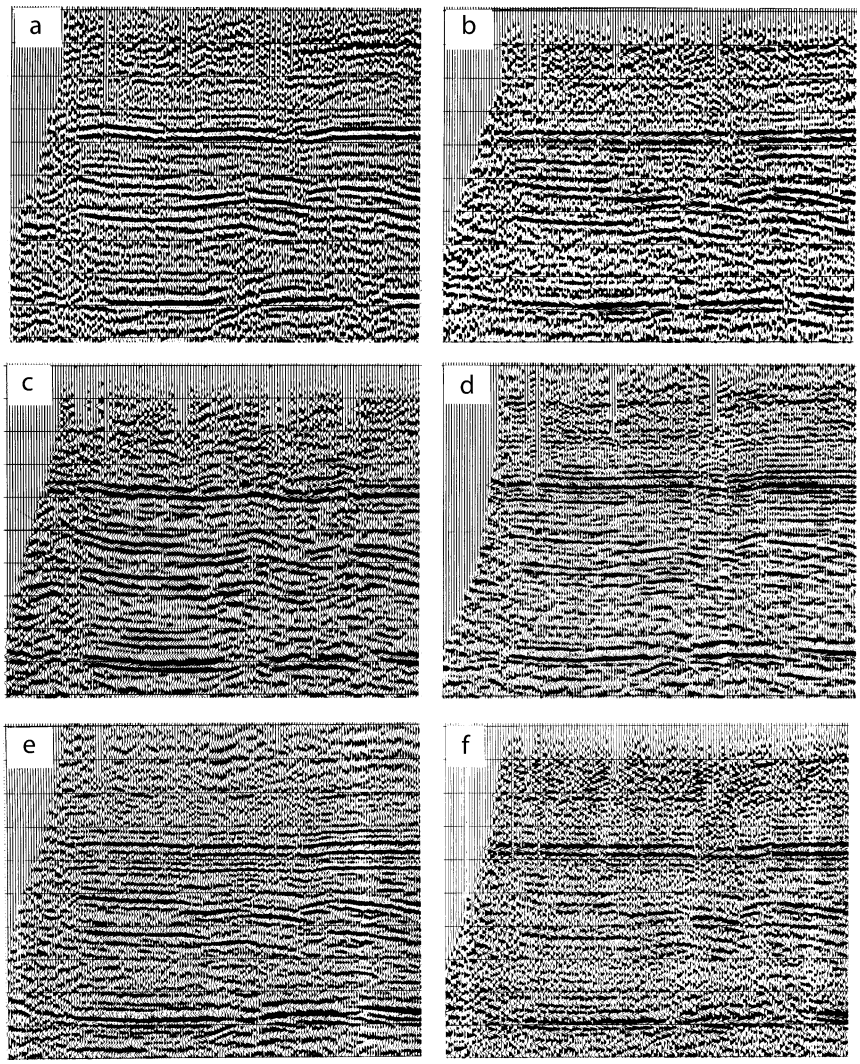
Input module → a series of processing-modules → Output module

Modules have been developed for carrying out specific tasks within the flow, e.g., static correction, band-pass filtering, stacking, migration, etc. Usually, there is a choice of modules (algorithms) available for a specific step – each with slightly different characteristics (and artifacts), and the proper selection of the modules for a flow needs both expertise and experience. This point is illustrated in the Figure 6, which shows six *different* results of processing the same data.

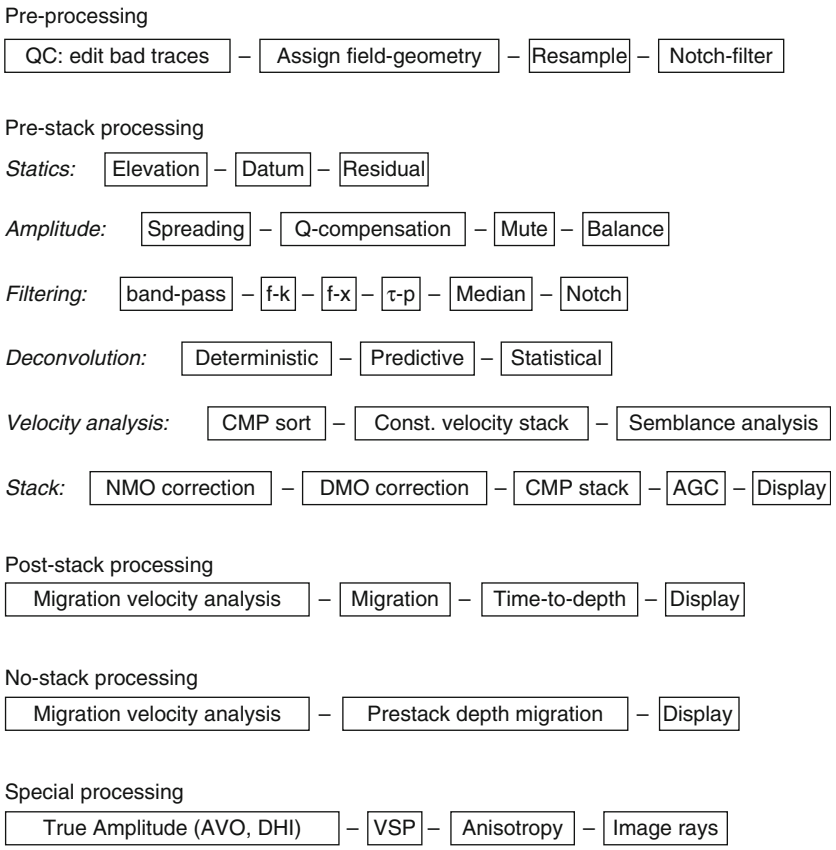
An overview of commonly applied *corrections* (processing modules) is shown in Figure 7. Space constraints will permit us to briefly describe only selected items from this list, which itself is not exhaustive; see Yilmaz (2001) for a more detailed treatment, and "*Seismic Imaging, Overview*" for additional information. Note that some modules may be applied more than once in the flow, and also, that parts of the flow may be iteratively repeated, till a reasonable result is obtained. The latter shows the importance of quality control (Q/C), by means of visual-display and other (quantitative) tools. The decision as to whether the processing of a dataset is finished depends often on the geological objectives, technical possibilities, and managerial constraints of time and money.

Pre-processing

Editing of seismic traces is an important first step, in view of the largely automated processing sequences later. *Geometry assignment* is also an essential step at this stage, and attaches acquisition information to the traces, e.g.,



Seismic Data Acquisition and Processing, Figure 6 Seismic Data Processing has no perfect answer. Seismic cross-sections produced from the same data processed by six different contractors. (Figure from Yilmaz, 2001 courtesy SEG and the author).



Seismic Data Acquisition and Processing, Figure 7 Components of seismic processing flow.

source- and receiver- coordinates. Each seismic trace is assigned a header, to store such and other information to enable efficient inter-process communication.

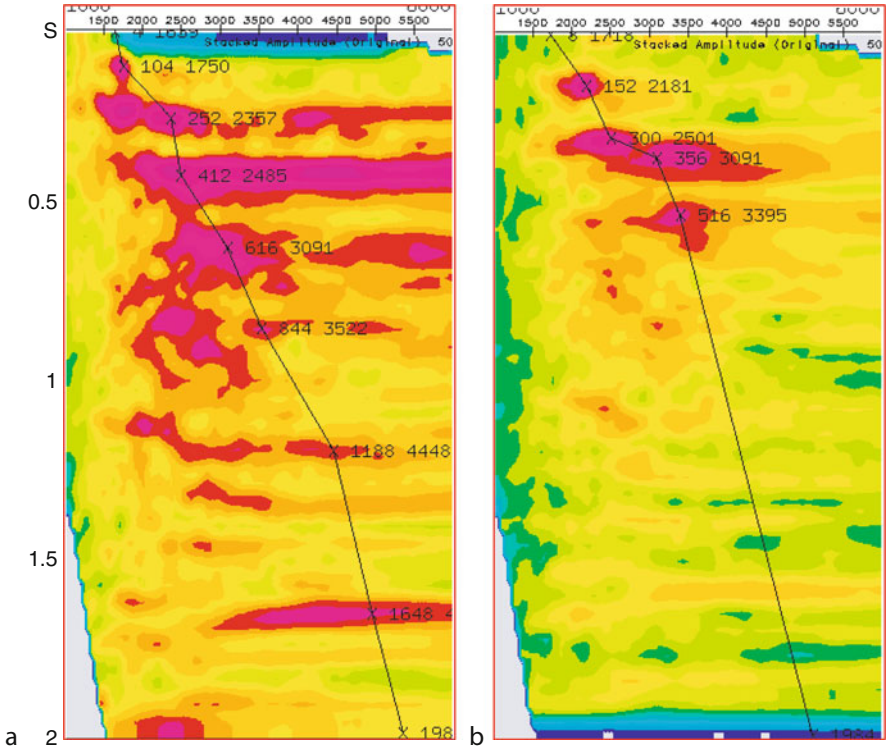
Pre-stack processing

Static corrections. These are time-invariant corrections applied to the traces, due to, e.g., elevation differences, and involve up-, or, down-shifting the entire trace in time; an example of their dramatic effect can be seen in Figure 8. The static effects due to slow lateral changes (long wavelength statics) are particularly difficult to model and can cause imaging problems. *Residual statics* involves *small* time-shifts applied late in the flow to improve the result; it uses the powerful concept of *surface-consistency* to try to correct for near-surface errors that were not modeled properly in the earlier stages. Its implementation by Rothman (1985) heralded the use of non-linear optimization (*simulated annealing, genetic algorithm*) in seismics.

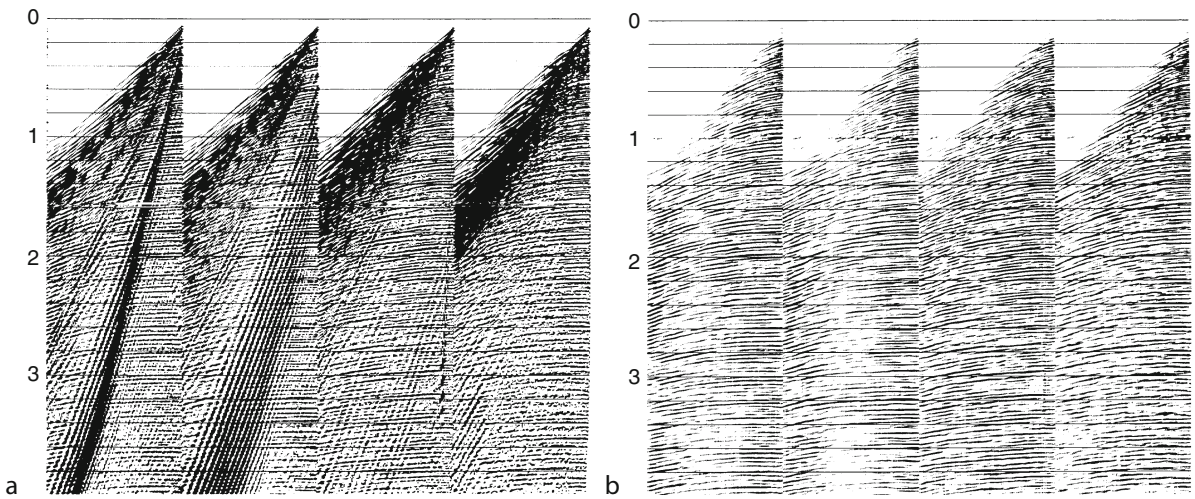
Amplitude corrections. Loss of amplitude due to *geometrical spreading* and *absorption* can be corrected for using the theory described earlier; the latter needs a Q-model, in the absence of which empirical relationships based on the total travel path/time are used. A part of the record may be removed from processing due to the presence of noise, or, suspected non-primaries; depending upon the part of the data volume removed, one then talks about *top-mute, bottom-mute, or a generalized mute*. Similarly, a *balancing* (amplitude equalization) may be applied to several adjacent traces to compensate, in an ad hoc manner, for local variations, e.g., bad receiver coupling.

Filtering, sharpness, taper. Any process that removes/reduces a part of the suspected *noise* from the data is a *filter*. Frequency-filters (high-cut, low-cut, band-pass) are the simplest examples. Data $f(t)$ is transformed using Fourier theory to its spectrum $F(\omega) = A(\omega)\ \exp^{-i\omega t}$ in the frequency domain, the amplitudes mainly corresponding to noise are zeroed-out, and the data is transformed back to the time-domain. Development of algorithms for fast and efficient Fourier transform of time-series (FFT) has caused large-scale application of digital filters.

Multi-channel data enables double-transformation of $f(x, t)$ to $F(\omega, k)$, making filtering possible based upon slopes (apparent velocities) in the $\omega - k$ plane; this is particularly effective in eliminating, e.g., slow traveling ground-roll (large amplitude surface waves), which often



Seismic Data Acquisition and Processing, Figure 8 Stacking velocity analysis using semblance (color contours). Semblance values are shown for a dataset for a range of trial velocities (*horizontal axis*), and enable interactive velocity picking as a function of TWT (*vertical axis*). The right panel shows a dramatic improvement in resolution as a result of proper static correction. (Figure from Yilmaz, 2001 courtesy SEG and the author).



Seismic Data Acquisition and Processing, Figure 9 Use of two-dimensional Fourier Transform as an apparent-velocity filter for four marine-seismic records brings out (weaker) reflections. (Figure from Yilmaz, 2001 courtesy SEG and the author).

mask the primaries. An example of such filtering is shown in Figure 9. *Notch* filters are used to remove a narrow band of frequencies, e.g., a 50-Hz noise from overhead transmission line. τ-p transforms are useful in filtering multiples, and in un-tangling far-offset data for velocity analysis, these use the Radon domain for the decomposition (Phinney et al., 1981). A few general comments apply to all filters:

- Filtering is effective only to the extent of signal-noise separation in the transformed domain.
- For any filtering, there is a trade-off between sharp cut-offs in the transform-domain and oscillatory artifacts in time-domain – and vice-versa. A compromise solution to this unavoidable problem is to apply *tapers* to smoothen the cut-off and thus minimize edge-effects.

Deconvolution. This important aspect has been dealt with in some detail in an earlier section.

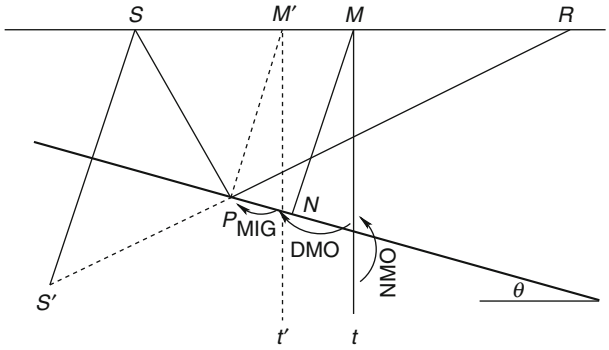
Stacking velocity analysis. This is almost always carried out in the CMP-domain, after *re-sorting* the data. The aim is to determine the velocity model, i.e., v_{rms} (TWT), to be used for computing the best move-out correction for the CMP-gathers. For a suite of velocity models, hyperbolic move-out curves are computed for the range of TWTs of interest; semblances are then computed to determine how well the arrivals in the gather line-up along these curves, and displayed in a contour plot in the v_{trial} -TWT domain, allowing interactive picking of an improved velocity model (Figure 8). The process is repeated at as many CMPs as possible – sometimes grouping neighboring CMPs together for averaging, velocity being a macroscopic property. The result is a laterally varying model of v_{rms}. Equation 6 can now be used to infer interval velocities.

CMP-stack. Once a reasonable velocity function has been determined, each trace in the CMP-gather (say, CMP_4 in Figure 1) is shifted in time by subtracting the corresponding move-out corrections. The move-out corrected traces in the CMP-gather are then added (*stacked*) together to produce one trace. This process, *CMP-stack*, reduces random noise – which does not line-up, while strengthening the reflection signal – which does, and thus improves S/N ratio of the data. Note that stacking reduces the data volume too – by a factor of *fold*! Much of the power of the seismic imaging derives from this simple step, which enhances the *primary* reflections (those only once reflected) at the expense of everything else.

Zero-offset traces/sections. The stack traces are also called *zero-offset* traces, the move-out correction having made the source and receiver coincident. A collection of stack traces is a stack- or zero-offset section, and represents the first (albeit approximate) 2-D cross-section of the subsurface. For display purposes, CMP-stack sections may be subjected to *automatic gain control* (AGC), an extremely non-linear time-variant amplitude scaling, to balance weaker/deeper signals and stronger/shallower ones.

Post-stack processing: positioning properly

The CMP-stack has one big drawback: dips were neglected throughout, which is what we are really after. This results in many artifacts in the section, e.g., crossing layering, diffraction tails, etc. Anticlinal structures are somewhat flattened, and synclinal structures could give rise to *bow-ties*.



Seismic Data Acquisition and Processing, Figure 10 Effect of dip in positioning the reflector.

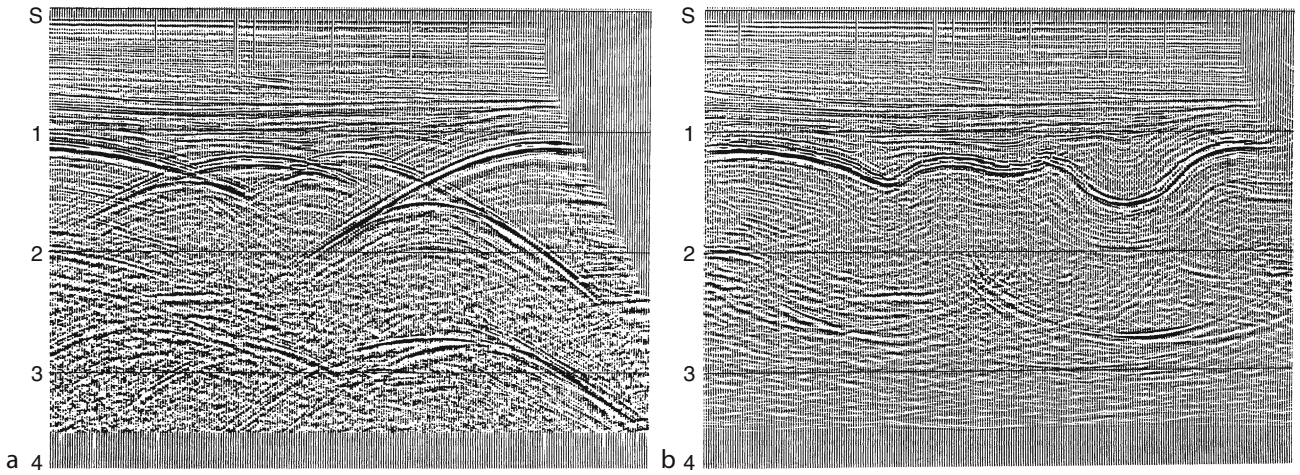
Migration. Figure 10 shows the problem schematically in the CMP-domain, for the case of a trace recorded from source S at receiver R from a reflector with a dip θ. After conventional pre-stack processing, the zero-offset trace would be plotted on the t axis below the mid-point M. This is clearly an error as the zero-offset ray for M should be incident normally on the reflector – at N; this correction is called *migration* (Yilmaz, 2001; Sheriff and Geldart, 1995).

Migration steepens and shortens energy alignments and moves these updip, clarifying the tangled image. Figure 11 shows an example of a successful migration. For further details, please see *Seismic, Migration*.

DMO. Figure 10 shows yet another error to be considered – the actual reflection point for the source-receiver combination $S-R$ is P, and not N. Even worse, for the different $S-R$ pairs making up the CMP-gather with mid-point M, the reflection points are all different, i.e., are *smeared* along the reflector, the amount of smear being dip-dependent. The process used to correct for this dip-dependent part of the move-out correction is called *DMO*. In practice, this step is applied before migration as indicated in Figure 7; Figure 10 shows the sequence:

- Reflection time is NMO corrected and plotted *below* M
- NMO corrected time is DMO corrected and plotted below M′, the *true* zero-offset point
- NMO+DMO corrected time is MIGRATED and plotted at P, the *reflection point*

Time-to-depth conversion. For final structural interpretation, the TWTs in the seismic section (stacked, or, migrated) need to be converted to depths, with velocity again playing the key role. For a homogeneous medium, this is just a rescaling of the vertical axis; with the velocity varying smoothly only in vertical direction (e.g., for flat sedimentary sequences), a nonuniform stretch of the



Seismic Data Acquisition and Processing, Figure 11 Migration positions energy from dipping structures properly. Here bow-tie like artifacts in the top part of panel (**a**) are imaged back into causative synclinal structures in panel (**b**). The artifacts persisting in the bottom of panel (**b**) probably point to lack of interest in imaging deeper structures in this case. (Figure from Yilmaz, 2001 courtesy SEG and the author).

vertical axis may suffice. Laterally varying velocities present depth-conversion problems though, increasing with the strength of the heterogeneity; ray-bending now needs to be taken into consideration.

No-stack processing: imaging complex structures
In the presence of strong lateral velocity variations (e.g., below salt structures), the conceptual model used to process CMP-gathers breaks down. Removing the simplifying assumptions makes the imaging physically more reasonable, albeit at the cost of substantially increased computational effort.

Pre-stack depth migration and migration velocity analysis. Simply put, this involves tracing the seismic energy from the source to the receiver for every recorded trace, with the philosophy that every seismic trace should be computable *if the structure and the velocity model were both known*. A detailed velocity model is essential for the success of PSDM; often a simplified model is assumed, and iteratively improved using *migration velocity analysis* (MVA). For details/issues regarding 2-D vs. 3-D, time- vs. depth- and post-stack vs. pre-stack migration, see *Seismic, Migration* and Yilmaz (2001).

Special processing
True amplitude: AVO, DHI. Observed variations of the *RC* with respect to angle of incidence may be interpreted in terms of changes in lithology across the reflecting boundary (*amplitude versus offset, or, AVO*), and may even indicate the nature of the pore-fluids. Such *direct hydrocarbon indicators* (DHI) include *bright-spots, flat-spots, polarity-reversals*, etc. (see Yilmaz, 2001; Sheriff and Geldart, 1995). A prerequisite for such analyses is *true amplitude processing*, avoiding modules that remove differential amplitude information, e.g., balancing, stacking, AGC, etc.

Converted waves. Using multi-component receivers, it is possible to identify waves that have been converted at the reflection boundary, and hence possess asymmetrical up- and down-ray-paths. Proper processing of such data, with CCP (*common conversion point*) replacing CDP, provides a better constraint for imaging.

VSP and cross-well tomography. Bore-holes can be used for placing receivers (and sources), resulting in significant noise-reduction. The first processing step now is to separate up- and down-going wave-fields, for details, see *Vertical Seismic Profiling*.

Anisotropy. Many seismic media are *anisotropic*, a common example being shales, which exhibit faster speeds parallel to the layering than across it, and require modification of procedures for proper imaging, e.g., the move-out curve would no more be hyperbolic. This field is proving important for reservoir studies too, see Helbig and Thomsen (2005) for an overview and also "Seismic Anisotropy."

Current investigations, controversies

Noise as a seismic source

This has become an exciting and active field of research (*seismic interferometry and daylight imaging*), connecting early conjectures (Claerbout, 1968) and more recent breakthroughs (Fink, 1993, 1997). See *Seismic Noise*; *Seismic, Ambient Noise Correlation* and Schuster et al. (2004) for an introduction to this evolving area.

Imaging versus inversion

Imaging tries to obtain useful (drillable) structural information using large data-redundancy and simple conceptual models, whereas inversion aims at getting values for the physical parameters of the medium, using more involved theory. *Full wave-form inversion* (see *Seismic, Waveform Modeling and Tomography*) i.e., predicting each seismic/seismological trace completely in terms of the viscoelastic properties of the medium is the ultimate goal. Note, that *perfect inversion implies perfect imaging – and vice-versa!*

Summary

The simple *echo-in-the-well* experiment mentioned at the start needs many physico-mathematical supports when applied to the earth's subsurface. Starting at data acquisition, the modules yielding the final image resemble a pipeline (flow). Several of these have been explained briefly; for others, cross-references elsewhere in this volume have been provided.

Bibliography

Aki, K., and Richards, P. G., 2002. *Quantitative Seismology*, 2nd edn. Sausalito: University Science Books.

Backus, M., 1959. Water reverberations – their nature and elimination. *Geophysics*, **24**(2), 233–261.

Claerbout, J. F., 1968. Synthesis of a layered medium from its acoustic transmission response. *Geophysics*, **33**, 264.

Claerbout, J. F., 1985a. *Fundamentals of Geophysical Data Processing*. Oxford: Blackwell. URL: http://sepwww.standford.edu/sep/prof/fgdp5.pdf.

Claerbout, J. F., 1985b. *Imaging the Earth's Interior*. Oxford: Blackwell. URL: http://sepwww.stanford.edu/sep/prof/iei2/.

Dennison, A. T., 1953. The design of electromagnetic geophones. *Geophysical Prospecting*, **1**, 3–28.

Dix, C., 1955. Seismic velocities from surface measurements. *Geophysics*, **20**, 68–86.

Fink, M., 1993. Time reversal mirrors. *Journal of Physics. D. Applied Physics*, **26**, 1333–1350. URL: http://iopscience.iop.org/0022-3727/26/9/001.

Fink, M., 1997. Time reversed acoustics. *Physics Today*, **50**, 34–40. URL: http://dx.doi.org/10.1063/1.881692.

Helbig, K., and Thomsen, L., 2005. 75-plus years of anisotropy in exploration and reservoir seismics: A historical review of concepts and methods. *Geophysics*, **70**, 9ND–23ND.

Lee, M., 2008. *Comparison of the modified Biot-Gassmann theory and the Kuster-Toksöz theory in predicting elastic velocities of sediments*. U.S. Geological Survey Scientific Investigations Report.

Liner, C., 2004. *Elements of 3D Seismology*. Tulsa: PennWell.

Menke, W., 1989. *Geophysical Data Analysis*, rev. edn. Academic Press, New York, USA.

Phinney, R., Roy Chowdhury, K., and Frazer, L. N., 1981. Transformation and analysis of record sections. *Journal of Geophysical Research*, **86**(B1), 359–377.

Robinson, E., 1967. Predictive decomposition of time series with application to seismic exploration. *Geophysics*, **32**, 418–484.

Robinson, E., 2005. The MIT Geophysical Analysis Group (GAG): 1954 and beyond. *Geophysics*, **70**, 7JA. URL: http://dx.doi.org/10.1190/1.2000287.

Robinson, E., and Treitel, S., 1964. Principles of digital filtering. *Geophysics*, **29**, 395–404.

Rothman, D., 1985. Nonlinear inversion, statistical mechanics, and residual statics estimation. *Geophysics*, **50**(12), 2784–2796.

Schuster, G. T., Yu, J., Sheng, J., and Rickett, J., 2004. Interferometric/daylight seismic imaging. *Geophysical Journal International*, **157**, 838–852.

Sheriff, R. E., and Geldart, L. P., 1995. *Exploration Seismology*, 2nd edn. Cambridge, UK: Cambridge University Press.

Treitel, S., 2005. The MIT Geophysical Analysis Group (GAG): 1954 and beyond. *Geophysics*, **70**, 31JA. URL: http://dx.doi.org/10.1190/1.1993707.

Vermeer, G. J., 2002. *3-D Seismic Survey Design*. Tulsa: Society of Exploration Geophysicists.

Yilmaz, O., 2001. *Seismic Data Analysis, Processing, Inversion and Interpretation of Seismic Data*, 2nd edn. Tulsa: Society of Exploration Geophysicists. Investigations in Geophysics, Vol. 1.

Cross-references

Energy Partitioning of Seismic Waves
Inverse Theory, Global Optimization
Ocean Bottom Seismics
Propagation of Elastic Waves: Fundamentals
Seismic Anisotropy
Seismic Diffraction
Seismic Imaging, Overview
Seismic Instrumentation
Seismic Noise
Seismic Properties of Rocks
Seismic Waves, Scattering
Seismic, Ambient Noise Correlation
Seismic, Migration
Seismic, Ray Theory
Seismic, Reflectivity Method
Seismic, Viscoelastic Attenuation
Seismic, Waveform Modeling and Tomography
Single and Multichannel Seismics
Vertical Seismic Profiling

SEISMIC DIFFRACTION

Enru Liu
China University of Mining and Technology, Xuzhou, China

Definition

Diffraction – Redistribution in space of the intensity of waves resulting from the presence of an object. It is also referred as the penetration of wave energy into areas forbidden by geometrical optics, e.g., the bending of wave energy around obstacles without obeying Snell's law as explained in Huygens' principle (generation of secondary sources).

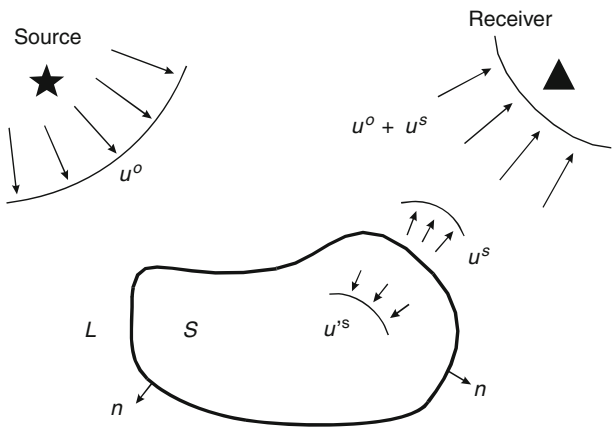
Diffraction wave-field – An event observed on seismic data produced by diffracted energy, resulting at the termination of reflectors (as at faults and other abrupt changes in seismic impedance), and it is characterized on seismic records and sections by a distinctive alignment.

Diffraction tomography – An inverse technique that is used in seismic exploration to reconstruct the physical properties under investigation using wave-equation propagation.

Introduction

When a wave (elastic wave, electromagnetic wave, or sound wave) meets an object, particle, or obstacle, it is diffracted due to scattering of energy of the propagating wave (Figure 1). However, in the literature, the terms of diffraction and scattering are often used interchangeably, and it can be sometimes confusing. Scattering and diffraction are two physical phenomena that any kind of waves can experience, but they are not the same thing. Scattering is understood in terms of particles, and behaves similarly for waves. Scattering is effectively bouncing off something. For waves, it is being absorbed and then almost immediately released in another direction. Scattering occurs because an object gets in the way of the wave. The part of the wave that strikes the object must either pass through (e.g., light through glass), be absorbed (sunburn), or be scattered (light bouncing off the wall, so we can see the wall). Diffraction is due to part of a wave being removed. It is an action taken by the part of the wave that does not strike an object. Here is an example of diffraction: Imagine a straight wave traveling forward along the surface of water. If you block the left half of the wave, the right half will not just keep moving forward, and it will expand toward the left, toward where the blocked wave would have been. A wave seems to continuously regenerate itself, always pushing forward on itself. When a section is removed, parts of the wave get pushed into the empty spot. This, in some ways, correlates to your pushing a wide row of blocks. If many people push on a few blocks each, a straight line can be maintained. If one person tries to do so, the blocks in front will tend to spread out.

In a sense, diffraction and scattering refer to a wave being redirected as a result of interacting with objects. However, a more precise definition used in optics distinguishes these two. In optics, scattering generally implies interaction of waves (or photons) with spatially uncoordinated (unordered) atoms, i.e., wave being scattered off the particles and surfaces. This means that if one looks at a picture of the scattered radiation, one would see a spatially continuous footprint. Diffraction, on the other hand, occurs when the object or part of the object is made up of ordered atoms. These atoms, being neatly arranged, "scatter" the waves or photons in a coordinated way, i.e., in specific directions, giving rise to what we can see on a film as bright spots rather. In other words, diffraction is a special type of scattering that leads to large-scale interference effects. Usually this is because the surface causing the scattering has some overall organization, such as a ruled diffraction grating or the knife-edge of a slit.



Seismic Diffraction, Figure 1 Problem configuration: a scattering object S bounded by the curve L with outward normal n. Upon an incidence of u_0 located at source, the total wave-field received at receiver is the superposition of the incident wave-field u_0 and the scattered wave-field u_s.

Although scattering and diffraction are not logically separate, the treatments tend to be separated, with diffraction being associated with departures from geometrical optics caused by the finite wavelength of the waves. Thus, diffraction traditionally involves apertures or obstacles whose dimensions are large compared to a wavelength. To lowest approximation the interaction of waves is described by ray tracing (geometrical optics). The next approximation involves the diffraction of the waves around the obstacles or through the apertures with a consequent spreading of the waves. Simple arguments show that the angle of diffraction of the waves are confined to the region $\theta \leq \lambda/a$ where λ is the wavelength and a is linear dimension of the aperture or obstacle (approximations considered work well if $\lambda/a \leq 1$). Note that diffraction may cause the localization of seismic waves – a phenomenon that is similar to the localization of lights in crystal (Larose et al., 2004). This phenomenon is caused by the focusing and defocusing of energy when seismic wave propagates through media with distributions of periodical or random distribution of scattering bodies.

In order to gain some perspective on these two extremely complex phenomena (diffraction and scattering), various theories and models have been developed in physics, and the limitations and validity of these theories are controlled by two ratios: object dimension α to wavelength (α/λ) and path-length L to wavelength λ (L/λ). In contrast to other branches of physics, in geophysics path-length is also important as we are interested both in near field as well as far field (and often near and far fields are treated differently). We also often use the dimensionless parameters $k\alpha$ and kL (where $k = \omega/\lambda$ is the wave number, ω is frequency).

Diffraction theories

Application of any diffraction model can be divided into two separate tasks. First, one must obtain the fields exiting a diffracting object (i.e., the near fields, or the boundary field values), or a reasonable approximation thereof. The second step involves propagating those fields to the

desired observation point. These are distinct and separate parts of the diffraction models. Most texts do not make this separation clear. Instead, the boundary value assumptions and subsequent propagation into the far field are lumped together into one theoretical treatment. If the resulting diffraction pattern is at all inaccurate, it is difficult to determine how much of that error is due to incorrect boundary fields and how much is the result of the propagation calculation. Because of this, it is often difficult to know which model is appropriate for a particular problem.

There are a number of different models to compute diffraction wave-field due to wave scattering, including the Huygens' Principle, the Rayleigh-Sommerfeld theory, the Kirchhoff's diffraction theory, Taylor series perturbation theory or the high-order Born approximation, Rytov phase approximation, and a model referred to as angular spectrum of plane waves. The well-known Fraunhofer and Fresnel approximations, as they appear in most introductory texts, are derived from the Kirchhoff model. Several other methods are also available for treating diffraction problems: discrete wave-number techniques; generalized ray techniques; and various numerical methods (finite difference methods; finite element method; and boundary integral or element methods). Each of these theoretical methods and models is based on some assumptions and has its strengths and weaknesses, and each can be satisfactorily employed for some ranges of problems. The choice of an appropriate model is based on what is known about a specific problem.

If there are several objects or particles, the scattering from one object will induce further scattered fields from all the other objects, which will induce further scattered fields, from all the other objects, and so on. This process is called multiple scattering. Not all theories are applicable to multiple scattering problems (in physics, this is called many-body problem). The Kirchhoff approximation ignores multiple scatterings between any two surface points. In general, it has been considered valid for the large-scale objects. Perturbation theory based on the Taylor series expansion, also sometimes called the high-order Born approximation, is valid for the small-scale objects whose dimensions are less than a wavelength (or objects whose physical properties are not too different from background solids). The Rytov phase approximation to large-scale object is not subject to the stringent restrictions that apply to the Kirchhoff approximation. Studies have shown that the Rytov approximation improves the Kirchhoff approximation in both amplitude and phase. To some degrees, the high-order Born series approximation can account for multiple scattering between closely-spaced objects. For instance, the second-order Born approximation might be sufficient to guarantee the accuracy for general rough surfaces without infinite gradients and extremely large surface heights. In contrast to other branch of field, e.g., optics, in seismology two kinds of waves exist, compressional and shear waves. These two waves can convert to each other when one meets an object. When multiple objects exist, the conversion and interaction between different wave types due to multiple scattering can be very complex. Therefore, care must be taken when one uses any diffraction theory to solve specific geophysical problems.

Geometrical theory of diffraction

The geometrical theory of diffraction (GTD) is an extension of geometrical optics that accounts for wave diffraction by edges. It was introduced in 1953 by Keller (the most commonly cited reference was published by Keller in 1962). The geometrical theory of diffraction was devised to eliminate many of the problems associated with geometrical optics. The strongest diffracted fields arise from edges, but ones of lesser strength originate from point discontinuities (tips and corners). The total field $u = (u_1, u_2, u_3)$ at an observation point $\vec{x}$ is decomposed into geometrical optic rays (the incident or reference field) u_i^0 and diffracted components u_i^d

$$u_i(\vec{x}) = u_i^0(\vec{x}) + u_i^d(\vec{x}). \tag{1}$$

The behavior of the diffracted field is based on the following postulates of GTD:

1. Wavefronts are locally plane waves.
2. Diffracted rays emerge radially from an edge.
3. Rays travel in straight lines in a homogeneous medium.
4. Polarization is constant along a ray in an isotropic medium.
5. The diffracted field strength is inversely proportional to the cross sectional area of the flux tube.
6. The diffracted field is linearly related to the incident field at the diffraction point by a diffraction coefficient (see Achenbach et al., 1982, for various analytic solutions).

GTD is a high frequency method for solving wave scattering problems from large-scale discontinuities or discontinuities in more than one dimension at the same point, and it uses ray diffraction to determine diffraction coefficients for each diffracting object-source combination. These coefficients are then used to calculate the field strength and phase for each direction away from the diffracting point. These fields are then added to the incident fields and reflected fields to obtain a total solution. Multiple scattering wave-fields cannot be easily computed using GTD.

Kirchhoff approximation

In the Kirchhoff representation of diffracted wave-fields, the ith component of diffracted wave-field u_i is computed using

$$u_i^d(\vec{x}) = -\int_{\sum} [u_k(\vec{x})] c_{kjpq}^0 \left[\frac{\partial G_i^p(\vec{x},\vec{X})}{\partial x_q}\right] n_j dS_{\vec{x}}, \tag{2}$$

where n_j is the jth component of the normal n to the surface of scattering object, $\mathbf{X}$ is a point on the face of the scattering object, $[u_k]$ is the kth displacement discontinuity across the object in the direction of n (object normal), and c_{kjpq}^0 is the elastic tensor of the background, which are often assumed to be isotropic. Equation 2 provides

a means of evaluating the diffracted field so long as the displacement discontinuity $[u]$ on the object can be estimated accurately. Since the amplitudes and phases of $[u]$ are unknown, in the Kirchhoff approximation, these are taken to be the same as if the object were infinitely long, that is, the effect of the boundary is ignored (see Douglas and Hudson, 1990; Liu et al., 1997). Therefore, the Kirchhoff approximation is a high frequency approximation, which is only applicable to diffraction from objects whose dimension is larger than the wavelength.

Perturbation theory: Born and Rytov approximation

The diffracted wave-field in Equation 1 can be written as an infinite series of the Taylor series expansion and is derived by assuming that the physical property of scattering objects is written as a perturbation of background media. The Born approximation consists of taking the incident field in place of the total field as the driving field at each point in the scatterer. It is the perturbation method applied to scattering by an extended body. It is accurate if the scattered field is small, compared to the incident field, in the scatterer. It is only valid for weak scattering or when the obstacles are small compared to both the wavelength and the spacing between the objects. Clearly, it has serious limitations when dealing with large-scale objects. The simplest approximation, called single scattering or the first-order Born approximation, is to ignore the multiply scattered field between objects. This approximation has been widely used in geophysics (Hudson and Heritage, 1981; Wu, 1982, 1989).

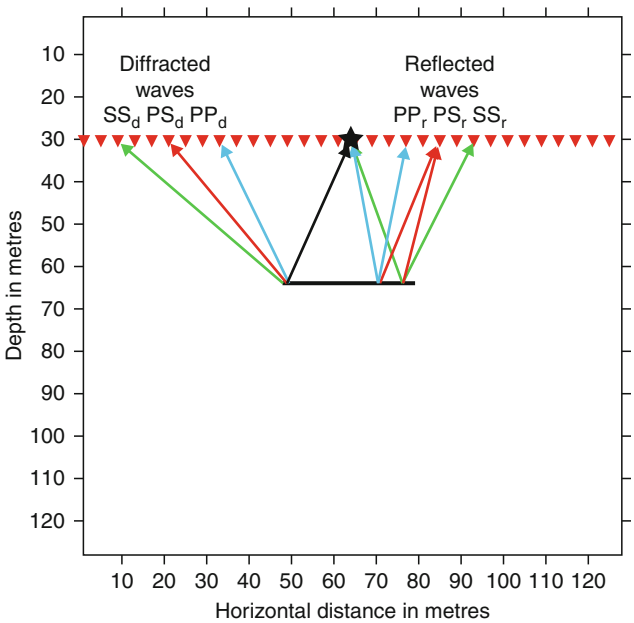
The Rytov approximation linearly perturbs the phase of a wave-field with respect to model parameters such as velocity whereas the Born approximation perturbs the amplitude. When the Green's functions for point sources are replaced by Eikonal approximations, the Rytov perturbed wave-field becomes a scaled, differentiated, time-delayed version of the reference or incident wave-field.

Numerical methods used to compute diffraction wavefield

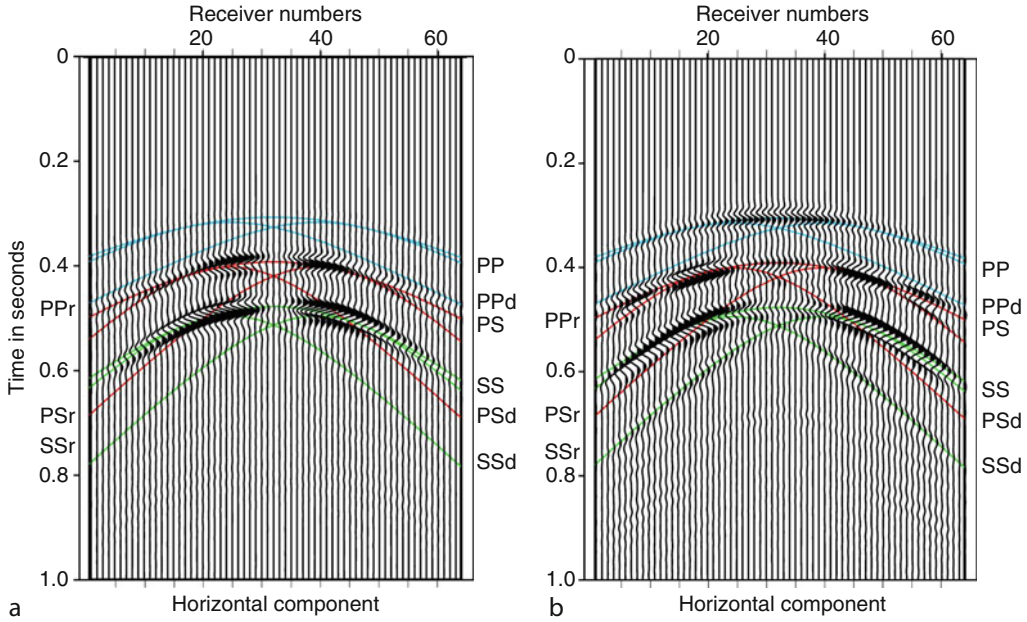
Various numerical methods can also be used to compute diffracted wave-fields, particularly now, when high-performance computers are widely available. In geophysics, finite difference methods have been used widely in the study of scattering of elastic waves by crustal heterogeneities with continuous variation of physical properties and they have also been used to model scattering by thin cracks and large fractures (see Coutant, 1989; Fehler and Aki, 1978; Coates and Schoenberg, 1995; Vlastos et al., 2003, 2007). Elastodynamic boundary integral equation or boundary element method has also been widely used to compute wave-fields from discrete inclusions with various spatial distributions as well as rough-surface topographic variations (see Bouchon, 1987; Pointer et al., 1998; Liu and Zhang, 2001; Sanchez-Sesma and Campillo, 1991, 1993).

An example of seismic diffraction from a single fracture

Some practical applications of diffraction theories in seismology include scattering from cavities; topographic variation, fractures, cracks (Figure 2). Here we give three examples to demonstrate the application of various theories in tacking diffraction problems. An example is given here for diffraction from a single fracture as computed using the finite difference method (Vlastos et al., 2003, 2007). The model geometry is shown in Figure 2. The source, receivers, and fracture are situated in an ideal elastic full space (Vp = 3,300 m/s, Vs = 1,800 m/s, density $\rho = 2.2$ g/m^3). The receiver array at which vertical and horizontal particle displacements are recorded is horizontal and 340 m above the fracture. The fracture is 300 m long. The source is located at the center of the receiver array. The source type is a vertical force. The source signal is a Ricker wavelet with a peak frequency of 25 Hz and a pulse initial time of 0.1 s. Figure 3 also shows the different kinds of waves generated by the interaction of the waves generated by the source and the fracture. The source generates both P and S waves. When they reach the fracture boundary those waves are reflected and we have PPr, PSr, SPr, and SSr waves. We calculate the theoretical ray travel-times and overlap them on the synthetic seismograms. Figures 3a and 3b show the horizontal (x) and the vertical (z) components, respectively, of the synthetic seismograms together with the theoretical ray travel-times. As we can see from both figures, we have very good agreement between the theoretical ray



Seismic Diffraction, Figure 2 Schematic representation showing diffraction from a fracture, and representation of the ray paths of the different kind of waves generated by the source that interact with the fracture.



Seismic Diffraction, Figure 3 Comparison between the theoretical ray travel-times and the synthetic seismograms generated by the modeling method: (**a**) the horizontal (*x*) and (**b**) the vertical (*z*) components.

travel-times and the synthetic seismograms. All types of waves are accurately represented in the synthetic seismograms. Owing to the type of source that we implement, we have strong arrivals at short offsets on the horizontal component and strong arrivals at long offsets on the vertical component. In addition to that, the diffracted waves from the tips of the fracture and the PPr and PPd waves are not visible in the horizontal component, but they are very clearly demonstrated in the vertical component and follow the theoretical travel-times. This is expected because the source causes vertical displacements on the medium, so very close to the source and very far away from it, the horizontal displacement is negligible. Another aspect of the comparison between the theoretical and the modeled data is that they give us further insight into the waveform patterns. For instance, we can see in both Figures 3a and 3b that in the areas of superposition between the reflected waves from the fractures and the diffracted waves from the tips we have maximum amplitude in the wave-field, as a result of constructive interference. This gives us valuable information concerning the medium we are examining.

Diffraction tomography

Seismic tomography is emerging as an imaging method for determining subsurface structure. When the view-angle coverage is limited and the scale of the medium inhomogeneities is comparable with the wavelength, as is often true in geophysical applications, the performance of ordinary ray tomography becomes poor. Other tomographic methods are needed to improve the imaging process, e.g., diffraction tomography. It has been widely used in surface reflection profiling (SRP), vertical seismic profiling (VSP), and cross-hole measurements. Theoretical formulations are derived by Wu and Toksoz (1987) for two-dimensional geometry in terms of line sources along a source line and line receivers along a receiver line. The theory for diffraction tomography is based on the Born or Rytov approximation. Multisource holography, which is similar to Kirchhoff-type migration, often gives distorted images of the object. This distortion causes long tails of the image in the case of SRP and a strong noise belt in the case of VSP and is due to incomplete and nonuniform coverage of the object spectrum. The filtering operation of diffraction tomography helps in correcting the nonuniform coverage (including duplication) of the object spectrum in the reconstruction process and therefore reduces the distortions. On the other hand, multisource holography is better suited for imaging sharp boundaries with large acoustic impedance contrasts since diffraction tomography is restricted, to weak inhomogeneities. In addition, multisource holography has the flexibility to be used with an arbitrary number of sources (including a single source). Its sampling interval is not restricted by the Nyquist frequency. Numerical examples show that combined data sets (such as surface reflection data combined with VSP data or cross-hole data combined with surface data) improve the image quality.

Summary

Diffraction refers to the spatial distribution of the intensity of seismic waves resulting from the presence of an object

(e.g., a hill- or valley-like topographic feature on the surface, or a fracture, cavity, or cave in the subsurface). It is also referred as the penetration of wave energy into areas forbidden by geometrical optics, e.g., the bending of wave energy around obstacles without obeying Snell's law as explained in Huygens' principle (as secondary sources). In the geophysical literature, the words of diffraction and scattering are often used interchangeably and it can be confusing. Diffraction and scattering are two different physical phenomena, but they are related to each other. Several analytic diffraction theories have been developed, e.g., geometrical theory of diffraction and Kirchhoff diffraction theory. More recently, numerical methods, such as finite difference and boundary element or boundary integral methods, are becoming increasingly used by geophysicists to simulate wave diffractions by complex variation of Earth's topography or subsurface cavies (cavities), fractures, irregular layers, etc. Geophysicists now often use the diffracted wave-field to reconstruct the subsurface physical properties (diffraction tomography) to solve the so-called inverse diffraction problem.

Bibliography

Achenbach, J. D., Gautesen, A. K., and McMaken, H., 1982. *Ray Methods for Waves in Elastic Solids, with Application to Scattering by Cracks*. London: Pitman Learning.

Bouchon, M., 1987. Diffraction of elastic waves by cracks or cavities using the discrete wave-number method. *The Journal of the Acoustical Society of America*, **81**, 1671.

Coates, R. T., and Schoenberg, M., 1995. Finite-difference modelling of faults and fractures. *Geophysics*, **60**, 1514.

Coutant, O., 1989. Numerical study of the diffraction of elastic waves by fluid-filled cracks. *Journal of Geophysical Research*, **94**, 17805.

Douglas, A., and Hudson, J. A., 1990. The effect on teleseismic P of the zone of damage created by an explosion. *Geophysical Journal International*, **103**, 111.

Fehler, M., and Aki, K., 1978. Numerical study of diffraction of plane elastic waves by a finite crack with application to location of a magma lens. *Bulletin of the Seismological Society of America*, **68**, 573.

Hudson, J. A., and Heritage, J. R., 1981. The use of the Born approximation in seismic scattering problems. *Geophysical Journal International*, **66**, 221.

Keller, J. B., 1962. Geometrical theory of diffraction. *Journal of the Optical Society of America*, **52**, 116.

Larose, E., Margerin, L., van Tiggelen, B. A., and Campillo, M., 2004. Weak localization of seismic waves. *Physical Review Letters*, **93**, 048501-1–048501-4.

Liu, E., and Zhang, Z. J., 2001. Numerical study of elastic wave scattering by distributed cracks or cavities using the boundary integral method. *Journal of Computational Acoustics*, **9**, 1039.

Liu, E., Crampin, S., and Hudson, J. A., 1997. Diffraction of seismic waves by cracks with application to hydraulic fracturing. *Geophysics*, **62**, 253.

Pointer, T., Liu, E., and Hudson, J. A., 1998. Numerical modelling of seismic waves scattered by hydrofractures: application of the indirect boundary element method. *Geophysical Journal International*, **135**, 289.

Sanchez-Seama, F. J., and Campillo, M., 1993. Topographic effects for incident P, SV, and Rayleigh waves. *Tectonophysics*, **218**, 113.

Sanchez-Sesma, F. J., and Campillo, M., 1991. Diffraction of P, SV, and Rayleigh waves by topographic features: a boundary integral formulation. *Bulletin of the Seismological Society of America*, **81**, 2234.

Vlastos, S., Liu, E., Main, I. G., and Li, X. Y., 2003. Numerical simulation of wave propagation in media with discrete distributions of fractures: effects of fracture sizes and spatial distributions. *Geophysical Journal International*, **152**, 649.

Vlastos, S., Liu, E., Main, I. G., and Narteau, C., 2007. Numerical simulation of wave propagation in fractured media: scattering attenuation at different stages of the growth of a fracture population. *Geophysical Journal International*, **171**, 865.

Wu, R. S., 1982. Attenuation of short period seismic waves due to scattering. *Geophysical Research Letters*, **9**, 9.

Wu, R. S., 1989. The perturbation method in elastic wave scattering. *Pure and Applied Geophysics*, **131**, 605.

Wu, R. S., and Toksoz, M. N., 1987. Diffraction tomography and multisource holography applied to seismic imaging. *Geophysics*, **52**, 11.

Cross-references

Deep Seismic Reflection and Refraction Profiling
Earthquake Rupture: Inverse Problem
Propagation of Elastic Waves: Fundamentals
Seismic, Migration
Seismic, Ray Theory
Seismic Waves, Scattering

SEISMIC DISCONTINUITIES IN THE TRANSITION ZONE

Lev P. Vinnik
Institute of Physics of the Earth, Moscow, Russia

Definition

The transition zone (TZ) is the mantle layer bounded by the 410– and 660–km seismic boundaries. The high P-wave and S-wave velocity gradients within the TZ are caused by a series of polymorphic phase transitions, the depths (pressures) of which are controlled by temperature and composition. Structure of the TZ plays an important role in the heat/mass transfer between the upper and the lower mantle.

Mineral physics data on the phase transitions in the TZ

The most frequently used model of mantle composition is pyrolite which contains ~60% of olivine $(Mg,Fe)_2SiO_4$. At a depth of ~410 km olivine (α) transforms to wadsleyite (β, modified spinel). The Clapeyron slope of this transition is positive (4.0 MPa/K, Katsura et al., 2004); the increase of the S-wave velocity is ~12%. At a depth of ~520 km wadsleyite transforms to ringwoodite (γ, silicate spinel). The velocity increase at this transition is by an order of magnitude less than at the α/β transition (e.g., Ita and Stixrude, 1992; Cammarano et al., 2005). At a depth of ~660 km ringwoodite transforms to a mixture of perovskite $(Mg,Fe)SiO_3$ and

magnesiowüstite (Mg,Fe)O. The S velocity contrast of this transition is comparable to that of the α/β transition. The post-spinel transition is sharp and has a negative Clapeyron slope of −2.8 MPa/K (Hirose, 2002). This value is disputed and some recent estimates are in a range from −0.4 to −1 MPa/K (Ohtani and Sakai, 2008).

The other components of pyrolite are orthopyroxene, clinopyroxene, and garnet. These components experience in the TZ more gradual transformations: the pyroxenes dissolve into garnet (majorite) and majorite transforms to perovskite near the bottom of the TZ. The Clapeyron slope of the post-majorite transition is 1.3 MPa/K (Hirose, 2002). The post-spinel and post-majorite phase boundaries cross each other at 1,700–1,800°C, and the corresponding seismic discontinuity is a combined effect of both transitions. At the temperatures less than 1,700°C the discontinuity is formed mainly by the post-spinel transition, whilst at the higher temperatures the post-majorite transition becomes dominant (Hirose, 2002).

Seismic methods

The actual composition and temperature of the TZ can be constrained by seismic data. An increase of the seismic velocity gradient with depth is mirrored by the increase of the apparent velocity of the P and S seismic arrivals with distance. Details can be obtained from the analysis of the related loop (triplication) in the travel times and modeling the related waveforms with synthetic seismograms (e.g., Grand and Helmberger, 1984). First indications of an increase of the velocity gradient near the top of the TZ were obtained from the analysis of incidence angles of the P waves (Galitzin, 1917). The 660-km discontinuity at the bottom of the TZ with a comparable velocity contrast was discovered in the mid-1960s (Anderson, 1965). Prior to the discovery of the 660-km boundary, the TZ was defined as the region between the 400- and 1,000-km depths (Bullen's region C).

The triplications corresponding to discontinuities in the TZ are observed at epicentral distances between ∼1,700 and ∼3,200 km, whereas properties of the mantle may change significantly on a scale of a few hundred kilometers. Observations of reflected and mode-converted phases can provide better resolution. In the neighboring field of exploration seismology, accurate mapping of discontinuities is based on observations of reflected phases. These signals are small relative to noise, and the detection is performed by using receiver arrays and stacking the recordings of different sensors with move-out time corrections. These methods are very efficient, but in seismology adequate receiver arrays are few. The idea of receiver function approach in seismology is to replace receiver arrays by seismic-source arrays that can be very large and dense. The differences in the waveforms of individual earthquakes that are recorded at the same station are eliminated by appropriate frequency filtering.

P-wave receiver functions (PRFs) present the oldest and most usable variety of receiver functions (Vinnik, 1977). This technique is based mainly on observations of Ps phases, converted from P to S. The delay of the Ps phase relative to P depends on the depth of the discontinuity and velocities along the wave-propagation paths. The amplitude of the Ps phase is proportional to the S velocity contrast at the discontinuity. A gradual transition is transparent for short-period P waves, and a shift of the spectrum of the Ps phase to lower frequencies indicates the width of the discontinuity. The detection of small seismic phases in PRFs can be accompanied by measurements of their slowness. In a laterally homogeneous Earth, Ps phases from the TZ discontinuities differ by slowness from lithospheric reverberations that arrive in the same time window, and this difference may help to separate the signals from noise. The best results can be obtained by combining source and receiver arrays. Unfortunately, practical detection of the TZ converted phases in noise is often based on relaxed criteria. Some of the widely cited results are obtained from noisy receiver functions which show a continuous train of positive "swells" and negative "troughs." The "swells" are brightly painted and the resulting optical illusion may convince readers that they see separate arrivals rather than an interference pattern.

S-wave receiver functions (SRFs) (Farra and Vinnik, 2000) are complementary to PRFs and deal with the Sp phases converted from S to P. Multiple scattering at shallow discontinuities, which presents the major source of noise in PRFs, is practically absent in SRFs, because the Sp phases from large depths arrive much earlier than the scattered phases from shallow discontinuities. Another useful method is based on observations of ScS reverberations (Revenaugh and Jordan, 1991). Detection of the phases reflected from TZ discontinuities in this method is in principle similar to that employed in receiver functions and this technique can be viewed as a variety of receiver function techniques. A shortcoming of this method is its low (in comparison to PRFs and SRFs) lateral resolution. Precursors to the seismic phases SS (SH component), PP and P′P′ (SH, P and P′ waves reflected from the Earth's surface between the source and the receiver) include phases reflected from TZ discontinuities. These phases can be detected by using the receiver function approach (Shearer, 1991). Lateral resolution of the SS-precursor technique (∼1,000 km) is by an order of magnitude lower than of PRFs, but the SS precursors are useful in the studies of the TZ of remote regions, where seismograph stations are too few for high-resolution studies.

Topography and sharpness of TZ discontinuities

The amplitudes of seismic phases converted or reflected from the major TZ discontinuities at 410- and 660-km depths vary laterally but on the average are approximately two times lower than predicted for olivine. This relationship implies that the actual composition of the TZ is broadly similar to pyrolite which contains ∼60% of olivine. A comparison of synthetic and actual wave fields for the TZ indicates that the potential temperature in the mantle differs for a homogeneous pyrolite mantle and

a mechanical mixture of basaltic and olivine components: 1,720 and 1,625 K, respectively (Ritsema et al., 2009).

The amplitudes of reflected and converted phases from the 520-km discontinuity in pyrolite should be about an order of magnitude lower than from the 660-km discontinuity and, for this reason, practically unobservable. However, there are reports on observations of a discontinuity near this depth, especially in the data on SS precursors (e.g., Flanagan and Shearer, 1998; Gu et al., 1998; Deuss and Woodhouse, 2001) and P wave recordings (e.g., Ryberg et al., 1997). Revenaugh and Jordan (1991) argued that the large 520-km discontinuity in seismic data is related to garnet/post-garnet transformation. Gu et al. (1998) discussed a possibility of very deep continental roots extending into the TZ. Deuss and Woodhouse reported splitting of the 520-km discontinuity into two discontinuities. Bock (1994) denied evidence for the 520-km discontinuity in precursors to SS.

The spectra of P410s in high-quality PRFs are usually close to the spectra of P, which indicates that the 410-km discontinuity is sharp, with a width of several kilometers (Petersen et al., 1993). Benz and Vidale (1993) observed high-frequency precursors to P′P′ with implication that the widths of both 410- and 660-km transitions are ~4 km or less. There are many other observations of short-period P′ waves reflected from the 660-km discontinuity, which indicate that this discontinuity is sharp. However, the spectra of P660s are shifted to longer periods relative to P (e.g., Petersen et al., 1993) with implication that the 660–km boundary is a few tens of kilometers wide. This controversy can be reconciled with the data of mineral physics on the two phase transitions in the same depth range.

Analysis of high-quality PRFs demonstrates that the large, in a range of a few seconds, variations of travel times of the TZ phases are caused mainly by lateral heterogeneity of the upper mantle outside the TZ (Chevrot et al., 1999). If the most heterogeneous structures are excluded from consideration, these variations are similar for P660s and P410s and can be removed by taking the difference between the P660s and P410s arrival times at the same station. Lateral variations of this difference are sensitive to changes in the TZ thickness. The best data on the P660s-P410s time difference suggest that the depths of the 410- and 660-km discontinuities are stable within several kilometers everywhere, except the anomalously hot and cold regions: hotspots and subduction zones, respectively (Chevrot et al., 1999). These normal or standard depths of the major boundaries practically coincide with those in the IASP91 model (410 and 660 km, Kennett and Engdahl, 1991), and the normal TZ thickness is 250 km. This should not be mistaken for the "average" thickness.

The results for hotspots, most of which are located in oceans, are controversial (e.g., Lawrence and Shearer, 2006; Tauzin et al., 2008; Deuss, 2007; Li et al., 2003). The related data are highly variable in quality. Therefore, instead of relying on statistics, I prefer to single out the results for Iceland, which is one of the most thoroughly investigated hotspots. The network of seismograph stations in Iceland is large, and there are several independent tomographic studies of this region. This means that the topography on the 410-km discontinuity can be separated from the effects of volumetric velocity variations. Seismic tomography reveals beneath Iceland a narrow (~200 km) columnar, low-velocity body extending into the TZ (e.g., Foulger et al., 2000), in which the 410-km discontinuity is depressed by ~15 km (Shen et al., 1998), whilst the 660-km discontinuity is at its normal depth (Du et al., 2006). No TZ anomaly could be found in the data of the station BORG in Iceland (Chevrot et al., 1999), just because the columnar body is narrow, and the station is located unfavorably with respect to it. The depressed 410-km discontinuity suggests that the TZ temperature is elevated by ~150°C. The standard depth of the 660 km discontinuity implies that either the temperature at this depth is normal, or, as indicated by high-pressure experiments (Hirose, 2002), a sensitivity of depth of the transition to temperature is low in this temperature range. Note, that the accurate estimate of depth of the 660-km discontinuity in the columnar body beneath Iceland required data on the P and S velocities in this body. For most other hotspots such data are unavailable. For the ray paths outside the columnar body, the time difference between P660s and P410s is close to the normal time with implication that the TZ of the normal oceanic mantle has the same 250-km thickness as beneath the continents.

Broadly similar results are obtained for several other hotspots. For example, similar measurements at several stations in the region of the South Pacific superswell (Suetsugu et al., 2007) show that beneath one station the thickness of the TZ is 216 ± 19 km. The anomalous region cannot be larger than a few hundred kilometers. The average for all other stations is 248 ± 5 km, very close to the standard value of 250 km. By comparison, the SS-precursor data portray in the South Pacific a large (several thousand kilometers) region, where the TZ thickness is reduced to ~230 km (Flanagan and Shearer, 1998; Gu et al., 1998; Houser et al., 2008). A similar region is found in the Atlantic. These data, if accepted at face value, imply that the TZ beneath oceans differs from that beneath continents, and the TZ beneath the hotspots is the normal oceanic TZ, contrary to the receiver function data. Substantial discrepancies between the results of the two methods exist in continental regions where numerous seismograph stations facilitate good-quality receiver function studies.

Subduction zones demonstrate another kind of anomaly. Temperature in subducted slabs is anomalously low and the equilibrium depth of the olivine-wadsleyite phase transition should be ~100 km less than in ambient mantle. However, the 410-km discontinuity in most of the presently active subduction zones apparently cannot be detected with either PRFs or other methods, most likely because the olivine-wadsleyite phase transformation is kinetically hindered and a wedge-shaped zone of olivine may persist at depths greatly exceeding 410 km (Kirby

et al., 1996). The signals from the 410-km discontinuity that are sometimes obtained in seismic studies of these regions are most likely generated outside the cold subducted lithosphere. Observations of short-period phases converted from S to P near the hypocenters of deep earthquakes show that the 660-km discontinuity in subducted slabs varies in depth but sometimes is depressed by up to ~50 km (e.g., Bock and Ha, 1984; Vidale and Benz, 1992; Wicks and Richards, 1993). This can be explained by the negative Clapeyron slope of the post-spinel transition and a low temperature of the subducted slab. The temperature anomalies can be up to ~600°C. Accurate calculations are hampered by uncertainty in the estimates of the Clapeyron slope. The studies conducted with high-resolution techniques (e.g., Niu and Kawakatsu, 1998) demonstrate that the lateral scales of the anomalies are many times smaller than imaged by SS precursors.

Robust evidence for an elevated differential time P660s-P410s and a depression of the 660-km discontinuity is provided by PRFs in the Alps where the active phase of Cenozoic subduction is over, but remnants of the oceanic lithosphere are present in the TZ (Lombardi et al., 2009). Similar anomalies are either found or can be expected in other regions of the Alpine belt.

The issue of water in the TZ

The cited results of mineral physics were obtained for the dry TZ. Many researchers examined solubility of water in minerals of the TZ (for reviews see, e.g., Bolfan-Casanova, 2005; Ohtani and Sakai, 2008; Smyth and Jacobsen, 2006). Hydration means modification of structures of the TZ minerals by incorporation of hydroxyl (OH). It appears that wadsleyite and ringwoodite may incorporate up to ~2.0 wt% of water. Pressure of the olivine-wadsleyite transition decreases by hydration by up to about 1 GPa (~30 km in depth), whereas the pressure of the post-spinel transition increases. The width of the two-phase loop between olivine and wadsleyite increases with increasing water content, and may reach a few tens of kilometers in depth relative to several kilometers for dry conditions. Hydration of 1 wt% lowers S velocity by 1–2%, whereas P velocities remain practically the same.

The expected anomalies in the depth (30 km) and thickness (a few tens of kilometers) of the 410-km discontinuity are sufficiently large to be easily detected by PRFs. However seismic observations of such effects in good-quality PRFs are practically unknown. In some regions, the seismic data for the TZ are very complicated, and hydration can be one of the reasons for the complexity, but this is hard to prove. Nolet and Zielhuis (1994) reported observations of anomalously low S velocity at depths of 300–500 km beneath the Tornquist-Teisseyre zone near the western boundary of the Russian platform. They attributed this effect to hydration at the time of closure of the Paleozoic Tornquist Ocean. This observation has been made by using surface waves. The anomalous area clearly presents a good target for application of other methods, but no such study is known yet. The lack of seismic evidence for the hydrated TZ is consistent with measurements of conductivity in the TZ (Yoshino et al., 2007), which suggest that the TZ is practically dry.

A low S velocity layer atop the 410-km discontinuity may present another possible effect of hydration of the TZ. Owing to the large water solubility, the TZ may have higher water concentration than water storage capacity of the upper mantle. Then upwelling mantle material entering the upper mantle from the TZ may undergo dehydration melting (e.g., Bercovici and Karato, 2003). Huang et al. (2005) argued that even ~0.1–0.2 wt% of water in the TZ of the Pacific is sufficient for partial melting at ~410-km depth.

The layer a few tens of kilometers thick with the S velocity reduction of a few percent atop the 410-km discontinuity was detected by Revenauh and Sipkin (1994) beneath eastern China and by Vinnik et al. (1996) beneath the Kaapvaal craton in southern Africa, from multiple ScS reverberations and PRFs, respectively. Both observations were confirmed by S receiver functions for the same locations and reproduced at a number of other locations that include Antarctica, Siberia, northern Africa, and Arabian Peninsula (Vinnik and Farra, 2007). This layer is also found in the west of North America (e.g., Song et al., 2004; Jasbinsek and Dueker, 2007). Most locations of this layer seem to be associated with Cenozoic and Mesozoic mantle upwellings (Vinnik and Farra, 2007), but hydrous melting as the reason for the low S velocity still is not proved. It might be proved by observations of anomalously high anelastic attenuation, but indications of it so far were reported only for southern Africa (Vinnik et al., 2009).

Beyond the layer atop the TZ, there are indications of a low S-wave velocity layer in a depth range between 450 and 520 km at several locations (e.g., Vinnik et al., 2009). This phenomenon might also be related to hydration and it requires further analysis.

Summary

Seismic observations of the major TZ discontinuities near 410- and 660-km depths on a global scale are broadly consistent with the pyrolite mantle model. The 410-km discontinuity is related to the olivine-wadsleyite phase transition with a positive Clapeyron slope. The 660-km discontinuity is related to the post-spinel transition in the olivine component and post-majorite transition in the other components, with a negative and positive Clapeyron slope, respectively. In the normal mantle, the depths of the discontinuities are stable and in good agreement with the IASP91 model, where the thickness of the TZ is 250 km. Anomalous topography of the TZ discontinuities is related to hot and cold regions (hotspots and subduction zones, respectively). The anomalies beneath oceans are related mainly to hotspots, where the 410-km discontinuity is depressed by ~20 km. The corresponding temperature anomalies are up to ~200°C. The lack of comparable topography on the 660-km discontinuity beneath hotspots

can be explained, among other reasons, by peculiarities of phase transitions near 660-km depth. Beneath continents, except Africa and Antarctica, the anomalies are related mainly to Cenozoic subduction zones, where the 660-km discontinuity can be depressed by up to ~50 km. The related temperature anomalies are in a range of several hundred degrees Celsius. In spite of high water solubility in wadsleyite and ringwoodite, credible observations of seismic effects of hydration in the TZ are practically unknown. The thin low-S-velocity layer atop the 410-km discontinuity, found at a number of locations, is probably the only exception. The low velocity can be an effect of a hydrous melt. On a regional scale the TZ may contain less well-understood complexities such as the 520-km discontinuity or a low S velocity layer between 450- and 520-km depths.

Bibliography

Anderson, D., 1965. Recent evidence concerning the structure and composition of the Earth's mantle. *Physics and Chemistry of the Earth*, **6**, 1–131.

Benz, H. M., and Vidale, J. E., 1993. Sharpness of upper-mantle discontinuities determined from high-frequency reflections. *Nature*, **365**, 147–150, doi:10.1038/365147a0.

Bercovici, D., and Karato, S.-I., 2003. Whole-mantle convection and the transition-zone filter. *Nature*, **425**, 39–44.

Bock, G., 1994. Synthetic seismogram images of upper mantle structure: No evidence for a 520-km discontinuity. *Journal of Geophysical Research*, **99**, 15843–15851.

Bock, G., and Ha, J., 1984. Short-period S-P conversion in the mantle at a depth near 700 km. *Geophysical Journal of the Royal Astronomical Society*, **77**(2), 593–615, doi:10.1111/j.1365-246X.1984.tb01951.x.

Bolfan-Casanova, N., 2005. Water in the Earth's mantle. *Mineralogical Magazine*, **69**(3), 229–257, doi:10.1180/0026461056930248.

Cammarano, F., Deuss, A., Goes, S., and Girardini, D., 2005. One-dimensional physical reference models for the mantle and transition zone: Combined seismic and mineral physics constraints. *Journal of Geophysical Research*, **110**, B01306, doi:10.1029/2004JB003272.

Chevrot, S., Vinnik, L., and Montagner, J.-P., 1999. Global scale analysis of the mantle Pds phases. *Journal of Geophysical Research*, **104**, 20203–20219.

Deuss, A., and Woodhouse, J., 2001. Seismic observations of splitting of the mid-transition zone discontinuity in Earth's mantle. *Science*, **294**(5541), 354–357.

Deuss, A., 2007. Seismic observations of transition-zone discontinuities beneath hotspot locations. In Foulger, G. R., and Jurdy, D. M. (eds.), *Plates, Plumes, and Planetary Processes*. Geological Society of America Special paper 430, pp. 121–131.

Du, Z., Vinnik, L. P., and Foulger, G. R., 2006. Evidence from P-to-S mantle converted waves for a flat "660-km" discontinuity beneath Iceland. *Earth and Planetary Science Letters*, **241**, 271–280.

Farra, V., and Vinnik, L. P., 2000. Upper mantle stratification by P and S receiver functions. *Geophysical Journal International*, **141**, 699–712.

Flanagan, M. P., and Shearer, P. M., 1998. Global mapping of topography of transition zone velocity discontinuities by stacking SS precursors. *Journal of Geophysical Research*, **103**, 2673–2692.

Foulger, G. R., Pritchard, M. J., Julian, B. R., Evans, J. R., Allen, R. M., Nolet, G., Morgan, W. J., Bergsson, B. H., Erlendsson, P., Jacobsdottir, S., Ragnarsson, S., Stefansson, R., and Vogfjord, K., 2000. The seismic anomaly beneath Iceland extends down to the mantle transition zone and no deeper. *Geophysical Journal International*, **142**, F1–F5.

Galitzin, B., 1917. Sur l' angle d'emergence des rayons sismiques, Nachr. D. seism. Komm., 7, Heft 2, Petersburg.

Grand, S. P., and Helmberger, D., 1984. Upper mantle shear structure of North America. *Geophysical Journal of the Royal Astronomical Society*, **76**(2), 399–438.

Gu, Y. J., Dziewonski, A. M., and Agee, C. B., 1998. Global decorrelation of the topography of transition zone discontinuities. *Earth and Planetary Science Letters*, **157**, 57–67.

Hirose, K., 2002. Phase transitions in pyrolitic mantle around 670-km depth: Implications for upwelling of plumes from the lower mantle. *Journal of Geophysical Research*, **107**(B4), 2078, doi:10.1029/2001JB000597.

Houser, C., Masters, G., Flanagan, M., and Shearer, P., 2008. Determination and analysis of long-wavelength transition zone structure using SS precursors. *Geophysical Journal International*, **174**, 178–194, doi:10.1111/j.1365-246x.2008.03719.x.

Huang, X., Xu, Y., and Karato, S., 2005. Water content in the transition zone from electrical conductivity of wadsleyite and ringwoodite. *Nature*, **434**, 746–749, doi:10.1038/nature03426.

Ita, J., and Stixrude, L., 1992. Petrology, elasticity, and composition of the mantle transition zone. *Journal of Geophysical Research*, **97**, 6849–6866.

Jasbinsek, J., and Dueker, K., 2007. Ubiquitous low-velocity layer atop the 410-km discontinuity in the northern Rocky Mountains. *Geochemistry, Geophysics, Geosystems*, **8**, Q10004, doi:10.1029/2007GC001661.

Katsura, T., Yamada, H., Nishikawa, O., Song, M., Kubo, A., Shinmei, T., Yokoshi, S., Aizawa, Y., Yoshino, T., Walter, M. J., Ito, E., and Funakoshi, K., 2004. Olivine-wadsleyite transformation in the system $(Mg,Fe)_2SiO_4$. *Journal of Geophysical Research*, **109**, B02209, doi:10.1029/2003JB002438.

Kennett, B. L. N., and Engdahl, E. R., 1991. Traveltimes for global earthquake location and phase identification. *Geophysical Journal International*, **105**, 429–465.

Kirby, S. H., Stein, S., Okal, E. A., and Rubie, D. C., 1996. Metastable mantle phase transformations and deep earthquakes in subducting oceanic lithosphere. *Reviews of Geophysics*, **34**(2), 261–306.

Lawrence, J. F., and Shearer, P. M., 2006. Constraining seismic velocity and density for the mantle transition zone with reflected and transmitted waveforms. *Geochemistry, Geophysics, Geosystems*, **7**, 1–19, doi:10.1029/2006GC001339.

Li, X., Kind, R., and Yuan, X., 2003. Seismic study of upper mantle and transition zone beneath hotspots. *Physics of the Earth and Planetary Interiors*, **136**, 79–92, doi:10.1016/S0031-9201(03)00021-9.

Lombardi, D., Braunmiller, J., Kissling, E., and Girardini, D., 2009. Alpine mantle transition zone imaged by receiver functions. *Earth and Planetary Science Letters*, **278**, 163–174.

Niu, F., and Kawakatsu, H., 1998. Determination of the absolute depths of the mantle transition zone discontinuities beneath China: Effect of stagnant slabs on transition zone discontinuities. *Earth Planets Space*, **50**, 965–975.

Nolet, G., and Zielhuis, A., 1994. Low S velocities under the Tornquist-Teisseyre zone: evidence for water injection into the transition zone by subduction. *Journal of Geophysical Research*, **99**, 15813–15820.

Ohtani, E., and Sakai, T., 2008. Recent advances in the study of mantle phase transitions. *Physics of the Earth and Planetary Interiors*, **170**, 240–247.

Petersen, N., Vinnik, L., Kosarev, G., Kind, R., Oreshin, S., and Stammler, K., 1993. Sharpness of the mantle discontinuities. *Geophysical Research Letters*, **20**(9), 859–862.

Revenaugh, J., and Jordan, T., 1991. Mantle layering from ScS reverberations 1. Waveform inversion of zeroth-order reverberations. *Journal of Geophysical Research*, **96**(B12), 19749–19762.
Revenaugh, J., and Sipkin, S., 1994. Seismic evidence for silicate melt atop the 410-km discontinuity. *Nature*, **369**, 474–476, doi:10.1038/369474a0.
Ritsema, J., Xu, W., Stixrude, L., and Lithgow-Bertelloni, C., 2009. Estimates of the transition zone temperature in a mechanically mixed upper mantle. *Earth and Planetary Science Letters*, **277**, 244–252.
Ryberg, T., Wenzel, F., Egorkin, A. V., and Solodilov, L., 1997. Short-period observation of the 520 km discontinuity in northern Eurasia. *Journal of Geophysical Research*, **102**(B3), 5413–5422.
Shearer, P., 1991. Constraints on upper mantle discontinuities from observations of long period reflected phases. *Journal of Geophysical Research*, **96**, 18147–18182.
Shen, Y., Solomon, S. C., Bjarnason, I Th, and Wolfe, C. J., 1998. Seismic evidence for a lower-mantle origin of the Iceland plume. *Nature*, **395**, 62–65.
Smyth, J. R., and Jacobsen, S. D., 2006. Nominally anhydrous minerals and Earth's deep water cycle. In Jacobsen, S. D., and van der Lee, S. (eds.), *Earth's Deep Water Cycle*. Washington, DC: AGU. Geophysical Monograph Series 168, doi:10.1029/168GM02.
Song, T. A., Helmberger, D. W., and Grand, S. P., 2004. Low-velocity zone atop the 410-km seismic discontinuity in the northwestern United States. *Nature*, **427**, 530–533, doi:10.1038/nature02231.
Suetsugu, D., Shiobara, H., Sigioka, H., Fukao, Y., and Kanazawa, T., 2007. Topography of the mantle discontinuities beneath the South Pacific superswell as inferred from broadband waveforms on seafloor. *Physics of the Earth and Planetary Interiors*, **160** (3–4), 310–318.
Tauzin, B., Debaule, E., and Wittlinger, G., 2008. The mantle transition zone as seen by global Pds phases: no clear evidence for a thin transition zone beneath hotspots. *Journal of Geophysical Research*, **113**, B08309, doi:10.1029/2007JB005364.
Vidale, J. E., and Benz, H. M., 1992. Upper-mantle seismic discontinuities and the thermal structure of subduction zones. *Nature*, **365**, 678–683.
Vinnik, L., 1977. Detection of waves converted from P to SV in the mantle. *Physics of the Earth and Planetary Interiors*, **15**, 39–45.
Vinnik, L., and Farra, V., 2007. Low S velocity atop the 410-km discontinuity and mantle plumes. *Earth and Planetary Science Letters*, **262**(3–4), 398–412.
Vinnik, L. P., Green, R. W. E., and Nicolaysen, L. O., 1996. Seismic constraints on dynamic of the mantle of the Kaapvaal craton. *Physics of the Earth and Planetary Interiors*, **95**, 139–151.
Vinnik, L., Oreshin, S., Kosarev, G., Kiselev, S., and Makeyeva, L., 2009. Mantle anomalies beneath southern Africa: Evidence from seismic S and P receiver functions. *Geophysical Journal International*, **179**, 279–298.
Wicks, W., and Richards, M. A., 1993. A detailed map of the 660-km discontinuity beneath the Izu-Bonin subduction zone. *Science*, **261**(5127), 1424–1427, doi:10.1126/science.261.5127.1424.
Yoshino, T., Manthilake, G., Matsuzaki, T., and Katsura, T., 2007. Dry mantle transition zone inferred from the conductivity of wadsleyite and ringwoodite. *Nature*, **451**, 326–329, doi:10.1038/nature06427.

Cross-references

Body Waves
Earth's Structure, Global
Earth's Structure, Upper Mantle
Geodynamics
Mantle Convection
Mantle Plumes
Seismic Phase Names: IASPEI Standard
Seismic, Receiver Function Technique
Seismology, Global Earthquake Model
Subduction Zones

SEISMIC HAZARD

Andrzej Kijko
Aon-Benfield Natural Hazard Centre, University of Pretoria, Pretoria, Republic of South Africa

Definition

Seismic hazard. Any physical phenomena associated with an earthquake (e.g., ground motion, ground failure, liquefaction, and tsunami) and their effects on land, man-made structure, and socioeconomic systems that have the potential to produce a loss. It is also used without regard to a loss to indicate the probable level of ground shaking occurring at a given point within a certain period of time.
Seismic hazard analysis. Quantification of the ground motion expected at a particular site.
Deterministic seismic hazard analysis. Quantification of a single or relatively small number of individual earthquake scenarios.
Probabilistic seismic hazard analysis. Quantification of the probability that a specified level of ground motion will be exceeded at least once at a site or in a region during a specified exposure time.
Ground motion prediction equation. A mathematical equation which indicates the relative decline of the ground motion parameter as the distance from the earthquake increases.

Introduction

The estimation of the expected ground motion which can occur at a particular site is vital to the design of important structures such as nuclear power plants, bridges, and dams. The process of evaluating the design parameters of earthquake ground motion is called seismic hazard assessment or seismic hazard analysis. Seismologists and earthquake engineers distinguish between seismic hazard and seismic risk assessments in spite of the fact that in everyday usage these two phrases have the same meaning. Seismic hazard is used to characterize the severity of ground motion at a site regardless of the consequences, while the risk refers exclusively to the consequences to human life and property loss resulting from the occurred hazard. Thus, even a strong earthquake can have little risk potential if it is far from human development and infrastructure, while a small seismic event in an unfortunate location may cause extensive damage and losses.

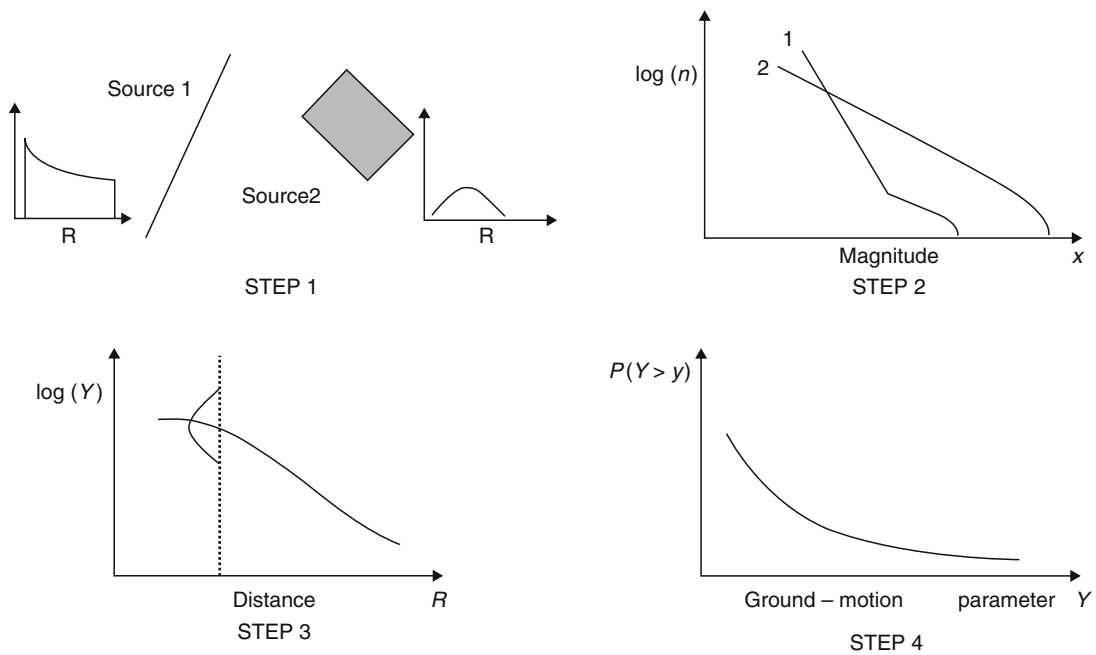
Seismic hazard analysis can be performed *deterministically*, when a particular earthquake scenario is considered, or *probabilistically*, when likelihood or frequency of specified earthquake size and location are evaluated.

The process of *deterministic* seismic hazard analysis (DSHA) involves the initial assessment of the maximum possible earthquake magnitude for each of the various seismic sources such as active faults or seismic source zones (SSHAC, 1997). An area of up to 450 km radius around the site of interest can be investigated. Assuming that each of these earthquakes will occur at the minimum possible distance from the site, the ground motion is calculated using appropriate attenuation equations. Unfortunately, this straightforward and intuitive procedure is overshadowed by the complexity and uncertainty in selecting the appropriate earthquake scenario, creating the need for an alternative, *probabilistic* methodology, which is free from discrete selection of scenario earthquakes. Probabilistic seismic hazard analysis (PSHA) quantifies as a probability whatever hazard may result from all earthquakes of all possible magnitudes and at all significant distances from the site of interest. It does this by taking into account their frequency of occurrence. Deterministic earthquake scenarios, therefore, are a special case of the probabilistic approach. Depending on the scope of the project, DSHA and PSHA can complement one another to provide additional insights to the seismic hazard (McGuire, 2004). This study will concentrate on a discussion of PSHA.

In principle, any natural hazard caused by seismic activity can be described and quantified by the formalism of the PSHA. Since the damages caused by ground shaking very often result in the largest economic losses, our presentation of the basic concepts of PSHA is illustrated by the quantification of the likelihood of ground shaking generated by earthquakes. Modification of the presented formalism to quantify any other natural hazard is straightforward.

The classic procedure for the PSHA includes four steps (Reiter, 1990), (Figure 1).

1. The first step consists of the identification and parameterization of the *seismic sources* (known also as *source zones*, *earthquake sources*, or *seismic zones*) that may affect the site of interest. These may be represented as area, fault, or point sources. Area sources are often used when one cannot identify a specific fault. In classic PSHA, a uniform distribution of seismicity is assigned to each earthquake source, implying that earthquakes are equally likely to occur at any point within the source zone. The combination of earthquake occurrence distributions with the source geometry, results in space, time, and magnitude distributions of earthquake occurrences. Seismic source models can be interpreted as a list of potential scenarios, each with an associated magnitude, location, and seismic activity rate (Field, 1995).
2. The next step consists of the specification of temporal and magnitude distributions of seismicity for each source. The classic Cornell–McGuire approach assumes that earthquake occurrence in time is random and follows the Poisson process. This implies that earthquake occurrences in time are statistically independent and that they occur at a constant rate. Statistical independence means that occurrence of future earthquakes does not depend on the occurrence of the past earthquake. The most often used model of earthquake magnitude recurrence is the



Seismic Hazard, Figure 1 Four steps of a PSHA (Modified from Reiter, 1990).

frequency-magnitude Gutenberg–Richter relationship (Gutenberg and Richter, 1944).

$$\log(n) = a - bm, \tag{1}$$

where n is the number of earthquakes with a magnitude of m and a and b are parameters. It is assumed that earthquake magnitude, m, belongs to the domain $<m_{\min}, m_{\max}>$, where $m_{\min}$ is the level of completeness of earthquake catalogue and magnitude $m_{\max}$ is the upper limit of earthquake magnitude for a given seismic source. The parameter a is the measure of the level of seismicity, while b describes the ratio between the number of small and large events. The Gutenberg–Richter relationship may be interpreted either as being a cumulative relationship, if n is the number of events with magnitude equal or larger than m, or as being a density law, stating that n is the number of earthquakes in a specific, small magnitude interval around m. Under the above assumptions, the seismicity of each seismic source is described by four parameters: the (annual) rate of seismicity λ, which is equal to the parameter of the Poisson distribution, the lower and upper limits of earthquake magnitude, $m_{\min}$ and $m_{\max}$, and the b-value of the Gutenberg–Richter relationship.

3. Calculation of ground motion prediction equations and their uncertainty. Ground motion prediction equations are used to predict ground motion at the site itself. The parameters of interest include peak ground acceleration, peak ground velocity, peak ground displacement, spectral acceleration, intensity, strong ground motion duration, etc. Most ground motion prediction equations available today are empirical and depend on the earthquake magnitude, source-to-site distance, type of faulting, and local site conditions. The choice of an appropriate ground motion prediction equation is crucial since, very often, it is a major contributor to uncertainty in the estimated PSHA.
4. Integration of uncertainties in earthquake location, earthquake magnitude and ground motion prediction equation into probability that the ground motion parameter of interest will be exceeded at the specified site during the specified time interval. The ultimate result of a PSHA is a *seismic hazard curve*: the annual probability of exceeding a specified ground motion parameter at least once. An alternative definition of the hazard curve is the frequency of exceedance versus ground motion amplitude (McGuire, 2004).

The following section provides the mathematical framework of the classic PSHA procedure, including its deaggregation. The most common modifications of the procedure will be discussed in the section (Some modifications of Cornell–McGuire PSHA procedure and alternative models).

The Cornell–McGuire PSHA methodology

Conceptually, the computation of a seismic hazard curve is fairly simple. Let us assume that seismic hazard is characterized by ground motion parameter Y. The probability of exceeding a specified value y, $P[Y \geq y]$, is calculated for an earthquake of particular magnitude located at a possible source, and then multiplied by the probability that that particular earthquake will occur. The computations are repeated and summed for the whole range of possible magnitudes and earthquake locations. The resulting probability $P[Y \geq y]$ is calculated by utilizing the Total Probability Theorem which is:

$$P[Y \geq y] = \sum P[Y \geq y|E_i] \cdot P[E_i], \tag{2}$$

where

$$P[Y \geq y|E_i] = \int \cdots \int P[Y \geq y|x_1, x_2, x_3 \ldots] \cdot f_i(x_1) \cdot f_i(x_2|x_1) \cdot f_i(x_3|x_1, x_2) \ \ldots \ dx_3\, dx_2\, dx_1. \tag{3}$$

$P[Y \geq y|E_i]$ denotes the probability of ground motion parameter $Y \geq y$, at the site of interest, when an earthquake occurs within the seismic source i. Variables x_i $(i = 1, 2, \ldots)$ are uncertainty parameters that influence Y. In the classic approach, as developed by Cornell (1968), and later extended to accommodate ground motion uncertainty (Cornell, 1971), the parameters of ground motion are earthquake magnitude, M, and earthquake distance, R. Functions $f(\cdot)$ are probability density functions (PDF) of parameters x_i. Assuming that indeed $x_1 \equiv M$ and $x_2 \equiv R$, the probability of exceedance (Equation 3) takes the form:

$$P[Y \geq y|E] = \int_{m_{\min}}^{m_{\max}} \int_{R|M} P[Y \geq y|m, r] \, f_M(m) f_{R|M}(r|m) \, dr \, dm, \tag{4}$$

where $P[Y \geq y|m, r]$ denotes the conditional probability that the chosen ground motion level y is exceeded for a given magnitude and distance; $f_M(m)$ is the PDF of earthquake magnitude, and $f_{R|M}(r|m)$ is the conditional PDF of the distance from the earthquake for a given magnitude. The $f_{R|M}(r|m)$ arises in specific instances, such as those where a seismic source is represented by a fault rupture. Since the earthquake magnitude depends on the length of fault rupture, the distance to the rupture and resulting magnitude are correlated.

If, in the vicinity of the site of interest, one can distinguish n_S seismic sources, each with average annual rate of earthquake occurrence λ_i, then the total average annual rate of events with a site ground motion level y or more, takes the form:

$$\lambda(y) = \sum_{i=1}^{n_S} \lambda \int_{m_{\min}}^{m_{\max}} \int_{R|M} P[Y \geq y|M, R] \, f_M(m) f_{R|M}(r|m) \, dr \, dm, \tag{5}$$

In Equation 5, the subscripts denoting seismic source number are deleted for simplicity, $P[Y \geq y|m,r]$ denotes the conditional probability that the chosen ground motion level y, is exceeded for a given magnitude m and distance r. The standard choice for the probability $P[Y \geq y|m,r]$ is a normal, complementary cumulative distribution function (CDF), which is based on the assumption that the ground motion parameter y is a log-normal random variable, $\ln(y) = g(m,r) + \varepsilon$, where ε is random error. The mean value of $\ln(y)$ and its standard deviation are known and are defined as $\overline{\ln(y)}$ and $\sigma_{\ln(y)}$, respectively. The function $f_M(m)$ denotes the PDF of earthquake magnitude. In most engineering applications of PSHA, it is assumed that earthquake magnitudes follow the Gutenberg–Richter relation (1), which implies that $f_M(m)$ is a negative, exponential distribution, shifted from zero to $m_{\min}$ and truncated from the top by $m_{\max}$

$$f_M(m) = \frac{\beta \exp[(-(m - m_{\min})]}{1 - \exp[(-\beta(m_{\max} - m_{\min})]}, \quad (6)$$

In Equation 6, $\beta = b \ln 10$, where b is the parameter of the frequency-magnitude Gutenberg–Richter relation (1).

After assuming that in every seismic source, earthquake occurrences in time follow a Poissonian distribution, the probability that y, a specified level of ground motion at a given site, will be exceeded at least once within any time interval t is

$$P[Y > y;\ t] = 1 - \exp[-\lambda(y) \cdot t]. \quad (7)$$

The Equation 7 is fundamental to PSHA. For $t = 1$ year, its plot versus ground motion parameter y, is the *hazard curve* – the ultimate product of the PSHA (Figure 2). For small probabilities,

$$P[Y > y; t = 1] = 1 - \exp(-\lambda)$$
$$\cong 1 - (1 - \lambda + \frac{1}{2}\lambda^2 - ...) \cong \lambda, \quad (8)$$

which means that the probability (Equation 7) is approximately equal to $\lambda(y)$. This proves that PSHA can be characterized interchangeably by the annual probability (Equation 7) or by the rate of seismicity (Equation 5).

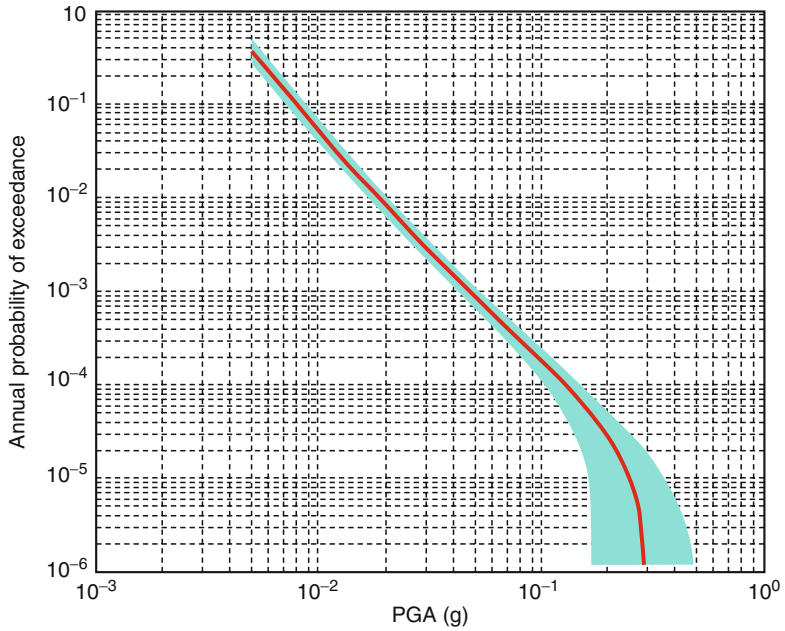
In the classic Cornell–McGuire procedure for PSHA, it is assumed that the earthquakes in the catalogue are independent events. The presence of clusters of seismicity, multiple events occurring in a short period of time or presence of foreshocks and aftershocks violates this assumption. Therefore, before computation of PSHA, these dependent events must be removed from the catalogue.

Estimation of seismic source parameters

Following the classic Cornell–McGuire PSHA procedure, each seismic source is characterized by four parameters:

- Level of completeness of the seismic data, $m_{\min}$
- Annual rate of seismic activity λ, corresponding to magnitude $m_{\min}$
- b-value of the frequency-magnitude Gutenberg–Richter relation (1)
- Upper limit of earthquake magnitude $m_{\max}$

Estimation of m_{min}. The level of completeness of the seismic event catalogue, $m_{\min}$, can be estimated in at least two different ways.



Seismic Hazard, Figure 2 Example of a Peak Ground Acceleration (PGA) seismic hazard curve and its confidence intervals.

The first approach is based on information provided by the seismic event catalogue itself, where $m_{\min}$ is defined as the deviation point from an empirical or assumed earthquake magnitude distribution model. Despite the fact that the evaluation of $m_{\min}$ based on information provided entirely by seismic event catalogue is widely used, it has several weak points. By definition, the estimated levels of $m_{\min}$ represent only the average values over space and time. However, most procedures in this category require assumptions on a model of earthquake occurrence, such as a Poissonian distribution in time and frequency-magnitude Gutenberg–Richter relation.

The second approach used for the estimation of $m_{\min}$ utilizes information on the detection capabilities of seismic stations. The approach release users from the assumptions of stationarity and statistical independence of event occurrence. The choice of the most appropriate procedure for $m_{\min}$ estimation depends on several factors, such as the knowledge of the history of the development of the seismic network, data collection, and processing.

Estimation of rate of seismic activity λ and b-value of Gutenberg-Richter. The accepted approach to estimating seismic source recurrence parameters λ and b is the maximum likelihood procedure. If successive earthquakes are independent in time, the number of earthquakes with magnitude equal to or exceeding a level of completeness, $m_{\min}$, follows the Poisson distribution with the parameter equal to the annual rate of seismic activity λ. The maximum likelihood estimator of λ is then equal to n/t, where n is number of events that occurred within time interval t.

For given $m_{\max}$, the maximum likelihood estimator of the b-value of the Gutenberg–Richter equation can be obtained from the recursive solution of the following:

$$1/\beta = \bar{m} - m_{\min} + \frac{(m_{\max} - m_{\min}) \cdot \exp[-\beta(m_{\max} - m_{\min})]}{1 - \exp[-\beta(m_{\max} - m_{\min})]}. \quad (9)$$

where $\beta = b \ln 10$, and $\bar{m}$ is the sample mean of earthquake magnitude. If the range of earthquake magnitudes $< m_{\max}, m_{\min} >$ exceeds 2 magnitude units, the solution of Equation 9 can be approximated by the well-known Aki-Utsu estimator (Aki, 1965; Utsu, 1965)

$$\beta = 1 \,/\, (\bar{m} - m_{\min}). \quad (10)$$

In most real cases, estimation of parameters λ and the b-value by the above simple formulas cannot be performed due to the incompleteness of seismic event catalogues. The alternative procedures, capable to utilize data incompleteness has been developed by Weichert (1980) and Kijko and Sellevoll (1992).

Estimation of m_{max}. The maximum magnitude, $m_{\max}$, is defined as the upper limit of magnitude for a given seismic source.

This terminology assumes a sharp cutoff magnitude at a maximum magnitude $m_{\max}$. Cognizance should be taken of the fact that an alternative, "soft" cutoff maximum earthquake magnitude is also being used (Main and Burton, 1984). The later formalism is based on the assumption that seismic moments of seismic events follow the Gamma distribution. One of the distribution parameters is called the maximum seismic moment, and the corresponding value of earthquake magnitude is called the "soft" maximum magnitude. Beyond the value of this maximum magnitude, the distribution decays much faster than the classical Gutenberg–Richter relation. Although this model has been occasionally used, the classic PSHA only considers models having a sharp cutoff of earthquake magnitude.

As a rule, $m_{\max}$ plays an important role in PSHA, especially in assessment of long return periods. At present, there is no generally accepted method for estimating $m_{\max}$. It is estimated by the combination of several factors, which are based on two kinds of information: seismicity of the area, and geological, geophysical, and structural information of the seismic source. The utilization of the seismological information focuses on the maximum observed earthquake magnitude within a seismic source and statistical analysis of the available seismic event catalogue. The geological information is used to identify distinctive tectonic features, which control the value of $m_{\max}$.

The current evaluations of $m_{\max}$ are divided between deterministic and probabilistic procedures, based on the nature of the tools applied.

Deterministic procedures. The deterministic procedure most often applied is based on the empirical relationships between magnitude and various tectonic and fault parameters, such as fault length or rupture dimension. The relationships are different for different seismic areas and different types of faults (Wells and Coppersmith, 1994 and references therein). Despite the fact that such empirical relationships are extensively used in PSHA (especially for the assessment of maximum possible magnitude generated by the fault-type seismic sources), the weak point of the approach is its requirement to specify the highly uncertain length of the future rupture. An alternative approach to the determination of earthquake recurrence on singular faults with a segment specific slip rate is provided by the so-called cascade model, where segment rupture is defined by the individual cascade-characteristic rupture dimension (Cramer et al., 2000).

Another deterministic procedure which has a strong, intuitive appeal is based on records of the largest historic or paleo-earthquakes (McCalpin, 1996). This approach is especially applicable in the areas of low seismicity, where large events have long return periods. In the absence of any additional tectono-geological indications, it is assumed that the maximum possible earthquake magnitude is equal to the largest magnitude observed, $m_{\max}^{obs}$, or the largest observed plus an increment. Typically, the increment varies from ¼ to 1 magnitude unit. The procedure is often used for the areas with several, small seismic sources, each having its own $m_{\max}^{obs}$ (Wheeler, 2009).

Another commonly used deterministic procedure for $m_{\max}$ evaluation, especially for area-type seismic sources, is based on the extrapolation of the frequency-magnitude

Gutenberg–Richter relation. The best known extrapolation procedures are probably those by Frohlich (1998) and the "probabilistic" extrapolation procedure applied by Nuttli (1981), in which the frequency-magnitude curve is truncated at the specified value of annual probability of exceedance (e.g., 0.001).

An alternative procedure for the estimation of $m_{\max}$ was developed by Jin and Aki (1988), where a remarkably linear relationship was established between the logarithm of coda Q_0 and the largest observed magnitude for earthquakes in China. The authors postulate that if the largest magnitude observed during the last 400 years is the maximum possible magnitude $m_{\max}$, the established relation will give a spatial mapping of $m_{\max}$.

Ward (1997) developed a procedure for the estimation of $m_{\max}$ by simulation of the earthquake rupture process. Ward's computer simulations are impressive; nevertheless, one must realize that all the quantitative assessments are based on the particular rupture model, postulated parameters of the strength and assumed configuration of the faults.

The value of $m_{\max}$ can also be estimated from the tectono-geological features like strain rate or the rate of seismic-moment release (WGCEP, 1995). Similar approaches have also been applied in evaluating the maximum possible magnitude of seismic events induced by mining (e.g., McGarr, 1984). However, in most cases, the uncertainty of $m_{\max}$ as determined by any deterministic procedure is large, often reaching a value of the order of one unit on the Richter scale.

Probabilistic procedures. The first probabilistic procedure for maximum regional earthquake magnitude was developed in the late 1960s, and is based on the formalism of the extreme values of random variables. A major breakthrough in the seismological applications of extreme-value statistics was made by Epstein and Lomnitz (1966), who proved that the Gumbel I distribution of extremes can be derived directly from the assumptions that seismic events are generated by a Poisson process and that they follow the frequency-magnitude Gutenberg–Richter relation. Statistical tools required for the estimation of the end-point of distribution functions (as, e.g., Cooke, 1979) have only recently been used in the estimation of maximum earthquake magnitude (Pisarenko et al., 1996; Kijko, 2004 and references therein).

The statistical tools available for the estimation of $m_{\max}$ vary significantly. The selection of the most suitable procedure depends on the assumptions of the statistical distribution model and/or the information available on past seismicity. Some of the procedures can be applied in the extreme cases when no information about the nature of the earthquake magnitude distribution is available. Some of the procedures can also be used when the earthquake catalogue is incomplete, i.e., when only a limited number of the largest magnitudes are known. Two estimators are presented here. Broadly speaking, the first estimator is straightforward and simple in application, while the second one requires more computational effort but provides more accurate results. It is assumed that both the analytical form and the parameters of the distribution functions of earthquake magnitude are known. This knowledge can be very approximate, but must be available.

Based on the distribution of the largest among n observations and on the condition that the largest observed magnitude $m_{\max}^{obs}$ is equal to the largest magnitude to be expected, the "simple" estimate of $m_{\max}$ is of the form

$$\hat{m}_{\max} = m_{\max}^{obs} + \frac{1}{n f_M(m_{\max}^{obs})}, \tag{11}$$

where $f_M(m_{\max}^{obs})$ is PDF of the earthquake magnitude distribution. If applied to the Gutenberg–Richter recurrence relation with PDF (Equation 6), it takes the simple form

$$\hat{m}_{\max} = m_{\max}^{obs} + \frac{1 - \exp[-\beta(m_{\max}^{obs} - m_{\min})]}{n\beta \exp[-\beta(m_{\max}^{obs} - m_{\min})]}. \tag{12}$$

The approximate variance of the estimator (Equation 12) is of the form

$$VAR(\hat{m}_{\max}) = \sigma_M^2 + \frac{1}{n^2} \left[\frac{1 - \exp[-\beta(m_{\max}^{obs} - m_{\min})]}{\beta \exp[-\beta(m_{\max}^{obs} - m_{\min})]}\right]^2, \tag{13}$$

where σ_M stands for epistemic uncertainty and denotes the standard error in the determination of the largest observed magnitude $m_{\max}^{obs}$. The second part of the variance represents the aleatory uncertainty of $m_{\max}$.

The second ("advanced") procedure often used for assessment of $m_{\max}$ is based on the formalism derived by Cooke (1979)

$$\hat{m}_{\max} = m_{\max}^{obs} + \int_{m_{\min}}^{m_{\max}^{obs}} [F_M(m)]^n \mathrm{d}m, \tag{14}$$

where $F_M(m)$ denotes the CDF of random variable m. If applied to the frequency-magnitude Gutenberg–Richter relation (1), the respective CDF is

$$F_M(m) = \begin{cases} 0, & \text{for } m < m_{\min}, \\ \frac{1-\exp[-\beta(m-m_{\min})]}{1-\exp[-\beta(m_{\max}-m_{\min})]}, & \text{for } m_{min} \le m \le m_{\max}, \\ 1, & \text{for } m > m_{\max}, \end{cases} \tag{15}$$

and the $m_{\max}$ estimator (Equation 14) takes the form

$$\hat{m}_{\max} = m_{\max}^{obs} + \frac{E_1(n_2) - E_1(n_1)}{\beta \exp(-n_2)} + m_{\min} \exp(-n), \tag{16}$$

where $n_1 = n/\{1 - \exp[-\beta(m_{\max}^{obs} - m_{\min})]\}$, $n_2 = n_1 \exp[-\beta(m_{\max}^{obs} - m_{\min})]$, and $E_1(\cdot)$ denotes an exponential integral function. The variance of estimator (Equation 16)

has two components, epistemic and aleatory, and is of the form

$$VAR(\hat{m}_{\max}) = \sigma_M^2 + \left[\frac{E_1(n_2) - E_1(n_1)}{\beta \exp(-n_2)} + m_{\min} \exp(-n)\right]^2, \quad (17)$$

where σ_M denotes standard error in the determination of the largest observed magnitude $m_{\max}^{obs}$.

Both above estimators of $m_{\max}$, by their nature, are very general and have several attractive properties. They are applicable for a very broad range of magnitude distributions. They may also be used when the exact number of earthquakes, n, is not known. In this case, the number of earthquakes can be replaced by λt. Such a replacement is equivalent to the assumption that the number of earthquakes occurring in unit time conforms to a Poisson distribution with parameter λ, where t is the span of the seismic event catalogue. It is also important to note that both estimators provide a value of $\hat{m}_{\max}$, which is never less than the largest magnitude already observed.

Alternative procedures are discussed by Kijko (2004), which are appropriate for the case when the empirical magnitude distribution deviates from the Gutenberg-Richter relation. These procedures assume no specific form of the magnitude distribution or that only a few of the largest magnitudes are known.

Despite the fact that statistical procedures based on the mathematical formalism of extreme values provide powerful tools for the evaluation of $m_{\max}$, they have one weak point: often available seismic event catalogues are too short and insufficient to provide reliable estimations of $m_{\max}$. Therefore the Bayesian extension of statistical procedures (Cornell, 1994), allowing the inclusion of alternative and independent information such as local geological conditions, tectonic environment, geophysical data, paleoseismicity, similarity with another seismic area, etc., are able to provide more reliable assessments of $m_{\max}$.

Numerical computation of PSHA

With the exception of a few special cases (Bender, 1984), the hazard curve (Equation 7) cannot be computed analytically. For the most realistic distributions, the integrations can only be evaluated numerically. The common practice is to divide the possible ranges of magnitude and distance into n_M and n_R intervals, respectively. The average annual rate (Equation 4) is then estimated as

$$\lambda(Y > y) \cong \sum_{i=1}^{n_S} \sum_{j=1}^{n_M} \sum_{k=1}^{n_R} \lambda_i P[Y > y | m_j, r_k] f_{M_j}(m_j) f_{R_k}(r_k) \Delta m \Delta r, \quad (18)$$

where $m_j = m_{\min} + (j - 0.5) \cdot (m_{\max} - m_{\min})/n_M$, $r_k = r_{\min} + (k - 0.5) \cdot (r_{\max} - r_{\min})/n_R$, $\Delta m = (m_{\max} - m_{\min})/n_M$, and $\Delta r = (r_{\max} - r_{\min})/n_R$.

If the procedure is applied to a grid of points, it will result in a map of PSHA, in which the contours of the expected ground motion parameter during the specified time interval can be drawn (Figure 3).

Deaggregation of seismic hazard

By definition, the PSHA aggregates ground motion contributions from earthquake magnitudes and distances of significance to a site of engineering interest. One has to note that the PSHA results are not representative of a single earthquake. However, an integral part of the design procedure of any critical structure is the analysis of the most relevant earthquake acceleration time series, which are generated by earthquakes, at specific magnitudes and distances. Such earthquakes are called "controlling earthquakes," and they are used to determine the shapes of the response spectral acceleration or PGA at the site.

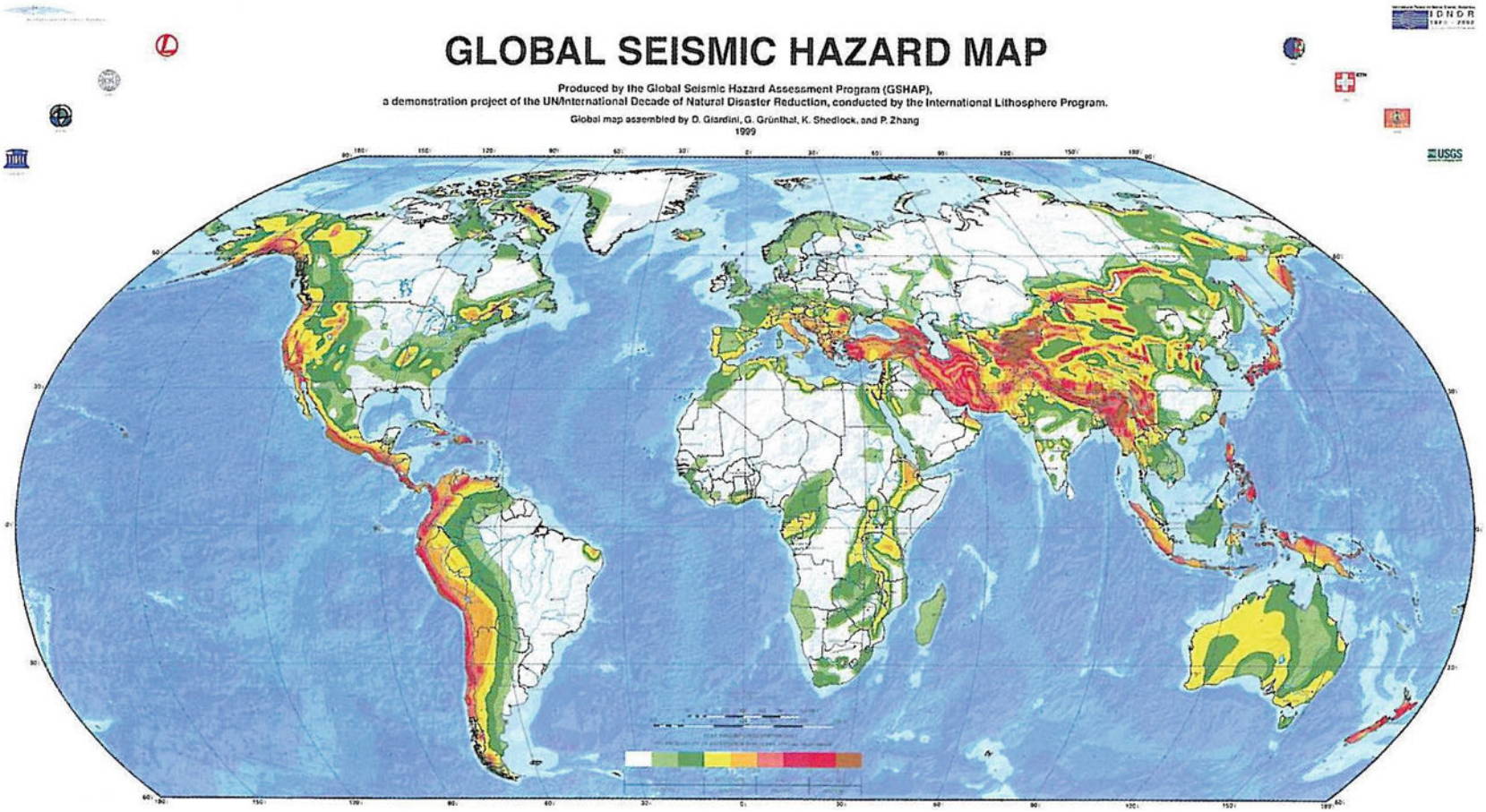
Controlling earthquakes are characterized by mean magnitudes and distances derived from so-called deaggregation analysis. During the deaggregation procedure, the results of PSHA are separated to determine the dominant magnitudes and the distances that contribute to the hazard curve at a specified (reference) probability. Controlling earthquakes are calculated for different structural frequency vibrations, typically for the fundamental frequency of a structure. In the process of deaggregation, the hazard for a reference probability of exceedance of specified ground motion is portioned into magnitude and distance bins. The relative contribution to the hazard for each bin is calculated. The bins with the largest relative contribution identify those earthquakes that contribute the most to the total seismic hazard.

Some modifications of Cornell–McGuire PSHA procedure and alternative models

Source-free PSHA procedures

The concept of seismic sources is the core element of the Cornell–McGuire PSHA procedure. Unfortunately, seismic sources or specific faults can often not be identified and mapped and the causes of seismicity are not understood. In these cases, the delineation of seismic sources is highly subjective and is a matter of expert opinion. In addition, often, seismicity within the seismic sources is not distributed uniformly, as it is required by the classic Cornell–McGuire procedure. The difficulties experienced in dealing with seismic sources have stimulated the development of an alternative technique to PSHA, which is free from delineation of seismic sources.

One of the first attempts to develop an alternative to the Cornell–McGuire procedure was made by Veneziano et al. (1984). Indeed, the procedure does not require the specification of seismic sources, is non-parametric, and, as input, requires only information about past seismicity. The empirical distribution of the specified seismic hazard parameter is calculated by using the observed earthquake magnitudes, epicentral distances, and assumed ground motion prediction equation. By normalizing this



Seismic Hazard, Figure 3 Example of product of PSHA. Map of seismic hazard of the world. Peak ground acceleration expected at 10% probability of exceedance at least once in 50 years. (From Giardini, 1999, http://www.gfz-potsdam.de/pb5/pb53/projects/gshap).

distribution for the duration of the seismic event catalogue, one obtains an annual rate of the exceedance for the required hazard parameter.

Another non-parametric PSHA procedure has been developed by Woo (1996). The procedure is also source-free, where seismicity distributions are approximated by data-based kernel functions. Comparison of the classic Cornell–McGuire-based and kernel-based procedures shows that the former yields a lower hazard.

By their nature, the non-parametric procedures work well in areas with a frequent occurrence of strong seismic events and where the record of past seismicity is considerably complete. At the same time, the non-parametric approach has significant weak points. Its primary disadvantage is a poor reliability in estimating small probabilities for areas of low seismicity. The procedure is not recommended for an area where the seismic event catalogues are highly incomplete. In addition, in its present form, the procedure is not capable of making use of any additional geophysical or geological information to supplement the pure seismological data. Therefore, a technique that accommodates the incompleteness of the seismic event catalogues and, at the same time, does not require the specification of seismic sources, would be an ideal tool for analyzing and assessing seismic hazard.

Such a technique, which can be classified as a *parametric-historic* procedure for PSHA has been successfully used in several parts of the world. Kijko (2008) used it for mapping seismic hazard of South Africa and sub-Saharan Africa. The procedure has been applied in selected parts of the world by the Global Seismic Hazard Assessment Program (GSHAP, Giardini, 1999), while Petersen et al. (2008) applied it for mapping the seismic hazard in the USA. In a series of papers, Frankel and his colleagues modified and substantially extended the original procedure. Their final approach is parametric and based on the assumption that earthquakes within a specified grid size are Poissonian in time, and that the earthquake magnitudes follow the Gutenberg–Richter relation truncated from the top by maximum possible earthquake magnitude $m_{\max}$.

In some cases, the frequency-magnitude Gutenberg–Richter relation is extended by characteristic events. The procedure accepts the contribution of seismicity from active faults and compensates for incompleteness of seismic event catalogues. Frankel's conceptually simple and intuitive parametric-historic approach combines the best of the deductive and non-parametric-historic procedures and, in many cases, is free from the disadvantages characteristic of each of the procedures. The rigorous mathematical foundations of the parametric-historic PSHA procedure have been given by Kijko and Graham (1999).

Alternative earthquake recurrence models

Time-dependent models. In addition to the classic assumption, that earthquake occurrence in time follows a Poisson process, alternative approaches are occasionally used. These procedures attempt to assess temporal, or temporal and spatial dependence of seismicity. Time-dependent earthquake occurrence models specify a distribution of the time to the next earthquake, where this distribution depends on the magnitude of the most recent earthquake. In order to incorporate the memory of past events, the non-Poissonian distributions or Markov chains are applied. In this approach, the seismogenic zones that recently produced strong earthquakes become less hazardous than those that did not rupture in recent history.

Clearly such models may result in a more realistic PSHA, but most of them are still only research tools and have not yet reached the level of development required by routine engineering applications.

Time-dependent occurrence of large earthquakes on segments of active faults is extensively discussed by Rhoades et al. (1994) and Ogata (1999) Also, a comprehensive review of all aspects of non-Poissonian models is provided by Kramer (1996). There are several time-dependent models which play an important role in PSHA. The best known models, which have both firm physical and empirical bases, are probably the models by Shimazaki and Nakata (1980). Based on the correlation of seismic activity with earthquake-related coastal uplift in Japan, Shimazaki and Nakata (1980) proposed two models of earthquake occurrence: a *time-predictable* and a *slip-predictable*.

The time-predictable model states that earthquakes occur when accumulated stress on a fault reaches a critical level; however, the stress drop and magnitudes of the subsequent earthquakes vary among seismic cycles. Thus, assuming a constant fault-slip rate, the time to the next earthquake can be estimated from the slip of the previous earthquake. The second, the slip-predictable model, is based on the assumption that, irrespective of the initial stress on the fault, an earthquake occurrence always causes a reduction in stress to the same level. Thus, the fault-slip in the next earthquake can be estimated from the time since the previous earthquake.

The second group of time-dependent models are less tightly based on the physical considerations of earthquake occurrence and attempt to describe intervals between the consecutive events by specified statistical distributions. Ogata (1999) considers five models: log-normal, gamma, Weibull, doubly exponential and exponential, which results in the stationary Poisson process. After application of these models to several paleo-earthquake data sets, he concluded that no one of the distributions is consistently the best fit; the quality of the fit strongly depends on the data. From several attempts to describe earthquake time intervals between consecutive events using statistical distributions, at least two play a significant role in the current practice of PSHA: the log-normal and the Brownian passage time (BPT) renewal model.

The use of a log-normal model is justified by the discovery that normalized intervals between the consecutive large earthquakes in the circum-Pacific region follow a log-normal distribution with an almost constant standard

deviation (Nishenko and Buland, 1987). The finite value for the intrinsic standard deviation is important because it controls the degree of aperiodicity in the occurrence of *characteristic earthquakes,* making accurate earthquake prediction impossible. Since this discovery, the log-normal model has become a key component of most time-dependent PSHA procedures and is routinely used by the Working Group on California Earthquake Probabilities (WGCEP, 1995).

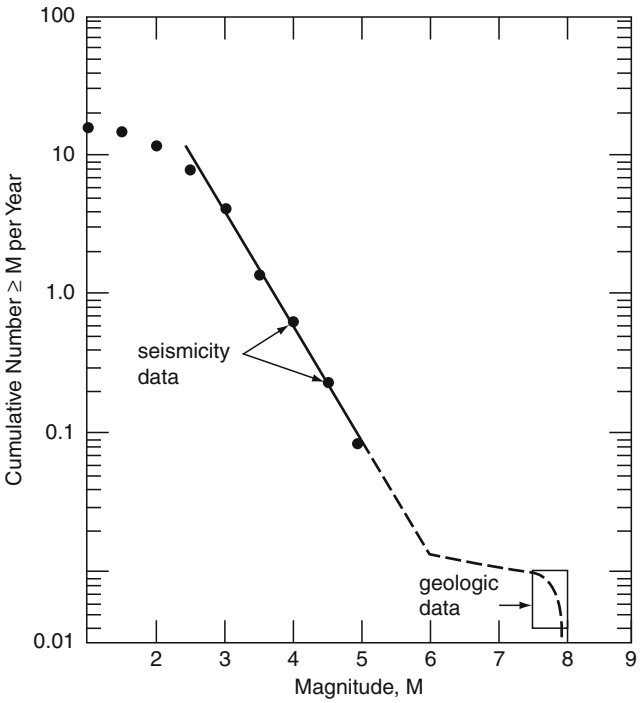
A time-dependent earthquake occurrence model which is applied more often is the Brownian passage time (BPT) distribution, also known as the inverse Gaussian distribution (Matthews et al., 2002). The model is described by two parameters: μ and σ, which, respectively, represent the mean time interval between the consecutive earthquakes and the standard deviation. The aperiodicity of earthquake occurrence is controlled by the variation coefficient $\alpha = \sigma/\mu$. For a small α, the aperiodicity of earthquake occurrence is small and the shape of distribution is almost symmetrical. For a large α, the shape of distribution is similar to log-normal model, i.e., skewed to the right and peaked at a smaller value than the mean. The straightforward control of aperiodicity of earthquake occurrence, by parameter α, makes the BPT model very attractive. It has been used to model earthquake occurrence in many parts of the world and has been applied by the Working Group on California Earthquake Probabilities (WGCEP, 1995).

Comparison of time-dependent with time-independent earthquake occurrence models have shown that the time-independent (Poissonian) model can be used for most engineering computations of PSHA. The exception to this rule is when the seismic hazard is dominated by a single seismic source, with a significant component of characteristic occurrence when the time interval from the last earthquake exceeds the mean time interval between consecutive events. Note that, in most cases, the information on strong seismic events provided by current databases is insufficient to distinguish between different models. The use of non-Poissonian models will therefore only be justified if more data will be available.

Alternative frequency-magnitude models. In the classic Cornell–McGuire procedure for PSHA assessment, it is assumed that earthquake magnitudes follows the Gutenberg–Richter relation truncated from the top by a seismic source characteristic, the maximum possible earthquake magnitude m_{max}. The PDF of this distribution is given by Equation 5.

Despite the fact that in many cases the Gutenberg–Richter relation describes magnitude distributions within seismic source zones sufficiently well, there are some instances where it does not apply. In many places, especially for areas of seismic belts and large faults, the Gutenberg–Richter relation underestimates the occurrence of large magnitudes. The continuity of the distribution (Equation 5) breaks down. The distribution is adequate only for small events up to magnitude 6.0–7.0. Larger events tend to occur within a relatively narrow range of magnitudes (7.5–8.0), but with a frequency higher than that predicted by the Gutenberg–Richter relation. These events are known as *characteristic earthquakes* (Youngs and Coppersmith, 1985, Figure 4). Often it is assumed that characteristic events follow a truncated Gaussian magnitude distribution (WGCEP, 1995).

There are several alternative frequency-magnitude relations that are used in PSHA. The best known is probably the relation by Merz and Cornell (1973), which accounts for a possible curvature in the log-frequency-magnitude relation (1) by the inclusion of a quadratic term of magnitude. Departure from linearity of the distribution (Equation 1) is built into the model by Lomnitz-Adler and Lomnitz (1979). The model is based on simple physical considerations of strain accumulation and release at plate boundaries. Despite the fact that m_{max} is not present in the model, it provides estimates of the occurrence of large events which are more realistic than those predicted by the Gutenberg–Richter relation (1). When seismic hazard is caused by induced seismicity, an alternative distribution to the Gutenberg–Richter model (1) is always required. For example, the magnitude distributions of tremors generated by mining activity are multimodal and change their shape in time (Gibowicz and Kijko, 1994).



Seismic Hazard, Figure 4 Gutenberg–Richter characteristic earthquake magnitude distribution. The model combines frequency-magnitude Gutenberg–Richter relation with a uniform distribution of characteristic earthquakes. The model predicts higher rates of exceedance at magnitudes near the characteristic earthquake magnitude. (After Youngs and Coppersmith, 1985).

Often, the only possible method that can lead to a successful PSHA for mining areas is the replacement of the analytical, parametric frequency-magnitude distribution by its model-free, non-parametric counterpart (Kijko et al., 2001).

Two more modifications of the recurrence models are regularly introduced: one when earthquake magnitudes are uncertain and the other when the seismic occurrence process is composed of temporal trends, cycles, short-term oscillations, and pure random fluctuations. The effect of error in earthquake magnitude determination (especially significant for historic events) can be minimized by the simple procedure of correction of the earthquake magnitudes in a catalogue (e.g., Rhoades, 1996). The modelling of random fluctuations in earthquake occurrence is often done by introducing compound distributions in which parameters of earthquake recurrence models are treated as random variables (Campbell, 1982).

Ground motion prediction equations

The assessment of seismic hazard at a site requires knowledge of the prediction equation of the particular strong motion parameter, as a function of distance, earthquake magnitude, faulting mechanism, and often the local site condition below the site. The most simple and most commonly used form of a prediction equation is

$$\ln(y) = c_1 - c_2 m - c_3 \ln(r) - c_4 r + c_5 F + c_6 S + \varepsilon, \tag{19}$$

where y is the amplitude of the ground motion parameter (PGA, MM intensity, seismic record duration, spectral acceleration, etc.); m is the earthquake magnitude; r is the shortest earthquake distance from the site to the earthquake source; F is responsible for the faulting mechanism; S is a term describing the site effect; and $\epsilon$is the random error with zero mean and standard deviation $\sigma_{\ln(y)}$, which has two components: epistemic and aleatory.

The coefficients $c_1, \ldots, c_6$ are estimated by the least squares or maximum likelihood procedure, using strong motion data. It has been found that the coefficients depend on the tectonic settings of the site. They are different for sites within stable continental regions, active tectonic regions, or subduction zone environments. Assuming that $\ln(y)$ has a normal distribution, regression of (Equation 19) provides the mean value of $\ln(y)$, the exponent of which corresponds to the median value of y, $\breve{y}$. Since the log-normal distribution is positively skewed, the mean value of y, $\bar{y}$, exceeds the median value $\breve{y}$ by a factor of $\exp(-0.5\sigma^2_{\ln(y)})$. This indicates that the seismic hazard for a particular site is higher when expressed in terms of $\bar{y}$, than the hazard for the same site expressed in terms of $\breve{y}$. It has been shown that the ground motion prediction equation remains a particularly important component of PSHA since its uncertainty is a major contributor to uncertainty of the PSHA results (SSHAC, 1997).

Uncertainties in PSHA

Contemporary PSHA distinguishes between two types of uncertainties: aleatory and epistemic.

The *aleatory uncertainty* is due to randomness in nature; it is the probabilistic uncertainty inherent in any random phenomenon. It represents unique details of any earthquake as its source, path, and site and cannot be quantified before the earthquake occurrence and cannot be reduced by current theories, acquiring addition data or information. It is sometimes referred as "randomness," "stochastic uncertainty," or "inherent variability" (SSHAC, 1997) and is denoted as U_R (McGuire, 2004). The typical examples of aleatory uncertainties are: the number of future earthquakes in a specified area; parameters of future earthquakes such as origin times, epicenter coordinates, depths and their magnitudes; size of the fault rupture; associated stress drop and ground motion parameters like PGA, displacement or seismic record duration at the given site. The aleatory uncertainties are characteristic to the current model and cannot be reduced by the incorporation of addition data. It can only be reduced by the conceptualization of a better model.

The *epistemic uncertainty*, denoted as U_K is the uncertainty due to insufficient knowledge about the model or its parameters. The model (in the broad sense of its meaning; as, e.g., a particular statistical distribution) may be approximate and inexact, and therefore predicts values that differ from the observed values by a fixed, but unknown, amount. If uncertainties are associated with numerical values of the parameters, they are also epistemic by nature. Epistemic uncertainty can be reduced by incorporating additional information or data. Epistemic distributions of a model's parameters can be updated using the Bayes' theorem. When new information about parameters is significant and accurate, these epistemic distributions of parameters become delta functions about the exact numerical values of the parameters. In such a case, no epistemic uncertainty about the numerical values of the parameters exists and the only remaining uncertainty in the problem is aleatory uncertainty.

In the past, epistemic uncertainty has been known as statistical or professional uncertainty. The examples of the epistemic uncertainties are: boundaries of seismic sources, distributions of seismic sources parameters (e.g., annual rate of seismic activity λ, b-value and m_{max}), or median value of the ground motion parameter given the source properties.

Aleatory uncertainties are included in the PSHA by means of integration (Equation 5) and they are represented by the hazard curve. In contrast, epistemic uncertainties are included through the use of an alternative hypothesis – different sets of parameters with different numerical values, different models, or through a *logic tree*. Therefore, by default, if in the process of PSHA, the logic tree formalism is applied, the resulting uncertainties of the hazard curve are of epistemic nature.

The major benefit of the separation of uncertainties into aleatory and epistemic is potential guidance in the preparation of input for PSHA and the interpretation of the results. Unfortunately, the division of uncertainties into aleatory and epistemic is model dependent and to a large extent arbitrary, indefinite, and confusing (Panel of Seismic Hazard Evaluation, 1997).

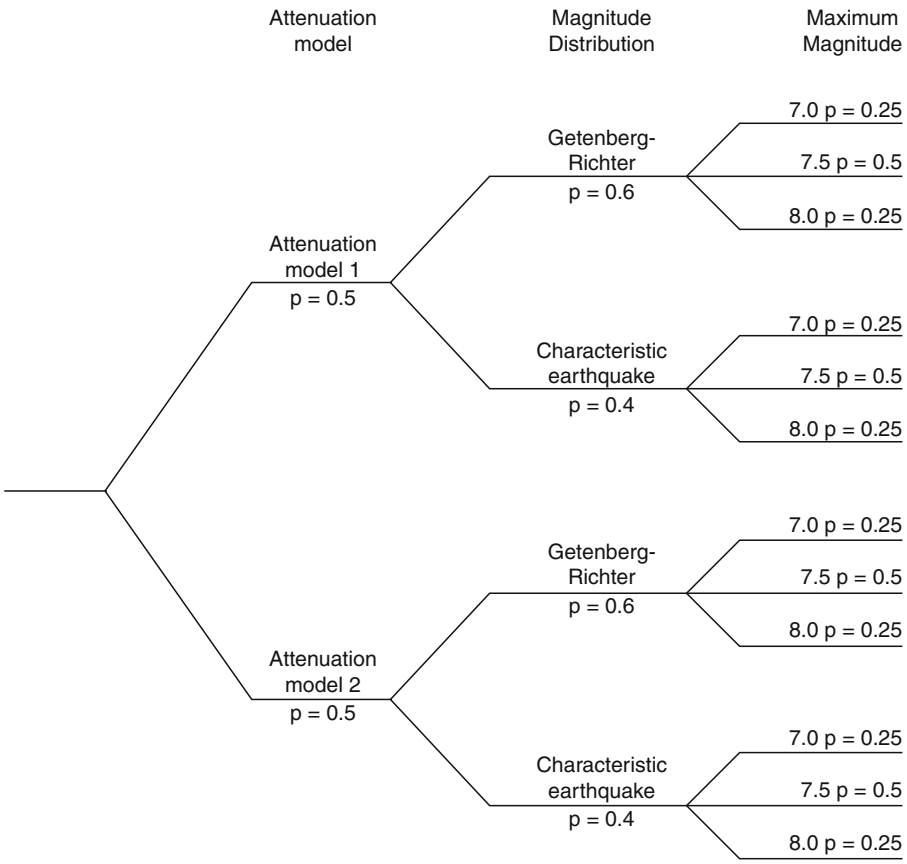
Logic tree

The mathematical formalism of PSHA computation, (Equations 7 and 9), integrates over all random (aleatory) uncertainties of a particular seismic hazard model. In many cases, however, because of our lack of understanding of the mechanism that controls earthquake generation and wave propagation processes, the best choices for elements of the seismic hazard model is not clear. The uncertainty may originate from the choice of alternative seismic sources, competitive earthquake recurrence models and their parameters, as well as from the choice of the most appropriate ground motion. The standard approach for the explicit treatment of alternative hypotheses, models, and parameters is the use of a *logic tree*. The logic tree formalism provides a convenient tool for quantitative treatment of any alternatives. Each node of the logic tree (Figure 5) represents uncertain assumptions, models, or parameters, and the branches extending from each node are the discrete uncertainty alternatives.

In the logic tree analysis, each branch is weighted according to its probability of being correct. As a result, each end branch represents a hazard curve with an assigned weight, where the sum of weights of all the hazard curves is equal to 1. The derived hazard curves are thus used to compute the final (e.g., mean) hazard curve and their confidence intervals. An example of a logic tree is shown in Figure 5. The alternative hypotheses account for uncertainty in the ground motion attenuation model, the magnitude distribution model and the assigned maximum magnitude m_{max}.

Controversy

Despite the fact that the PSHA procedure, as we know it in its current form, was formulated almost half of century ago, it is not without controversy. The controversy surrounds questions such as: (1) the absence of the upper limit of ground motion parameters, (2) division of uncertainties between aleatory and epistemic, and (3) methodology itself, especially the application of the logic tree formalism.



Seismic Hazard, Figure 5 An example of a simple logic tree. The alternative hypothesis accounts for uncertainty in ground motion attenuation relation, magnitude distribution model, and the assigned maximum magnitude m_{max}.

In most currently used Cornell–McGuire-based PSHA procedures, the ground motion parameter used to describe the seismic hazard is distributed log-normally. Since the log-normal distribution is unlimited from the top, it results in a nonzero probability of unrealistically high values for the ground motion parameter, e.g., PGA ≈ 20g, obtained originally from a PSHA for a nuclear-waste repository at Yucca Mountain in the USA (Corradini, 2003). The lack of the upper bound of earthquake-generated ground motion in current hazard assessment procedures has been identified as the "missing piece" of the PSHA procedure (Bommer et al., 2004).

Another criticism of the current PSHA procedure concerns portioning of uncertainties into aleatory and epistemic. As noted in the section (Uncertainties in PSHA) above, the division between aleatory and epistemic uncertainty remains an open issue.

A different criticism comes from the ergodic assumptions which underlie the formalism of the PSHA procedure. The ergodic process is a random process in which the distribution of a random variable in space is the same as distribution of that variable at a single point when sampled as a function of time (Anderson and Brune, 1999). It has been shown that the major contribution to PSHA uncertainty comes from uncertainty of the ground motion prediction equation. The uncertainty of the ground motion parameter y, is characterized by its standard deviation, $\sigma_{\ln(y)}$, which is calculated as the misfit between the observed and predicted ground motions at several seismic stations for a small number of recorded earthquakes.

Thus, $\sigma_{\ln(y)}$ mainly characterizes the spatial and not the temporal uncertainty of ground motion at a single point. This violates the ergodic assumption of the PSHA procedure. According to Anderson and Brune (1999), such violation leads to overestimation of seismic hazard, especially when exposure times are longer than earthquake return times. In addition, Anderson et al. (2000) shows that high-frequency PGAs observed at short distances do not increase as fast as predicted by most ground motion relations. Therefore, the use of the current ground motion prediction equations, especially relating to seismicity recorded at short distances, results in overestimation of the seismic hazard.

A similar view has been expressed by Wang and Zhou (2007) and Wang (2009). Inter alia, they argue that in the Cornell–McGuire-based PSHA procedure, the ground motion variability is not treated correctly. By definition, the ground motion variability is implicitly or explicitly dependent on earthquake magnitude and distance; however, the current PSHA procedure treats it as an independent random variable. The incorrect treatment of ground motion variability results in variability in earthquake magnitudes and distance being counted twice. They conclude that the current PSHA is not consistent with modern earthquake science, is mathematically invalid, can lead to unrealistic hazard estimates, and causes confusion. Similar reservations have been expressed in a series of papers by Klügel (Klügel, 2007 and references therein).

Equally strong criticism of the currently PSHA procedure has been expressed by Castanos and Lomnitz (2002). The main target of their criticism is the logic tree, the key component of the PSHA. They describe the application of the logic tree formalism as a misunderstanding in probability and statistics, since it is fundamentally wrong to admit "expert opinion as evidence on the same level as hard earthquake data."

The science of seismic hazard assessment is thus subject to much debate, especially in the realms where instrumental records of strong earthquakes are missing. At this time, PSHA represents a best-effort approach by our species to quantify an issue where not enough is known to provide definitive results, and by many estimations a great deal more time and measurement will be needed before these issues can be resolved.

Further reading: There are several excellent studies that describe all aspects of the modern PSHA. McGuire (2008) traces the intriguing historical development of PSHA. Hanks and Cornell (1999), and Field (1995) present an entertaining and unconventional summary of the issues related to PSHA, including its misinterpretation. Reiter (1990) comprehensively describes both the deterministic as well as probabilistic seismic hazard procedures from several points of view, including a regulatory perspective. Seismic hazard from the geologist's perspective is described in the book by Yeats et al. (1997). Kramer (1996) provides an elegant, coherent, and understandable description of the mathematical aspects of both DSHA and PSHA. Anderson et al. (2000), Gupta (2002), and Thenhaus and Campbell (2003) present excellent overviews covering theoretical, methodological, as well as procedural issues of modern PSHA. Finally, the most comprehensive treatment to date of all aspects of PSHA, including treatment of *aleatory* and *epistemic* uncertainties, is provided by the SSHAC – Senior Seismic Hazard Committee (1997) – report and in book form by McGuire (2004). The presentations here benefited from all quoted above sources, especially the excellent book by Kramer (1996).

Summary

Seismic hazard is a term referring to any physical phenomena associated with an earthquake (e.g., ground motion, ground failure, liquefaction, and tsunami) and their effects on land, man-made structures, and socioeconomic systems that have the potential to produce a loss. The term is also used, without regard to a loss, to indicate the probable level of ground shaking occurring at a given point within a certain period of time. Seismic hazard analysis is an expression referring to quantification of the expected ground motion at the particular site. Seismic hazard analysis can be performed deterministically, when a particular earthquake scenario is considered, or probabilistically, when the likelihood or frequency of a specified level of ground motion at a site during a specified exposure time is evaluated. In principle, any natural hazard caused by

seismic activity can be described and quantified in terms of the probabilistic methodology. Classic probabilistic seismic hazard analysis (PSHA) includes four steps: (1) identification and parameterization of the seismic sources, (2) specification of temporal and magnitude distributions of earthquake occurrence, (3) calculation of ground motion prediction equations and their uncertainty, and (4) integration of uncertainties in earthquake location, earthquake magnitude, and ground motion prediction equations into the hazard curve.

An integral part of PSHA is the assessment of uncertainties. Contemporary PSHA distinguishes between two types of uncertainties: aleatory and epistemic. The aleatory uncertainty is due to randomness in nature; it is the probabilistic uncertainty inherent in any random phenomenon. The aleatory uncertainties are characteristic to the current model and cannot be reduced by the incorporation of addition data. The epistemic uncertainty is the uncertainty due to insufficient knowledge about the model or its parameters. Epistemic uncertainty can be reduced by incorporating additional information or data. Aleatory uncertainties are included in the probabilistic seismic hazard analysis due to the integration over these uncertainties, and they are represented by the hazard curve. In contrast, epistemic uncertainties are included through the use of alternative models, different sets of parameters with different numerical values or through a logic tree.

Unfortunately, the PSHA procedure, as we know it in its current form, is not without controversy. The controversy arises from questions such as: (1) the absence of the upper limit of ground motion parameter, (2) division of uncertainties between aleatory and epistemic, and (3) methodology itself, especially the application of the logic tree formalism.

Bibliography

Aki, K., 1965. *Maximum Likelihood Estimate of b in the Formula logN = a−bM and its Confidence Limits*. University of Tokyo: Bulletin Earthquake Research Institute, Vol. 43, pp. 237–239.

Anderson, J. G., and Brune, J. N., 1999. Probabilistic seismic hazard analysis without the ergodic assumptions. *Seismological Research Letters*, **70**, 19–28.

Anderson, J. G., Brune, J. N., Anooshehpoor, R., and Shean-Der, Ni., 2000. New ground motion data and concepts in seismic hazard analysis. *Current Science, Special Section: Seismology*, **79**, 1278–1290.

Bender, B., 1984. Incorporation acceleration variability into seismic hazard analysis. *Bulletin Seismological Society of America*, **74**, 1451–1462.

Bommer, J. J., Abrahamson, N. A., Strasser, F. O., Pecker, A., Bard, P. Y., Bugnum, H., Cotton, F., Fäh, D., Sabette, F., Scherbaum, F., and Studer, J., 2004. The challenge of defining upper bounds on earthquake ground motions. *Seismological Research Letters*, **75**, 82–95.

Campbell, K. W., 1982. Bayesian analysis of extreme earthquake occurrences. Part I. Probabilistic hazard model. *Bulletin Seismological Society of America*, **72**, 1689–1705.

Castanos, H., and Lomnitz, C., 2002. PSHA: is it science? opinion paper. *Engineering Geology*, **66**, 315–317.

Cooke, P., 1979. Statistical inference for bounds of random variables. *Biometrika*, **66**, 367–374.

Cornell, C. A., 1968. Engineering seismic risk analysis. *Bulletin Seismological Society of America*, **58**, 1583–1606.

Cornell, C. A., 1971. Bayesian Statistical Decision Theory and Reliability Based Design. In Freudenthal, A. M. (ed.), *Proceedings of the International Conference on Structural Safety and Reliability*, April 9-11, 1969. Smithsonian Institute: Washington DC, pp. 47–66.

Cornell, C. A., 1994. Statistical analysis of maximum magnitudes. In Johnston, A. C., Coppersmith, K. J., Kanter, L. R., and Cornell, C. A. (eds.), *The Earthquakes of Stable Continental Regions*. Palo Alto: Electric Power Research Institute. Assessment of Large Earthquake Potential, Vol. 1, pp. 5–1, 5–27.

Corradini, M. L., 2003. Letter from chairman of the US Nuclear Waste Technical Review Board to the director of the Office of Civilian Radioactive Waste Management; available at: http://www.nwtrb.gov/corr/mlc010.pdf

Cramer, C. H., Petersen, M. D., Cao, T., Toppozada, T. R., and Reichle, M., 2000. A time dependent probabilistic seismic hazard model for California. *Bulletin Seismological Society of America*, **90**, 1–21.

Epstein, B., and Lomnitz, C., 1966. A model for the occurrence of large earthquakes. *Nature*, **211**, 954–956.

Field, D. H., 1995. Probabilistic seismic hazard analysis. A primer. (http://www.relm.org.tutorial_materials).

Frohlich, C., 1998. Does maximum earthquake size depend on focal depth? *Bulletin Seismological Society of America*, **88**, 329–336.

Giardini, D., 1999. The Global Seismic hazard Assessment Program (GSHAP) 1992–1999. *Annali di Geofisica*, **42**, 957–1230.

Gibowicz, S. J., and Kijko, A., 1994. *An Introduction to Mining Seismology*. San Diego: Academic.

Gupta, L. D., 2002. The state of the art in seismic hazard analysis. *ISET Journal of Earthquake Technology*, **39**, 311–346. Paper No. 428.

Gutenberg, B., and Richter, C. F., 1944. Frequency of earthquakes in California. *Bulletin Seismological Society of America*, **34**, 185–188.

Hanks, T. C., and Cornell, C. A., 1999. Probabilistic Seismic Hazard Analysis: A Beginners Guide (available from T.C. Hanks at thanks@usgs.gov).

Jin, A., and Aki, K., 1988. Spatial and temporal correlation between coda Q and seismicity in China. *Bulletin Seismological Society of America*, **78**, 741–769.

Kijko, A., 2004. Estimation of the maximum earthquake magnitude m_{max}. *Pure and Applied Geophysics*, **161**, 1–27.

Kijko, A., 2008. *Data Driven Probabilistic Seismic Hazard Assessment Procedure for Regions with Uncertain Seimogenic Zones*. NATO Monograph, Husebye E. S. (ed.), Earthquake monitoring and seismic hazard mitigation. Seismic hazard analysis and assessment. Springer Science, Chap. 16, pp. 237–251.

Kijko, A., and Graham, G., 1999. Parametric-historic. Procedure for probabilistic seismic hazard analysis. Part II: assessment of seismic hazard at specified site. *Pure and Applied Geophysics*, **154**, 1–22.

Kijko, A., and Sellevoll, M. A., 1992. Estimation of earthquake hazard parameters from incomplete data files, part II, incorporation of magnitude heterogeneity. *Bulletin Seismological Society of America*, **82**, 120–134.

Kijko, A., Lasocki, S., and Graham, G., 2001. Nonparametric seismic hazard analysis in mines. *Pure and Applied Geophysics*, **158**, 1655–1675.

Klügel, J.-U., 2007. Error inflation in probabilistic seismic hazard analysis. *Engineering Geology*, **90**, 186–192.

Kramer, S. L., 1996. *Geotechnical Earthquake Engineering*. Englewood Cliffs: Prentice-Hill.

Lomnitz-Adler, J., and Lomnitz, C., 1979. A modified form of the Gutenberg-Richter magnitude-frequency relation. *Bulletin Seismological Society of America*, **69**, 1209–1214.

Main, I. G., and Burton, P. W., 1984. Information theory and the earthquake frequency-magnitude distribution. *Bulletin Seismological Society of America*, **74**, 1409–1426.
Matthews, M. V., Ellsworth, W. L., and Reasenberg, P. A., 2002. A Brownian model for recurrent earthquakes. *Bulletin Seismological Society of America*, **92**, 2233–2250.
McCalpin, J. P. (ed.), 1996. *Paleoseimology*. New York: Academic.
McGarr, A., 1984. Some applications of seismic source mechanism studies to assessing underground hazard. In Gay, N. C., and Wainwright, E. H. (eds.), *Rockburst and Seismicity in Mines*. Johannesburg: South African Institute of Mining and Metallurgy (Symposium Series No. 6, pp. 199–208).
McGuire, R. K., 2004. *Seismic Hazard and Risk Analysis*. Oakland: Earthquake Engineering Research Institute, MNO-10.
McGuire, R. K., 2008. Review. Probabilistic seismic hazard analysis: early history. *Earthquake Engineering and Structural Dynamics*, **37**, 329–338.
Merz, H. A., and Cornell, C. A., 1973. Seismic risk based on quadratic magnitude frequency low. *Bulletin Seismological Society of America*, **69**, 1209–1214.
Muir-Wood, R., 1993. From global seismotectonics to global seismic hazard. *Annali di Geofisica*, **36**, 153–168.
Nishenko, S. P., and Buland, R., 1987. A generic recurrence interval distribution for earthquake forecasting. *Bulletin Seismological Society of America*, **77**, 1382–1399.
Nuttli, O. W., 1981. *On the Problem of Maximum Magnitude of Earthquakes*. U.S. Geological Survey Open-file Report 81:1–13.
Ogata, Y., 1999. Estimating the hazard of rupture using uncertain occurrence times of paleoearthquakes. *Journal of Geophysical Research*, **104**(17), 995–18014.
Panel of Seismic Hazard Analysis, 1997. *Review of Recommendations for Probabilistic Seismic Hazard Analysis: Guidance on Uncertainty and Use of Experts*. Washington DC: National Academy.
Petersen, M. D., Frankel, A. D., Harmsen, S. C., Mueller, C. S., Haller, K. M., Wheeler, R. L., Wesson, R. L., Zeng, Y., Boyd, O. S., Perkins, D. M., Luco, N., Field, E. H., Wills, C. J., and Rukstales, K. S., 2008. Documentation for the 2008 update of the United States national seismic hazard maps. *U.S. Geological Survey Open-File Report 2008–1128*, 61 p.
Pisarenko, V. F., Lyubushin, A. A., Lysenko, V. B., and Golubieva, T. V., 1996. Statistical estimation of seismic hazard parameters: maximum possible magnitude and related parameters. *Bulletin Seismological Society of America*, **86**, 691–700.
Reiter, L., 1990. *Earthquake Hazard Analysis: Issues and Insights*. New York: Columbia University Press.
Rhoades, D. A., 1996. Estimation of the Gutenberg-Richter relation allowing for individual earthquake magnitude uncertainties. *Tectonophysics*, **258**, 71–83.
Rhoades, D., Van Dissen, R. J., and Dowrick, D. J., 1994. On the handling of uncertainties in estimating the hazard rupture on a fault. *Journal of Geophysical Research*, **99**, 13,701–13,712.
Shimazaki, K., and Nakata, T., 1980. Time-Predictable recurrence model for large earthquakes. *Geophysical Research Letters*, **7**, 279–282.
SSHAC - Senior Seismic Hazard Committee, 1997. *Recommendations for Probabilistic Seismic Hazard Analysis: Guidance on Uncertainty and Use of Experts*. NUREG/CR-6372, UCR-ID-122160, Main Report 1. Prepared for Lawrence Livermore National Laboratory.
Thenhaus, P. C., and Campbell, K. W., 2003. Seismic hazard analysis. In Chen, W. F., and Scawthorn, C. (eds.), *Earthquake Engineering Handbook*. Boca Raton: CRC Press, pp. 8-1–8-50.
Utsu, T., 1965. A method for determining the value of *b* on the formula log $n = a - bM$ showing the magnitude–frequency relation for earthquakes. *Geophysical Bulletin Hokkaido University*, **13**, 99–103. In Japan: Engl. abstr.
Veneziano, D., Cornell, C. A., and O'Hara, T., 1984. *Historic Method for Seismic Hazard Analysis*. Electric Power Research Institute, Report, NP-3438, Palo Alto.
Wang, Z., 2009. Comment on "Sigma: issues, insights, and challenges" by F. O. Strasser, N. A. Abrahamson, and J. J. Bommer. *Seismological Research Letters*, **80**, 491–493.
Wang, Z., and Zhou, M., 2007. Comment on "why do modern probabilistic seismic-hazard analyses often lead to increased hazard estimates?" by Julian J. Bommer, and Norman A. Abrahamson. *Bulletin Seismological Society of America*, **97**, 2212–2214.
Ward, S. N., 1997. More on M_{max}. *Bulletin Seismological Society of America*, **87**, 1199–1208.
Weichert, D. H., 1980. Estimation of the earthquake recurrence parameters for unequal observation periods for different magnitudes. *Bulletin Seismological Society of America*, **70**, 1337–1346.
Wells, D. L., and Coppersmith, K. J., 1994. New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement. *Bulletin Seismological Society of America*, **84**, 974–1002.
WGCEP (Working Group on Central California Earthquake Probabilities), 1995. Seismic hazard in southern California: probable earthquakes, 1994 to 2024. *Bulletin Seismological Society of America*, **85**, 379–439.
Wheeler, R. L., 2009. *Methods of Mmax estimation East of Rocky mountains*. USGS, Open-File Report 2009–1018.
Woo, G., 1996. Kernel estimation methods for seismic hazard area source modeling. *Bulletin Seismological Society of America*, **86**, 353–362.
Yeats, R. S., Sieh, K., and Allen, C. R., 1997. *The Geology of Earthquakes*. New York: Oxford University Press.
Youngs, R. R., and Coppersmith, K. J., 1985. Implications of fault slip rates and earthquake recurrence models to probabilistic seismic hazard estimates. *Bulletin Seismological Society of America*, **75**, 939–964.

Cross-references

Characteristic Earthquakes and Seismic Gaps
Earthquake, Magnitude
Earthquakes, Early and Strong Motion Warning
Earthquakes, Intensity
Earthquakes, Shake Map
Earthquakes, Strong-Ground Motion
Seismic Zonation
Statistical Seismology

SEISMIC IMAGING, OVERVIEW

Gerard T. Schuster
Division of Environmental and Earth Sciences, King Abdullah University of Science and Technology, Thule, Saudi Arabia

Definition

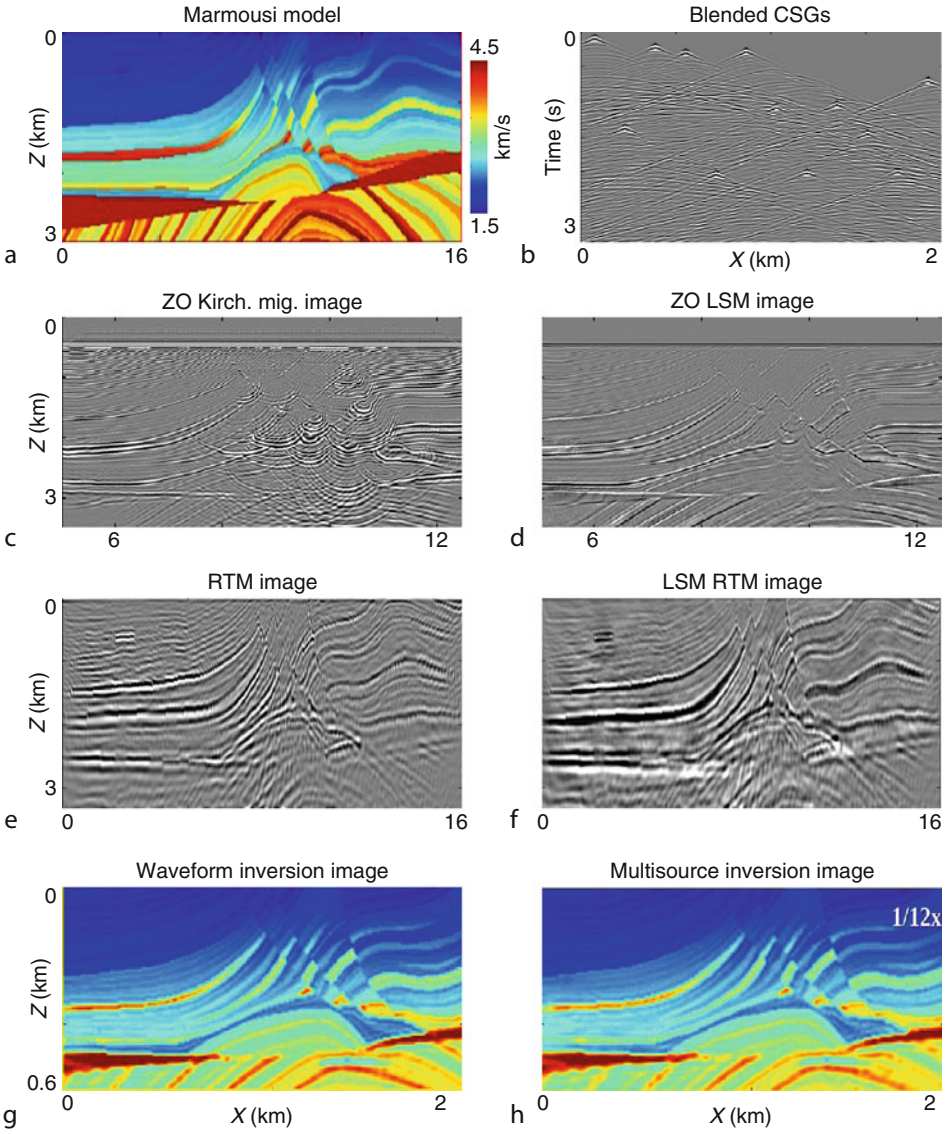
Seismic imaging or tomography (tomo = slice and graph = picture) is a procedure for estimating the earth's rock parameters from seismic data. These rock parameters can be represented by the spatial distribution of, e.g., P-wave velocity, S-wave velocity, porosity, density, or anisotropic parameters. The result of inversion is graphically presented as a 2-D or 3-D grid of pixels, where each pixel

contains the value of the model parameter of interest, which is P velocity in Figure 1g–h. Such tomograms are used to estimate the geometry and lithology of geologic layers, and can help exploration geophysicists and earthquake seismologists understand the evolution of the earth's interior.

There are five main types of seismic imaging: seismic migration, least squares migration, full waveform inversion (FWI), phase-like inversion, and migration velocity analysis. Four of these methods can be derived as special cases of finding the optimal model that minimizes a waveform or a phase-related misfit function. This entry will present these methods as they are applied to body wave events in seismic data, and overview the current best practices in waveform tomography.

Inversion theory

There are four steps to inverting for the model **m** from seismic data **d**. We will assume that the input data are either seismic traces, generated by man-made or earthquake sources, or some skeletonized part of the traces, such as first arrival traveltimes or the phase of an event at some frequency.



Seismic Imaging, Overview, Figure 1 (**a**) Marmousi velocity model, (**b**) time-shifted and blended shot gathers, (**c**) zero-offset Kirchhoff migration image, (**d**) zero-offset least squares migration image after 30 iterations (courtesy of Naoshi Aoki), (**e**) reverse time migration (RTM) image, (**f**) least squares RTM image (Dai and Schuster, 2009), (**g**) waveform tomogram after 50 iterations using the prestack gathers, and (**h**) waveform tomogram where the inputs are eight supergathers, each with 12 phase-encoded shot gathers (Zhan and Schuster, in press); images are at different scales from one another.

Step 1: $\mathbf{L}(\mathbf{m}) = \mathbf{d}$. Establish the mathematical relationship $\mathbf{L}(\mathbf{m}) = \mathbf{d}$ between the seismic data $\mathbf{d}$ and model $\mathbf{m}$.

$$\mathbf{d} = \mathbf{L}(\mathbf{m}), \tag{1}$$

where $\mathbf{L}$ represents the forward modeling operator for the actual model. Equations A1–A3 in Table 1 show the three steps in establishing a linearized version of Equation 1 for the acoustic wave equation, where the extensions to the elastic (Mora, 1987) and anisotropic (Barnes et al., 2008; Operto et al., 2009) wave equations are tedious but straightforward.

Step 2: Discretize $\mathbf{m}$, $\mathbf{d}$, and $\mathbf{L}$. Discretize the 3D earth model into a 3D grid of N physical parameters (e.g., unknown slowness in each cell) and assemble the unknowns into the $N \times 1$ vector $\mathbf{m}$. Discretize the seismic traces in both space and time (or frequency) into an $M \times 1$ vector $\mathbf{d}$ of data measurements. In this case, $\mathbf{L}$ reduces to a $M \times N$ matrix. The forward modeling for FWI requires knowledge of the source wavelet, which can be estimated in a variety of ways: stacking of the direct arrival at different near-offset hydrophones in marine data, iterative inversion of the source wavelet (Mora, 1987; Zhou et al., 1997; Pratt, 1999), or deconvolution of the shot gather with a time-windowed near-offset trace (Sheng, personal communication). The time window can be a few periods long centered about the direct arrival.

Step 3. Linearize $\mathbf{L}\delta\mathbf{m} \approx \delta\mathbf{d}$. Linearize the nonlinear relationship between the data and model. Expanding the ith data measurement $d_i(\mathbf{m})$ to first order in $\delta\mathbf{m}$ by a Taylor series about a first-guess model $\mathbf{m}_o$ (close to the true model) gives the linearized estimate:

$$\begin{aligned} d_i(\mathbf{m}) &\approx d_i(\mathbf{m}_o) + \sum_j \frac{\delta d_i(\mathbf{m}_o)}{\delta m_j}\,\delta m_j \rightarrow \delta d_i(\mathbf{m}) \\ &= \sum_j \overbrace{\left[\partial d_i(\mathbf{m}_o)/\partial m_j\right]}^{\text{wavepath function}}\,\delta m_j, \end{aligned} \tag{2}$$

or in matrix-vector notation

$$\delta\mathbf{d} = \mathbf{L}\delta\mathbf{m}. \tag{3}$$

Here, $\partial d_i(\mathbf{m}_o)/\partial m_j = \lim_{\Delta m_j \rightarrow 0}[d_i(\mathbf{m}_o + \hat{\mathbf{j}}\Delta m_j) - d_i(\mathbf{m}_o)]/\Delta m_j$ is the Fréchet derivative with respect to the jth parameter; the data residual $\delta d_i = [d_i(\mathbf{m}) - d_i(\mathbf{m}_o)]$ is the difference between the ith components of the predicted data vector $\mathbf{d}(\mathbf{m}_o)$ and the observed data vector $\mathbf{d}(\mathbf{m})$; and $\hat{\mathbf{j}}$ is the jth unit vector in the finite-dimensional model space. The model perturbation $\delta\mathbf{m} = \mathbf{m} - \mathbf{m}_o$ is the difference between the actual model $\mathbf{m}$ and the guessed model $\mathbf{m}_o$, and $\mathbf{L}$ is now interpreted as the Jacobian matrix. Its elements $[\mathbf{L}]_{ij} = \partial d_i(\mathbf{m}_o)/\partial m_j$ determine the sensitivity of the data to the perturbations in the model. For a windowed arrival, the Jacobian plots out in model space as a wavepath for a single source-receiver pair (Woodward, 1992; Luo, 1992; Dahlen et al., 2002; Marquering et al., 2002; Montelli et al., 2004; van der Hilst and de Hoop, 2006; Xu and Xie, 2009); and the velocity variations within its first Fresnel zone mostly influence the event of interest. Equation A3 shows that the linearized equations take the form of the Born approximation (Stolt and Benson, 1986) to the

Seismic Imaging, Overview, Table 1 Symbols and specific formulas for inverting the Helmholtz equation for the perturbed slowness distribution $\delta s(\mathbf{x})$, where $\delta s(\mathbf{x})$ is assumed to be small compared to the background slowness model $s(\mathbf{x})$. The harmonic pressure field is denoted by $p(\mathbf{x}|\mathbf{s})$ for a source localized about $\mathbf{s}$, and the Green's function for the Helmholtz equation is given by $G(\mathbf{x}|\mathbf{s})$ for a point source at $\mathbf{s}$ and observer at $\mathbf{x}$. The source wavelet spectrum is $W(\omega)$ at the angular frequency ω and the body force term localized about the point $\mathbf{s}$ is denoted by $f(\mathbf{x}|\mathbf{s})$. The predicted (observed) traveltime of an event received at $\mathbf{g}$ for a source at $\mathbf{s}$ is denoted as $\tau(\mathbf{g}|\mathbf{s})$ ($\tau(\mathbf{g}|\mathbf{s})^{\text{obs.}}$); and $\varepsilon = ||\mathbf{d} - \mathbf{d}^{\text{obs.}}||^2$ is the misfit function for waveform inversion

Symbol	Mathematical formula
(A1) Helmholtz equation: $\mathbf{A}(\mathbf{d},\mathbf{m}) = \mathbf{f}$	$(\nabla^2 + \omega^2 s(\mathbf{x})^2)p(\mathbf{x}\|\mathbf{s}) = f(\mathbf{x}\|\mathbf{s})$ $\mathbf{m} \rightarrow s(\mathbf{x}); \mathbf{d} \rightarrow p(\mathbf{x}\|\mathbf{s})$
(A2) Linearized Helmholtz equation: $\delta\mathbf{A}(\mathbf{d},\mathbf{m}) = 0$	$\left(\nabla^2 + \omega^2 s(\mathbf{x})^2\right)\delta p(\mathbf{x}\|\mathbf{s}) = -2\omega^2 s(\mathbf{x})\delta s(\mathbf{x}) p(\mathbf{x}\|\mathbf{s})$ $\delta\mathbf{m} \rightarrow \delta s(\mathbf{x}); \delta\mathbf{d} \rightarrow \delta p(\mathbf{x}\|\mathbf{s})$
(A3) Lippmann–Schwinger equation with Born approx.: $\delta\mathbf{d} = \mathbf{L}\delta\mathbf{m}$	$\delta p(\mathbf{g}\|\mathbf{s}) = -2\omega^2 \int G(\mathbf{g}\|\mathbf{x}) s(\mathbf{x}) \delta s(\mathbf{x}) W(\omega) G(\mathbf{x}\|\mathbf{s}) \mathrm{d}x^3$ where $p(\mathbf{x}\|\mathbf{s}) = W(\omega)G(\mathbf{x}\|\mathbf{s})$
(A4) Jacobian or wavepath function: $\delta\mathbf{d}/\delta\mathbf{m}$	$\delta p(\mathbf{g}\|\mathbf{s})/\delta s(\mathbf{x}) = -2s(\mathbf{x})\omega^2 W(\omega) G(\mathbf{g}\|\mathbf{x}) G(\mathbf{x}\|\mathbf{s})$
(A5) Misfit gradient or reverse time migration or $\mathbf{L}^\dagger[\mathbf{L}\delta\mathbf{m} - \delta\mathbf{d}]$: $\delta\varepsilon/\delta\mathbf{m}$	$\delta\varepsilon/\delta s(\mathbf{x}) = \gamma \int G(\mathbf{g}\|\mathbf{x})^* \Delta d(\mathbf{g}\|\mathbf{s}) G(\mathbf{x}\|\mathbf{s})^* \mathrm{d}g\mathrm{d}s$ where $\gamma = -2\omega^2 s(\mathbf{x}) W(\omega)$ and $\Delta d(\mathbf{g}\|\mathbf{s}) = p(\mathbf{g}\|\mathbf{s}) - p(\mathbf{g}\|\mathbf{s})^{\text{obs.}}$
(A6) Least squares migration or linearized inversion: $\mathbf{m}^{(k)} - \delta\varepsilon/\delta\mathbf{m}$	$m(\mathbf{x})^{(k)} - \gamma \int G(\mathbf{g}\|\mathbf{x})^* \Delta d(\mathbf{g}\|\mathbf{s})^{(k)} G(\mathbf{x}\|\mathbf{s})^* \mathrm{d}g\mathrm{d}s$
(A7) Waveform inversion: $\mathbf{m}^{(k)} - \delta\varepsilon/\delta\mathbf{m}$	$m(\mathbf{x})^{(k)} - \gamma \int G^{(k)}(\mathbf{g}\|\mathbf{x})^* \Delta d(\mathbf{g}\|\mathbf{s})^{(k)} G^{(k)}(\mathbf{x}\|\mathbf{s})^* \mathrm{d}g\mathrm{d}s$
(A8) Wave equation traveltime inversion: $\mathbf{m}^{(k)} - \delta\varepsilon/\delta\mathbf{m}$	$m(\mathbf{x})^{(k)} - \gamma \int G^{(k)}(\mathbf{g}\|\mathbf{x})^* \Delta d(\mathbf{g}\|\mathbf{s})^{(k)} G^{(k)}(\mathbf{x}\|\mathbf{s})^* \mathrm{d}g\mathrm{d}s$ where $\Delta d(\mathbf{g}\|\mathbf{s}) \approx p(\mathbf{g}\|\mathbf{s})^{\text{obs.}} \left(\tau(\mathbf{g}\|\mathbf{s}) - \tau(\mathbf{g}\|\mathbf{s})^{\text{obs.}}\right)$

Lippmann–Schwinger equation, and its kernel is the Jacobian given by Equation A4.

For a single source and a single receiver in a smoothly increasing velocity medium, Equation A4 plots out as a curved "fat" ray (trace a ray that connects the source and receiver and honors Snell's law. Surround this ray with a "fat finite-frequency ray" such that the propagation of events from the source to receiver within the "fat" ray differs in traveltime by no more than 1/2 the period of the source wavelet. This fat ray region is that portion of the earth which mostly influences the traveltime of the event of interest) that connects the source and receiver points. This fat ray is denoted as a wavepath by Woodward (1992) and forms the basis of finite-frequency traveltime tomography. There seems to be a general (Montelli et al., 2004), but not a universal (van der Hilst and de Hoop, 2006), agreement that finite-frequency tomography can be superior to that of ray-based tomography. As an example, earthquake seismologists use these wavepaths (renamed as banana-doughnuts) with finite-frequency tomography to invert earthquake traveltimes and surface wave data for deep mantle velocity variations attributed to, for example, ascending plumes of rocks (Dahlen et al., 2002; Marquering et al., 2002; Montelli et al., 2006).

Step 4: Solve $\mathbf{L}\delta\mathbf{m} = \delta\mathbf{d}$ by an iterative gradient method. Equation 3 is typically an overdetermined, inconsistent, and poorly conditioned system of equations. Therefore, the solution we seek is the one that minimizes the sum of the data misfit $||\mathbf{Lm} - \mathbf{d}||_K^2$ and model penalty $\lambda^2||\mathbf{Cm}||_W^2$ functions in the K and W norms (Clinthorne et al., 1993):

$$\varepsilon = \frac{1}{2}||\mathbf{L}\delta\mathbf{m} - \delta\mathbf{d}||_K^2 + \frac{\lambda^2}{2}||\mathbf{C}\delta\mathbf{m}||_W^2. \tag{4}$$

where $||\mathbf{L}\delta\mathbf{m} - \delta\mathbf{d}||_K^2 = (\mathbf{L}\delta\mathbf{m} - \delta\mathbf{d})^\dagger\mathbf{K}(\mathbf{L}\delta\mathbf{m} - \delta\mathbf{d})$ and $\mathbf{K}$ is sometimes taken as the inverse of the data covariance matrix (Tarantola, 1987) that is real and symmetric. The penalty function (also known as the regularization term) is defined as $||\delta\mathbf{Cm}||_W^2 = (\mathbf{Cm})^\dagger\mathbf{W}(\mathbf{Cm})$, where the real symmetric matrix $\mathbf{W}$ might strongly weight certain regions in the model because they mostly influence the data. The matrix $\mathbf{C}$ might be a spatial second-derivative operator in one or several directions so as to encourage solutions with smoothly varying model parameters along selected directions.

A simplified steepest descent solution for the ith model parameter m_i is given by

$$\begin{aligned} m_i^{(k+1)} &= m_i^{(k)} - \alpha \overbrace{\frac{\delta\varepsilon}{\delta m_i}}^{\text{gradient Equation A5}} \rightarrow \mathbf{m}^{(k+1)} \\ &= \mathbf{m}^{(k)} - \overbrace{\alpha\mathbf{L}^\dagger\delta\mathbf{d}^{(k)}}^{\text{migration of data residual}}, \end{aligned} \tag{5}$$

where $\lambda = 0$; $\mathbf{K} = \mathbf{I}$; α is the step length; and preconditioning is used (Beydoun and Mendes, 1989; Clinthorne et al., 1993; Causse et al., 1995) to accelerate convergence. A sequence of models is generated until the data residual falls below some acceptable level. In practice, a preconditioned conjugate gradient method (Mora, 1987; Luo and Schuster, 1991; Epanomeritakis et al., 2008), a Gauss-Newton Krylov solver (Erlangga and Hermann, 2009), or a limited memory quasi-Newton method (Pratt et al., 1998; Plessix, 2006) for different implementations of the adjoint method (Plessix, 2009) is implemented; and sometimes a direct matrix solver is used to find the Hessian inverse if the problem size is small enough (Pratt and Goulty, 1991). Approximations to the Hessian by a systematic procedure can be found in Thierry et al. (1999). There is strong evidence (Crase et al., 1990; Brossier et al., 2010) that using the L_1 norm misfit function is noticeably more resistant to data noise than the L_2 norm misfit function, and the combination of the two is sometimes the best choice.

Five types of seismic imaging methods

Five types of seismic imaging methods and their resolution properties will be discussed (there are many seismic inversion methods, but the five types discussed here are often used in the geophysical community): migration (modest resolution of reflectivity), least squares migration (high resolution of reflectivity), full waveform inversion (high resolution of velocity), phase-like inversion (modest resolution of velocity), and migration velocity analysis (modest resolution of velocity). The high-resolution methods typically pay the price of increased computational cost and decreased robustness, compared to the moderate-resolution methods with relatively low cost and desirable robustness. For reflection migration and inversion, the spatial resolution limits are approximately defined by a generalized Radon transform analysis (Beylkin et al., 1985); and the resolution limits for ray-based transmission tomography (Williamson, 1991) can be estimated by considering the width of the transmission Fresnel zone.

Migration is the first iterate solution of Equation 5, and least squares migration, sometimes known as linearized inversion, is the final iterative solution where the operator $\mathbf{L}$ is not updated after each iteration; also, $\mathbf{m}(\mathbf{x})$ represents the reflectivity model at the position $\mathbf{x}$. In contrast to migration, the waveform inversion tomogram is the final iterative solution, where the velocity model $\mathbf{m}$ and $\mathbf{L}$ are updated after every iteration; waveform inversion falls under the class of nonlinear optimization methods. It can be shown under certain assumptions that waveform inversion reduces to either wave equation traveltime tomography (Luo and Schuster, 1991) or ray-based tomography. These last two methods are classified as phase-like inversion methods. Unlike minimizing the data misfit function in Equation 4, migration velocity analysis updates the velocity model to minimize a *model misfit function*, which is the normed difference between the predicted migration image and the actual migration image

in the, e.g., common image gather. Convergence problems associated with local minima in the misfit function are reduced by emphasizing *flatness* in the CIG misfit function (Symes and Carazone, 1991; Shen et al., 1993).

Migration

If the starting model is a smoothly varying velocity distribution that only generates the accurate first arrival, then the input data $\mathbf{d} - \mathbf{d}^{(0)}$ residual becomes the scattered data. In this case the desired model is the reflectivity distribution, which is similar to the slowness perturbation function $\delta s(\mathbf{x})$, and the background velocity model is the inverse of slowness $s(\mathbf{x})$. If a finite-difference solution to the wave equation, for example, is used to calculate $G(\mathbf{x}|\mathbf{s})$ and $G(\mathbf{g}|\mathbf{x})$ in Equation A5, then the first model update $\delta\mathbf{m}^{(1)}$ in Equation 5 is known as the reverse time migration image (Whitmore, 1983; McMechan, 1983). If a one-way wave equation method is used to generate the Green's function, then the migration method is a phase-shift or FX-type algorithm (Stolt and Benson, 1986; Claerbout, 1992; Etgen et al., 2009). A diffraction-stack migration method results if the Green's function is replaced by its asymptotic approximation $G(\mathbf{x}|\mathbf{s}) = A(\mathbf{x},\mathbf{s})e^{i\omega\tau_{xs}}$ (here, $A(\mathbf{x}, \mathbf{s})$ accounts for geometric spreading losses, and τ_{xs} is the first arrival time for a ray that connects the source at $\mathbf{s}$ with the observer at $\mathbf{x}$. A ray-tracing method can be used to compute these traveltimes for a sufficiently high frequency and smoothly varying medium [Bleistein et al., 2001]). An example of standard poststack migration is depicted in Figure 1c, where the velocity model is shown in Figure 1a. The migration image depicts the reflectivity distribution computed by a Kirchhoff migration method with an eikonal traveltime solver.

Seismic migration is also used to migrate teleseismic body waves processed to form so-called receiver functions. Examples include Ryberg and Weber (2000), Sheehan et al. (2000), and also Bostock et al. (2001) who used a ray-Born inversion approach. More recently, Nowack et al. (2007) applied the Gaussian beam migration approach of Hill (2001) to the migration of teleseismic body waves. Despite its widespread use and its robustness, migration is considered to be a moderate-resolution method because it approximates the inverse Hessian matrix $[\mathbf{L}^{\dagger}\mathbf{L}]^{-1}$ by a diagonal matrix.

Least squares migration

If the background model $1/s(\mathbf{x})$ is not updated after each iteration (i.e., $\mathbf{L}$ in Equation 5 is independent of the k index), then $\mathbf{m}^{(k+1)}$ for large k is known as the least squares migration image (Nemeth et al., 1999; Duquet et al., 2000). As in standard migration, the model to be iteratively updated is the reflectivity distribution and not the velocity model. The iterative least squares migration Equation A6 is interpreted as a sequence of standard migrations, where the data residual is backprojected into the earth model by the migration operator $\mathbf{L}^{\dagger}$. Least squares migration (LSM) is also known as linearized waveform inversion (Lailly, 1984; Tarantola, 1986, 1987; Jin et al., 1992; Lambaré et al., 1992) and is superior to standard migration by reducing migration artifacts caused by a poor acquisition geometry; it also can provide a spatial resolution that is more than twice (Nemeth et al., 1999; Yu et al., 2006) that of standard migration if the migration velocity is sufficiently accurate. Its main drawbacks are that its effectiveness is very sensitive to the accuracy of the migration velocity model, and it can be more than an order of magnitude more expensive than standard migration. (Recent developments [Dai and Schuster, 2009] in phase-encoded migration suggest a great reduction in the cost of LSM.) As an example, Figure 1d depicts the LSM image obtained from zero-offset (ZO) data, which is more accurate than the ZO standard migration image in Figure 1c. Figure 1e–f depict the RTM and LSM RTM images obtained from the prestack shot gathers.

Full waveform inversion

If the background model is updated after each iteration, then Equation A7 is known as nonlinear waveform inversion (Tarantola, 1987; Mora, 1987; Mora, 1989; Song et al., 1995; and many others); a common designation for waveform inversion is FWI or full waveform inversion. Unlike LSM or standard migration, Equation A7 for FWI iteratively updates the velocity and reflectivity models so that the final velocity image can be much improved in both resolution and accuracy. Its main drawbacks are a tendency to get stuck in local minima, it is computationally expensive compared to standard migration, and there might be more than one model that can explain the same data, i.e., a nonunique solution. For example, elastic isotropic modeling codes can sometimes generate predicted traces that adequately fit the observed data, but the estimated isotropic velocity model is inconsistent with the actual anisotropic rocks in the real earth.

Successful examples of waveform tomography images are shown in Figure 1g–h and were obtained from the same shot gathers used for migration. The final tomogram shows the velocity distribution that is almost identical to that of the actual model. One of the keys to success in waveform inversion is that a good starting model is often required for an accurate result. This starting model can be obtained by migration velocity analysis (Stork, 1992; Jousselin et al., 2009), reflection traveltime tomography (Langan et al., 1985; Bishop et al., 1985; Nolet, 1987; Zelt, 2002), or refraction tomography (Pratt and Goulty, 1991; Luo and Schuster, 1991; Min and Shin, 2006; Sheng et al., 2006).

A major challenge to the success of waveform inversion is the limited offset range between sources and receivers and the lack of low-frequency information in the recorded data (Sirgue and Pratt, 2004; Barnes et al., 2008; Boonyasiriwat et al., 2009; Kelly et al., 2009). Such deficiencies can prevent waveform inversion from reconstructing the low wave-number parts of the model, and sometimes prevent convergence to any type of

reasonable model. Remedies being explored include the possibility of recording data at much lower frequencies with more capable recording devices, and obtaining very wide-offset data. Other challenges address the validity of the acoustic approximation versus the reality that the recorded data are largely of elastic (Vigh et al., 2009) or viscoelastic nature (Causse et al., 1999; Askan, 2006). Elastic inversion of elastic seismograms (Mora, 1987; Mora, 1989; Zhou et al., 1997; Brossier et al., 2009) have been presented, but the acoustic approximation can still provide useful results. One approach to viscoelastic inversion is to invert for the acoustic velocity first, then follow this with inversion for attenuation parameters (Kamei and Pratt, 2008); another approach is to use a ray-based method to invert for the attenuation factor Q and then use this Q to correct for the attenuation in the data (Pratt et al., 2005; Sheng et al., 2007; Boonyasiriwat et al., 2009).

Wave equation traveltime inversion

If the data residual $\delta d(\mathbf{g}|\mathbf{s})$ is replaced by the traveltime residual $\delta\tau(\mathbf{g}|\mathbf{s})$ weighted by the recorded trace $d(\mathbf{g}|\mathbf{s})^{\text{obs.}}$, then this is known as wave equation traveltime (Luo and Schuster, 1991) tomography (WT); it is a variant of Rytov inversion (Woodward, 1992) and updates the velocity model by smearing weighted traveltime (not waveform) residuals over the associated wavepaths (Woodward, 1992; Luo, 1992). In the high-frequency limit, it reduces to ray-based traveltime tomography (RT). The advantage of WT over RT is that it does not require a high-frequency approximation and accounts for the band-limited nature of waves as they propagate through the earth. Shin et al. (2002, 2007) use a modified logarithm norm to invert for phase data, which bears a close relationship to the Rytov inversion method; and Effelsen (2009) compares phase inversion to traveltime tomography for inverting refraction events. Figure 2 illustrates how the above seismic imaging methods are related to one another.

The main disadvantage of WT is that it is at least an order of magnitude more expensive than RT because it requires a finite-difference solution of the wave equation rather than a ray-traced approximation. Earthquake seismologists often see wave equation traveltime as a major improvement in estimating the earth's velocity parameters from earthquake records (Dahlen et al., 2002; Marquering et al., 2002; Montelli et al., 2006; van der Hilst and de Hoop, 2006).

Migration velocity analysis

The previous imaging methods can be described as estimating the earth model by predicting synthetic data that *best matches the observed data in the data domain*. In contrast, migration velocity analysis (MVA) finds the velocity model that *best flattens or focuses the migration sections in the image domain*; here, the image domain is the migration cube in the (x, y, z) indices and some other index s such as shot index, receiver index, source-receiver offset, (Yilmaz and Chambers, 1984; Faye and Jeannot, 1986; Al-Yahya, 1989; Toldi, 1989; Stork, 1992; Lafond and Levander, 1993), common angle parameter (Xu et al., 1998), or subsurface offset coordinate (Rickett and Sava, 2002; Sava and Biondi, 2004; Robein, 2010).

To understand why we seek a velocity model that flattens events in the image domain, consider the 2D migration image $m(x, z, s_0)$ in Figure 3a obtained by migrating one shot gather (the source is at the surface with shot index s_0). If the migration velocity is correct, then all of the single-shot migration images should be similar in appearance. This means that all of the migrated reflection events should be flat in the common image gather (CIG) given by $m(x_{\text{CIG}}, z, s)$ for all values of s, z, and a fixed value of x_{CIG} (see Figure 3b). Note that if the migration velocity is accurate, the reflector boundary denoted by the dotted horizontal line will be flat for a common image gather. If the migration velocity is too slow, then the imaged reflector boundary will curve upward as illustrated by the curved dashed line, and if too fast, the imaged reflector will curve downward.

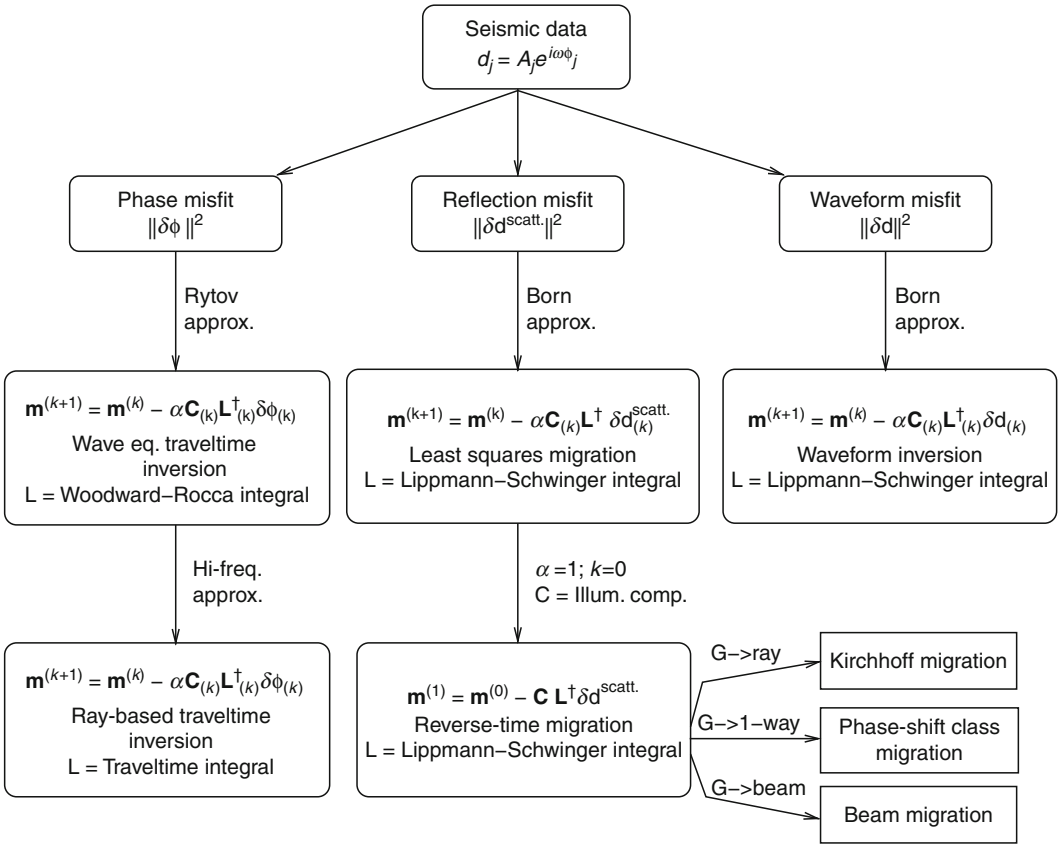
With MVA, the goal is to find the velocity model so that, ideally, the predicted migration image $\mathbf{m}_{\text{mig.}}^{\text{pred.}} = \mathbf{L}^{\dagger}\mathbf{L}\mathbf{M}$ best fits the actual migration image $\mathbf{m}_{\text{mig.}} = \mathbf{L}^{\dagger}\mathbf{d}$. The associated misfit function can be constructed so that it is similar to that of Equation 4, except that the norm of the *migration* residual

$$\varepsilon = \frac{1}{2}||\mathbf{m}_{\text{mig.}}^{\text{pred.}} - \mathbf{m}_{\text{mig.}}||_K^2 + \frac{\lambda^2}{2}||\mathbf{C}\mathbf{m}||_W^2 \tag{6}$$

is minimized rather than the *data* residual; in this case, $\mathbf{m}$ represents the velocity model. To find the velocity model, we can use the unregularized steepest descent equation:

$$\mathbf{m}^{(k+1)} = \mathbf{m}^{(k)} - \overbrace{\alpha\mathbf{L}^{\dagger}\mathbf{L}\delta\mathbf{m}_{\text{mig.}}^{(k)}}^{\text{migration}-\text{modeling of migration residual}}, \tag{7}$$

where $\delta\mathbf{m}_{\text{mig.}}^{(k)} = \mathbf{m}_{\text{mig.}}^{\text{pred.}} - \mathbf{m}_{\text{mig.}}$ is the migration residual at the kth iteration. The importance of this formula is that it shows that the gradient term is computed by modeling the migration residual to get the *filtered* data residual $\mathbf{L}\delta\mathbf{m}_{\text{mig.}}^{(k)}$, and then the velocity model is updated by smearing this filtered (the filtered data residual is constructed by migrating and modeling the actual data residual; hence, it is a filtered version of the actual data residual) data residual into the model by the migration operation $\mathbf{L}^{\dagger}[\mathbf{L}\delta\mathbf{m}_{\text{mig.}}^{(k)}]$. These filtered residuals are smeared along wavepaths for each source-receiver pair of traces, and the migration and modeling operators are updated after each iteration. This leads to a moderate resolution of the velocity model because the filtered, not unfiltered, data are kinematically fitted. Moreover, MVA



Seismic Imaging, Overview, Figure 2 Three types of seismic imaging methods: phase inversion, least squares migration (LSM), and nonlinear waveform inversion. Note that LSM is a linearized inversion so that the modeling operator **L** does not get updated after each iteration. This compares to nonlinear waveform inversion which updates **L** after each iteration. The Lippmann–Schwinger-type and Woodward–Roca-type integrals are displayed in Table 1, and details about various migration methods are in the Seismic Migration section in the *Encyclopedia of Solid Earth Geophysics*. The traces that only contain scattered arrivals are obtained by muting the direct arrivals; and the ensemble of these traces is symbolized by the vector $\mathbf{d}^{scatt.}$. Instead of computing the Green's functions in the integral equations by a finite-difference solution to the wave equation, various approximations such as ray-based Green's functions (Kirchhoff migration), one-way wave equation approximations (phase-shift type migration), and Gaussian Beam (beam-like migration) are used for migration.

is largely a curve fitting exercise that is mostly insensitive to subtle amplitude variations in the migration traces; ignoring such information will blind MVA to subtle variations in the impedance distribution.

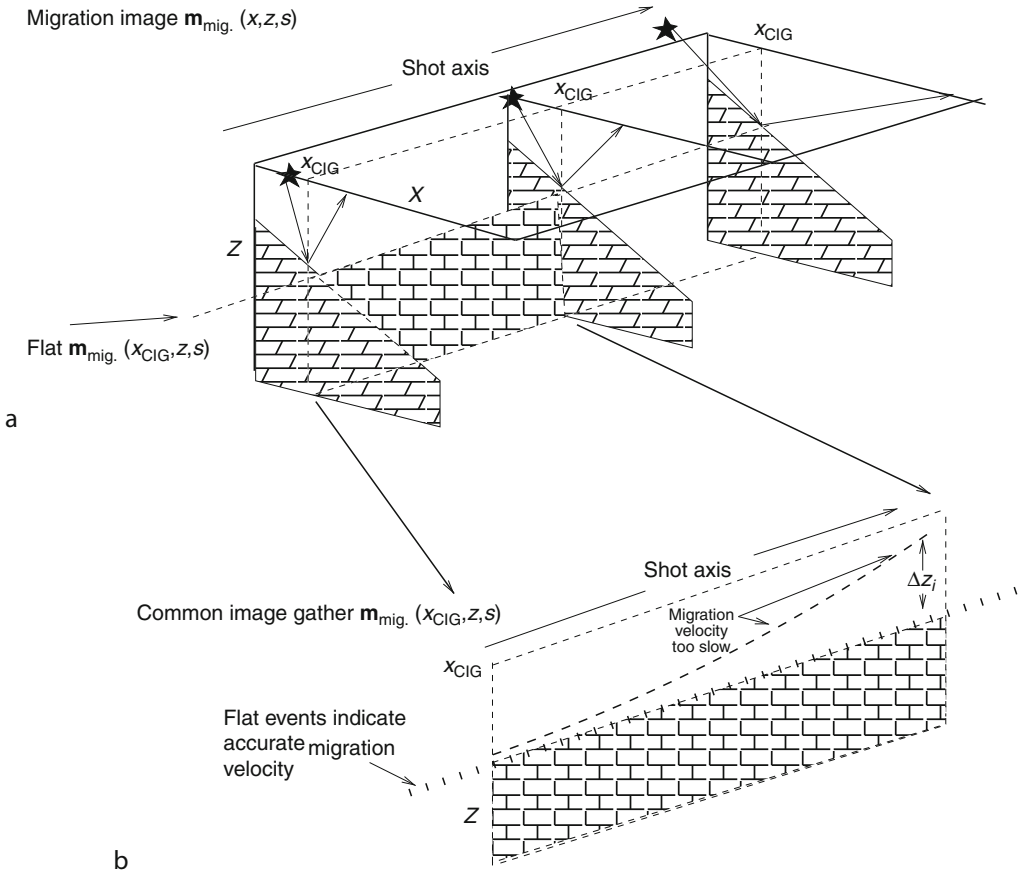
To reduce computational costs and increase robustness at the expense of reduced model resolution, MVA is sometimes implemented with the following steps:

1. Automatically pick the depth residual $\Delta z_i = z_i - z_i^{\text{ref.}}$ of a coherent CIG event in Figure 3b at the ith shot position; here, z_i is the picked depth of the targeted reflection at the x_{CIG} offset in the ith migrated shot gather. The depth $z_i^{\text{ref.}}$ of the reference event for that reflector is estimated from the near-offset trace in that CIG. A computer algorithm can window about the near-offset reflection of interest and use crosscorrelation with neighboring traces in the CIG to estimate the depth lag $z_i - z_i^{\text{ref.}}$ associated with the strongest correlation energy.
2. The misfit function is then defined as

$$\varepsilon = 1/2 \sum_i (\Delta z_i)^2 + \text{model smoothness constraints}, \tag{8}$$

for each CIG and the summation is over the shot index in Figure 3b. Sometimes the misfit function is the sum over all CIGs.
3. The velocity model is iteratively updated by a gradient optimization method until ε is minimized.

An example of the above procedure is shown in Figure 4, where the left column of images depict the (a) migration image obtained with an inaccurate velocity



Seismic Imaging, Overview, Figure 3 (**a**) Migration image $\mathbf{m}_{\text{mig.}}(x, z, s)$ cube in the model coordinates of x and z and the surface shot offset index s; the common image gather $\mathbf{m}_{\text{mig.}}(x_{CIG}, z, s)$ is computed by migrating the shot gathers and displaying the migration image at the fixed value of $x = x_{CIG}$ for all shot s and depth z values. (**b**) Common image gather $\mathbf{m}_{\text{mig.}}(x_{CIG}, z, s)$ for the common image point at x_{CIG}. The migration image should only be non-zero along the interface between the brick region and open regio shown above.

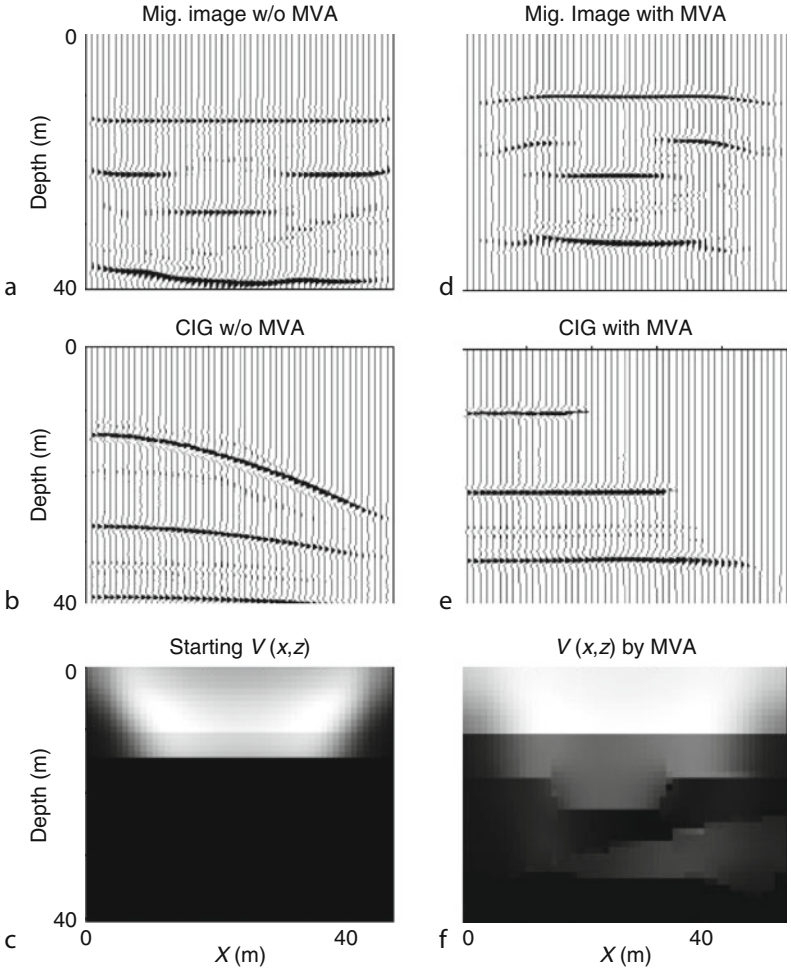
model, (b) the CIG with curved events, and (c) the migration velocity model. After seven iterations of MVA, the right column of figures is obtained. Note the flat events in the CIG, and the final velocity model is almost the same as the actual velocity model.

The Fréchet derivative $\frac{\partial z_i}{\partial c(\mathbf{X})}$ associated with the gradient of ε can be numerically computed by using a ray-tracing scheme (Stork, 1992; Chauris et al., 1998) to determine the change in depth z_i of a migrated reflection point with respect to a change in the velocity parameter at $\mathbf{x}$. Sometimes, the depth residual Δz_i for the ith ray is converted into a time residual $\Delta\tau_i$ and, similar to traveltime tomography, this *converted* time residual is smeared along the reflection ray to iteratively update the velocity model. For arbitrary reflector geometries in a homogeneous media, an analytic conversion formula was derived by Al-Yahya (1989), and a ray-tracing method was used by Stork (1992) and others (Robein, 2010).

Similar to waveform inversion, MVA seeks to predict the wiggly migration traces seen in the migration image $\mathbf{L}^{\dagger}\mathbf{d}$ and, therefore, can easily get stuck in the many local minima of ε. To avoid this problem, Symes and Carazone (1991) proposed the smoother differential semblance misfit function (Chauris et al., 1998; Shen et al., 2003):

$$\varepsilon = 1/2||\sum_{x,y}[\gamma\sum_{z}\sum_{h}\partial m(x,y,z,h)_{\text{mig.}}/\partial h^2]||^2 + \text{constraints}. \tag{9}$$

that rewards flat events (i.e., accurate velocity models) in the CIG domain. Here, γ is a normalization term that depends on (x, y). Empirical tests by Chauris et al. (1998) suggest that the semblance based ε is much smoother than that in Equation 6 and is nearly devoid of local minima in misfit function. Some links between MVA and FWI are established in Symes (2008), and MVA compared to several tomography methods is presented by Le Bégat et al. (2004). The subsurface offset domain can be exploited for MVA (Sava and Biondi, 2004) and extraction of scattering angle information



Seismic Imaging, Overview, Figure 4 (**a**) Migration image (Sun, 2001) obtained by prestack migration using the smooth homogeneous velocity model in (**c**) that is far from the true model approximated in (**f**); (**b**) common image gather in the shot offset index, where the curved events indicate an incorrect velocity model; (**c**) incorrect migration velocity model used to compute (**a**); (**d**) migration image obtained after seven iterations of MVA; (**e**) CIG after seven iterations of MVA; and (**f**) velocity model inverted by seven iterations of MVA. This result closely closely resembles the true velocity model.

(Rickett and Sava, 2002), and an excellent summary of MVA research is given in Robein (2010).

In summary, the exploration community heavily favors MVA over many other velocity estimation methods because it is robust, efficient, and the picking of depth residuals is easily automated in the migration image domain. (Traveltimes picked from traces are notoriously difficult to automate because waveforms often interfere with one another. In comparison, migration untangles these events and focuses them to their origin along the reflectors so that automatic picking is easier in the CIG. Sometimes semblance methods are used to find the best fit hyperbolic or parabolic curve to the data [Robein, 2010].) Its chief disadvantage is that an MVA image lacks the detailed resolution of waveform inversion, which suggests that MVA should be used to estimate the starting velocity models for waveform tomography.

Recent advances in seismic imaging

In the last 15 years, several breakthroughs have enabled practical waveform inversion of seismic data. One of these advances is the relentless increase in computational capabilities of cluster computers and GPU-like processors, and two others are multiscale inversion and phase-encoded waveform inversion.

Multiscale waveform inversion

One of the main difficulties with waveform inversion is that the misfit function is plagued by many local minima. If the starting velocity model is moderately far from the actual model (an erroneous velocity might be one where the modeled events arrive by more than a period after the actual arrivals), then the iterative gradient solution gets stuck there and never reaches the global minimum or actual model.

The partial cure to this local-minima problem is a multiscale approach (Bunks et al., 1995), where the initial iterations estimate a coarse-grid velocity model from low-frequency data. For a reasonable starting model, this often prevents getting stuck in local minima because the predicted low-frequency arrivals are more likely to arrive within a period of the arrivals in the low-pass filtered data. After a number of iterations, the velocity model is refined to a finer scale (the grid interval is halved) and intermediate frequency data are iteratively inverted to update the velocity model. After suitable reduction of the data misfit, the model grid is refined again and higher-frequency data are inverted until a satisfactory model is reconstructed. One of the first relevant demonstrations of multiscale FWI applied to real data in a complex environment was performed by Ravaut et al. (2004). Other results with both synthetic and field data (Sirgue and Pratt, 2004; Sirgue et al., 2007; Plessix, 2009; Vigh and Starr, 2007, 2008; Sirgue et al., 2010) convincingly demonstrate the effectiveness of this approach. As mentioned earlier, improved model reconstructions can be achieved if more accurate starting models are employed, lower frequency data are recorded, and wider offset data are acquired. The new challenges are to employ modeling and inversion codes that robustly take into account the effects of viscoelasticity and anisotropy in the data.

Phase-encoded multisource waveform inversion

A major difficulty, until recently, is the enormous computational expense of waveform inversion. Each shot gather of residuals must be migrated at each iteration, which for 3D problems can be too demanding even for the most powerful computers. To relieve this problem, Krebs et al. (2009), Virieux and Operto (2009), and Dai and Schuster (2009) proposed summing phase-encoded shot gathers into supergathers and migrating the supergathers at each iteration; a supergather is modeled with just one finite-difference simulation, where the computational cost is the same as that for one shot gather. This is similar to the phase-encoding RTM of Romero et al. (2000), except iterations are used to reduce the crosstalk noise in supergather migration. The result can be an enormous cost savings compared to conventional waveform inversion or migration.

Figure 1h shows an example of phase-encoding shot gathers, where each trace in a shot gather has the same random time shift but different shots have different time shifts. In this case, 12 shot gathers were time-shifted and blended together into one supergather; there were 192 traces per shot gather. A total of 16 nonoverlapping supergathers were input into the iterative waveform inversion code, where each supergather migration costs the same as the migration of one shot gather because 12 shots were excited at nearly the same time. Hence, the Figure 1h tomogram costs 1/12 that of the Figure 1g tomogram. For 3D inversion, the computational cost savings can be more than two orders of magnitude compared to conventional waveform inversion.

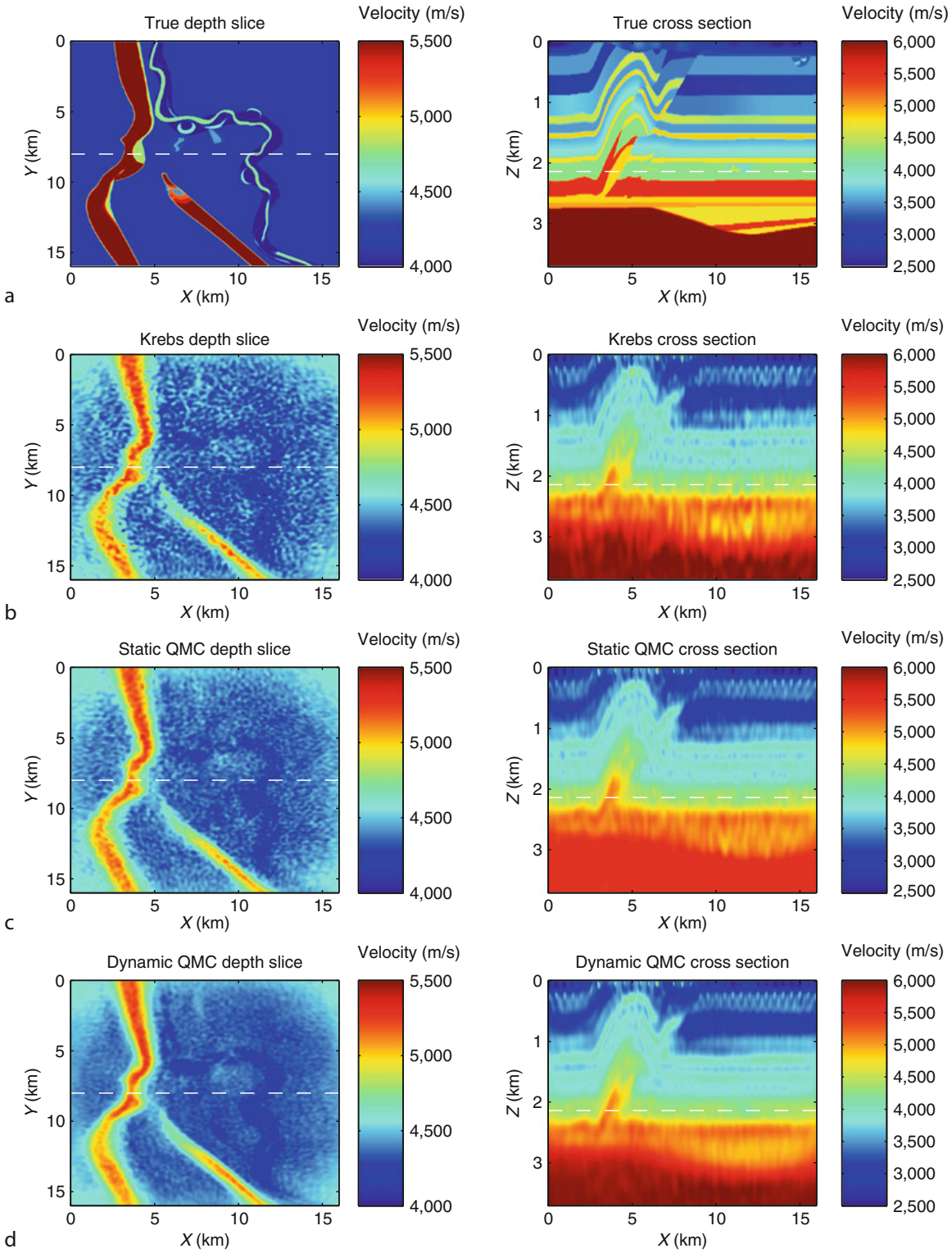
For a 3D example, Figure 5a shows results after waveform inversion of 3D synthetic data, where the model size is 800 × 800 × 186 grid points and the spatial sampling interval is 20 m. There are 1,089 sources evenly distributed along the surface with an interval of 500 m in the inline (X) and crossline (Y) directions. Multisource waveform inversion using the Krebs method (Krebs et al., 2009), static Quasi-Monte Carlo (QMC) phase encoding, and dynamic QMC phase encoding (Boonyasiriwat and Schuster, 2010) are applied to this data set.

Figure 5b–d show that dynamic QMC phase encoding provides a higher-quality tomogram than the other methods, yet the computational cost is two orders of magnitude less than that of conventional waveform inversion. The Krebs strategy is the most efficient because 1,089 shot gathers were blended into one supergather, compared to the QMC strategies that used 99 CSGs/supergather. Dynamic phase encoding changed the time shifts of each shot gather after each iteration, while the static strategy kept the phase encoding the same.

Current status and future of seismic imaging

Reverse time migration is becoming the preferred means for seismic imaging beneath complex regions such as salt bodies. Earthquake seismologists are now recognizing the benefits of migration imaging earthquake records for tectonic structures such as subduction zones (Bostock et al., 2001) or using wide-angle seismic experiments to image the crust (Brenders and Pratt, 2007) and mantle (Montelli et al., 2006). Least squares migration and waveform inversion are now being considered as viable upgrades to standard migration because of the tremendous speedup from phase-encoded multisource methodology. There is still a debate about whether waveform inversion should be computed in the frequency (there are additional advantages by formulating the problem in the Laplace transform domain [Shin and Ha, 2008]) or time domains (Vigh and Starr, 2007; Warner, 2008), but there is no debate that we eventually need to account for viscoelastic and anisotropic effects in the data.

3D waveform inversion of earthquake records for whole earth tomograms greater than 1 Hz is still too computationally demanding except at very low frequencies, and the same can be said for 3D exploration geophysics at frequencies above 50 Hz. Challenges still remain, especially in the critically important area of anisotropic RTM (Zhang and Zhang, 2009; Fowler et al., 2010) and waveform inversion; estimation of accurate anisotropic migration velocity models is an ongoing line of research. Earthquake seismologists are now testing the possibility of using earthquake records for inverting basin structures and velocity models so as to improve their simulation-based predictions of earthquake hazard. Passive seismic recordings and the use of interferometry to extract surface wave records (Shapiro and Campillo, 2004; Shapiro et al., 2005), followed by inversion for S-velocity tomograms are playing an increasingly important role in earthquake seismology.



Seismic Imaging, Overview, Figure 5 Depth slices at $z = 2.1$ km and cross-sections at $y = 8$ km of (**a**) true model, and inverted models using (**b**) the Krebs method (1089 CSGs/supergather), (**c**) static QMC method (99 CSGs/supergather), and (**d**) dynamic QMC method (99 CSGs/supergather) after 40 iterations. A greater number of CSGs/supergather require a greater number of iterations to get the same accuracy in the final tomogram; the dynamic encoding strategy is more effective than the static strategy. (Figures courtesy of C. Boonyasiriwat).

Bibliography

Al-Yahya, K., 1989. Velocity analysis by iterative profile migration. *Geophysics*, **54**, 718–729.

Aoki, N., and Schuster, G. T., 2009. Fast least-squares migration with a deblurring filter. *Geophysics*, **74**, WCA83–WCA93.

Askan, A., 2006. *Full Waveform Inversion for Seismic Velocity and Anelastic Losses in Heterogeneous Structures*. PhD dissertation, Carnegie Mellon University.

Barnes, C., Charara, M., and Tsuchiya, T., 2008. Feasibility study for an anisotropic full waveform inversion of cross-well seismic data. *Geophysical Prospecting*, **56**, 897–906.

Beydoun, W., and Mendes, M., 1989. Elastic ray-born L2-migration/inversion. *Geophysical Journal International*, **97**, 151–160.

Beylkin, G., Oristaglio, M., and Miller, D., 1985. Spatial resolution of migration algorithms. In Berkhout, A. J., Ridder, J., and van der Walls, L. F. (eds.), *Proceeding of the 14th International Symposium on Acoustical Imaging*. pp. 155–167.

Bishop, T., Bube, K. P., Cutler, R. T., Langan, R. T., Love, P. L., Resnick, J. R., Shuey, R. T., Spindler, D. A., and Wyld, H. W., 1985. Tomographic determination of velocity and depth in laterally varying media. *Geophysics*, **50**, 903–923.

Bleistein, N., Cohen, J. K., and Stockwell, J. W., 2001. *Mathematics of multidimensional seismic inversion*. New York: Springer.

Boonyasiriwat, C., and Schuster, G. T., 2010. 3D multisource ful-waveform inversion using dynamic QMC phase encoding. In *Expanded Abstracts of SEG International Meeting*, p. 29.

Boonyasiriwat, C., Valasek, P., Routh, P., Cao, W., Schuster, G. T., and Macy, B., 2009. An efficient multiscale method for time-domain waveform tomography. *Geophysics*, **74**, WCC59–WCC68, doi:10.1190/1.3151869.

Bostock, M. G., Rondenay, S., and Shragge, J., 2001. Multiparameter two-dimensional inversion of scattered teleseismic body waves: Theory for oblique incidence. *Journal of Geophysical Research*, **106**, 771–782.

Brenders, A. J., and Pratt, R. G., 2007. Efficient waveform tomography for lithospheric imaging: Implications for realistic 2D acquisition geometries and low frequency data. *Geophysical Journal International*, **168**, 152–170.

Brossier, R., Operto, S., and Virieux, J., 2009. Seismic imaging of complex onshore structures by 2D elastic frequency-domain full-waveform inversion. *Geophysics*, **74**, 105–118.

Brossier, R., Operto, S., and Virieux, J., 2010. Which data residual norm for robust elastic frequency-domain full waveform inversion? *Geophysics*, **75**, R37–R46.

Bunks, C., Salek, F. M., Zaleski, S., and Chavent, G., 1995. Multiscale seismic waveform inversion. *Geophysics*, **60**, 1457–1473.

Causse, E., Mittet, R., and Ursin, B., 1999. Preconditioning for full-waveform inversion in viscoacoustic media. *Geophysics*, **64**, 130–145.

Chauris, H., Noble, M., and Podvin, P., 1998. Testing the behavior of differential semblance for velocity estimation. In *Expanded Abstracts of SEG International Meeting*, pp. i1305–1308.

Claerbout, J., 1992. *Earth Soundings Analysis: Processing versus Inversion*. Cambridge, MA: Blackwell Scientific.

Clinthorne, N., Pan, T., Chiao, P., Rogers, W., and Stamos, J., 1993. Preconditioning methods for improved convergence rates in iterative reconstructions. *IEEE Transactions on Medical Imaging*, **12**, 78–83.

Crase, E., Pica, A., Noble, M., McDonald, J., and Tarantola, A., 1990. Robust elastic non-linear waveform inversion: Application to real data. *Geophysics*, **55**, 527–538.

Dahlen, F., Hung, S., and Nolet, G., 2002. Frechet kernels for finite-frequency traveltimes I. Theory. *Geophysical Journal International*, **141**, 157–174.

Dai, W., and Schuster, G. T., 2009. Least-squares migration of simultaneous sources data with a deblurring filter. In *Expanded Abstracts of SEG International Meeting*, pp. 2990–2993.

Duquet, B., Marfurt, K., and Dellinger, J., 2000. Kirchhoff modeling, inversion for reflectivity, and subsurface illumination. *Geophysics*, **65**, 1195–1209.

Effelsen, K., 2009. A comparison of phase inversion and traveltime tomography for processing of near-surface refraction traveltimes. *Geophysics*, **74**, WCB11–WCB24.

Epanomeritakis, I., Akçelik, V., Ghattas, O., and Bielak, J., 2008. A Newton-CG method for large-scale three-dimensional elastic full waveform seismic inversion. *Inverse Problems*, **24**, 975–987.

Erlangga, Y., and Hermann, F., 2009. Seismic waveform inversion with Gauss-Newton-Krylov method. In *Expanded Abstracts of SEG International Meeting*, pp. 2357–2361.

Etgen, J., Gray, S., and Zhang, Y., 2009. An overview of depth imaging in exploration geophysics. *Geophysics*, **74**, WCA5–WCA17.

Faye, J.-P., and Jeannot, J.-P., 1986. Prestack migration velocities from focusing depth analysis. In *Expanded Abstracts of SEG International Meeting*, pp. 438–440.

Fowler, P., Du, X., and Fletcher, R., 2010. Coupled equations for reverse time migration in transversely isotropic media. *Geophysics*, **75**, S11–S22.

Hill, N. R., 2001. Prestack Gaussian beam depth migration. *Geophysics*, **66**, 1240–1250.

Jin, S., Madariaga, R., Virieux, J., and Lambaré, G., 1992. Two-dimensional asymptotic iterative elastic inversion. *Geophysical Journal International*, **108**, 575–588.

Jousselin, P., Duquet, B., Audebert, F., and Sirgue, J., 2009. Bridging the gap between ray-based tomography and wave-equation migration image gathers. In *Expanded Abstracts of SEG International Meeting*, pp. 3979–3983.

Kamei, R., and Pratt, G., 2008. Waveform tomography strategies for imaging attenuation structure for cross-hole data. In *70th Conference and Technical Exhibition, EAGE Expanded Abstracts*, p. F019.

Kelly, S., Ramos-Martinez, J., and Tsimelzon, B., 2009. The effect of improved, low-frequency bandwidth in full-wave form inversion for velocity. In *Expanded Abstracts of SEG International Meeting*, pp. 3974–3977.

Krebs, J. R., Anderson, J. E., Hinkley, D., Neelamani, R., Lee, S., Baumstein, A., and Lacasse, M. D., 2009. Fast full-wavefield seismic inversion using encoded sources. *Geophysics*, **74**, WCC177–WCC188.

Lafond, C., and Levander, A., 1993. Migration moveout analysis and depth focusing. *Geophysics*, **58**, 91–100.

Lailly, P., 1984. Migration methods: partial but efficient solutions to the seismic inverse problem. In Santosa, F., Pao, Y. H., Symes, W., and Holland, C. H. (eds.), *Inverse Problems of Acoustic and Elastic Waves*. Philadelphia: SIAM.

Lambaré, Virieux, J., Madariaga, R., and Jin, S., 1992. Iterative asymptotic inversion in the acoustic approximation. *Geophysics*, **57**, 1138–1154.

Langan, R., Lerche, I., and Cutler, R. T., 1985. Tracing of rays through heterogeneous media: An accurate and efficient procedure. *Geophysics*, **50**, 1456–1465.

Le Bégat, S., Chauris, H., Devaux, V., Nguyen, S., and Noble, M., 2004. Velocity model estimation for depth imaging: Comparison of three tomography methods on a 2D real data set. *Geophysical Prospecting*, **52**, 427–438.

Luo, Y., 1992. Calculation of wavepaths for band-limited seismic waves. In *Expanded Abstracts of SEG International Meeting*, pp. 1509–1512.

Luo, Y., and Schuster, G. T., 1991. Wave-equation traveltime inversion. *Geophysics*, **56**, 645–653.

Marquering, H., Dahlen, F. A., and Nolet, G., 2002. Three-dimensional sensitivity kernels for finite-frequency traveltimes: the banana-doughnut paradox. *Geophysical Journal International*, **137**, 805–815.

McMechan, G., 1983. Migration by extrapolation of time-dependent boundary values. *Geophysical Prospecting*, **31**, 413–420.

Min, D., and Shin, C., 2006. Refraction tomography using a waveform-inversion back-propagation technique. *Geophysics*, **71**(3), R21–R30.

Montelli, R., Nolet, G., Masters, G., Dahlen, F., and Hung, S. H., 2004. Global P and PP traveltime tomography: rays versus waves. *Geophysical Journal International*, **158**, 637–654.

Montelli, R., Nolet, G., and Dahlen, F., 2006. Comment on "Banana-doughnut kernels and mantle tomography" by van der Hilst and de Hoop. *Geophysical Journal International*, **167**, 1204–1210.

Mora, P., 1987. *Elastic Wavefield Inversion*. PhD dissertation, Stanford University, pp. 143.

Mora, P., 1989. Inversion=migration+tomography. *Geophysics*, **54**, 1575–1586.

Nemeth, T., Wu, C., and Schuster, G. T., 1999. Least-squares migration of incomplete reflection data. *Geophysics*, **64**, 208–221.

Nolet, G., 1987. *Seismic Tomography: with Applications in Global Seismology and Exploration*. Springer.

Nowack, R. L. W. P., Chen, U. K., and Dasgupta, S., 2007. Imaging offsets in the Moho: synthetic tests using Gaussian beams with teleseismic waves. *Pure and Applied Geophysics*, **164**, 1921–1936.

Operto, S., Virieux, J., Ribodetti, A., and Anderson, J., 2009. Finite-difference frequency-domain modeling of viscoelastic wave propagation in two-dimensional tilted transversely isotropic media. *Geophysics*, **74**, T75–T95.

Plessix, R. E., 2006. A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. *Geophysical Journal International*, **167**, 495–503.

Plessix, R. E., 2009. 3D frequency-domain full-waveform inversion with an iterative solver: Geophysics. *Geophysics*, **74**, WCC149–WCC157.

Pratt, G., 1999. Seismic waveform inversion in the frequency domain, part I: theory and verification in a physical scale model. *Geophysics*, **64**, 888–901.

Pratt, R. G., and Goulty, N. R., 1991. Combining wave-equation imaging with traveltime tomography to form high-resolution images from crosshole data. *Geophysics*, **56**, 208–224.

Pratt, R. G., Shin, C., and Hicks, G. J., 1998. Gauss-Newton and full Newton methods in frequency-space seismic waveform inversion. *Geophysical Journal International*, **133**, 341–362.

Pratt, R. G., Hou, F., Bauer, K., and Weber, M., 2005. Waveform tomography images of velocity and inelastic attenuation from the Mallik 2002 crosshole seismic surveys. In Dallimore, S. R., and Collett, T. S. (eds.), *Scientific Results from the Mallik 2002 Gas Hydrate Production Research Well Program*. Mackenzie Delta, Northwest Territories, Canada: Geological Survey of Canada.

Ravaut, C., Operto, S., Improta, L., Virieux, J., Herrero, A., and dell'Aversana, P., 2004. Multi-scale imaging of complex structures from multi-fold wide-aperture seismic data by frequency-domain full-wavefield inversions: application to a thrust belt. *Geophysical Journal International*, **159**, 1032–1056.

Rickett, J., and Sava, P., 2002. Offset and angle-domain common-image gathers for shot-profile migration. *Geophysics*, **67**, 883–889.

Robein, E., 2010. *Seismic Imaging*. EAGE Publications.

Romero, L., Ghiglia, D., Ober, C., and Morton, S., 2000. Phase encoding of shot records in prestack migration. *Geophysics*, **65**, 426–436.

Ryberg, T., and Weber, M., 2000. Receiver function arrays: a reflection seismic approach:. *Geophysical Journal International*, **141**, 1–11.

Sava, P., and Biondi, B., 2004. Wave-equation migration velocity analysis-1: theory. *Geophysical Prospecting*, **52**, 593–606.

Shapiro, N., and Campillo, M., 2004. Emergence of broadband Rayleigh waves from correlations of the ambient seismic noise. *Geophysical Research Letters*, **31**, L07614, doi:10.1029/2004GL019491.

Shapiro, N., Campillo, M., Stehly, L., and Ritzwoller, M., 2005. High-resolution surface-wave tomography from ambient seismic noise. *Science*, **307**, 1615–1618.

Sheehan, A. F., Shearer, P. M., Gilbert, H. J., and Dueker, K. G., 2000. Seismic migration processing of P-SV converted phases for mantle discontinuity structure beneath the Snake River Plain, Western United States. *Journal of Geophysical Research*, **105**, 055–065.

Shen, P., Symes, W., and Stolk, C., 2003. Differential semblance velocity analysis by wave equation migration. In *Expanded Abstracts of SEG International Meeting*, pp. 2135–2139.

Sheng, J., Leeds, A., Buddensiek, M., and Schuster, G. T., 2006. Early arrival waveform tomography on near-surface refraction data. *Geophysics*, **71**(4), U47–U57.

Shin, C., and Ha, W., 2008. A comparison between the behavior of objective functions for waveform inversion in the frequency and Laplace domains. *Geophysics*, **73**, VE119–VE133.

Shin, C., Min, D.-J., Marfurt, K. J., Lim, H. Y., Yang, D., Cha, Y., Ko, S., Yoon, K., Ha, T., and Hong, S., 2002. Traveltime and amplitude calculations using the damped wave solution. *Geophysics*, **67**, 1637–1647.

Shin, C., Pyun, S., and Bednar, B., 2007. Comparison of waveform inversion, part 1: conventional wavefield vs logarithmic wavefield. *Geophysical Prospecting*, **55**, 449–464.

Sirgue, L., and Pratt, R. G., 2004. Efficient waveform inversion and imaging: A strategy for selecting temporal frequencies. *Geophysics*, **69**, 231–248.

Sirgue, L., Etgen, J., and Albertin, U., 2007. 3D full-waveform inversion: wide-versus narrow-azimuth acquisitions. In *Expanded Abstracts of SEG International Meeting*, pp. 1760–1764.

Sirgue, L., Barkved, O. I., Dellinger, J., Etgen, J., Albertin, U., and Kommedal, J. H., 2010. Full waveform inversion: the next leap forward in imaging at Valhall. *First Break*, **28**, 65–70.

Song, Z., Williamson, P., and Pratt, G., 1995. Frequency-domain acoustic-wave modeling and inversion of crosshole data, Part 2: Inversion method, synthetic experiments and real-data results. *Geophysics*, **60**, 786–809.

Stolt, R., and Benson, A., 1986. Seismic migration: theory and practice. In *Handbook of Geophysical Exploration*. London, UK: Geophysical Press, Vol. 5.

Stork, C., 1992. Reflection tomography in the postmigrated domain. *Geophysics*, **57**, 680–682.

Sun, H., 2001. *Wavepath Migration for Depth Imaging and Velocity Analysis*. PhD dissertation, University of Utah.

Symes, W., 2008. Migration velocity analysis and waveform inversion. *Geophysical Prospecting*, **56**, 765–790.

Symes, W., and Carazone, J., 1991. Velocity inversion by differential semblance optimization. *Geophysics*, **56**, 654–663.

Tarantola, A., 1986. Linearized inversion of seismic reflection data. *Geophysical Prospecting*, **32**, 998–1015.

Tarantola, A., 1987. *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation*. Amsterdam: Elsevier Science.

Thierry, P., Operto, S., and Lambaré, G., 1999. Fast 2D ray-Born inversion/migration in complex media. *Geophysics*, **64**, 162–181.

Toldi, J., 1989. Velocity analysis without picking. *Geophysics*, **54**, 191–199.
van der Hilst, R., and de Hoop, M., 2006. Reply to comment by R. Montelli, G. Nolet, and F.A. Dahlen on "Banana-doughnut kernels and mantle tomography". *Geophysical Journal International*, **167**, 1211–1214.
Vigh, D., and Starr, E. W., 2007. Comparisons for waveform inversion, time domain or frequency domain? In *Expanded Abstracts of SEG International Meeting*, pp. 1890–1894.
Vigh, D., and Starr, E. W., 2008. 3D prestack plane-wave full-waveform inversion. *Geophysics*, **73**, 135–144.
Vigh, D., Starr, E. W., and Elapavuluri, P., 2009. Acoustic waveform inversion vs. elastic data. In *Expanded Abstracts of SEG International Meeting*, pp. 2298–2301.
Virieux, J., and Operto, S., 2009. An overview of full-waveform inversion in exploration geophysics. *Geophysics*, **74**, WCC1–WCC26.
Warner, 2008. 3D wavefield tomography: synthetic and field data examples. In *Expanded Abstracts of SEG International Meeting*, pp. 3330–3334.
Whitmore, N. D., 1983. Iterative depth migration by backward time propagation. In *Expanded Abstracts of SEG International Meeting*, pp. 827–830.
Williamson, P., 1991. A guide to the limits of resolution imposed by scattering in ray tomography. *Geophysics*, **56**, 202–207.
Woodward, M. J., 1992. Wave-equation tomography. *Geophysics*, **57**, 1526.
Xu, W., and Xie, X., 2009. How serious is the nonlinear effect on traveltime delays predicted by sensitivity kernels. In *Expanded Abstracts of SEG International Meeting*, pp. 4049–4053.
Xu, S., Chauris, H., Lambar, G., and Noble, M., 1998. Common angle image gather: a strategy for imaging complex media. In *Expanded Abstracts of SEG International Meeting*, pp. 1538–1541.
Yilmaz, O., and Chambers, R., 1984. Migration velocity analysis by wave-field extrapolation. *Geophysics*, **49**, 1664–1674.
Yu, J., Hu, J., Schuster, G., and Estill, R., 2006. Prestack migration deconvolution. *Geophysics*, **71**, S53, doi:10.1190/1.2187783.
Zelt, C., 2002. Modelling strategies and model assessment for wide-angle seismic traveltime data. *Geophysical Journal International*, **139**, 183–204.
Zhan, G., and Schuster, G. T., (in press). Multisource phase-encoded waveform inversion. *Journal of Exploration Geophysics*.
Zhang, Y., and Zhang, H., 2009. A stable TTI reverse time migration and its implementation. In *Expanded Abstracts of SEG International Meeting*, pp. 2794–2798.
Zhou, C., Schuster, G. T., Hassanzadeh, S., and Harris, J. M., 1997. Elastic wave-equation traveltime and waveform inversion of crosshole seismic data. *Geophysics*, **62**, 853–868.

Cross-references

Body Waves
Free Oscillations of the Earth
Inverse Theory, Global Optimization
Inverse Theory, Linear
Inverse Theory, Monte Carlo Method
Numerical Methods, Finite Difference
Seismic Tomography
Seismic, Ambient Noise Correlation
Seismic, Migration
Seismic, Waveform Modeling and Tomography
Single and Multichannel Seismics
Traveltime Tomography Using Controlled-Source Seismic Data
Vertical Seismic Profiling

SEISMIC INSTRUMENTATION

Duncan Carr Agnew
Institute of Geophysics and Planetary Physics, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA

Synonyms

Seismometry

Definition

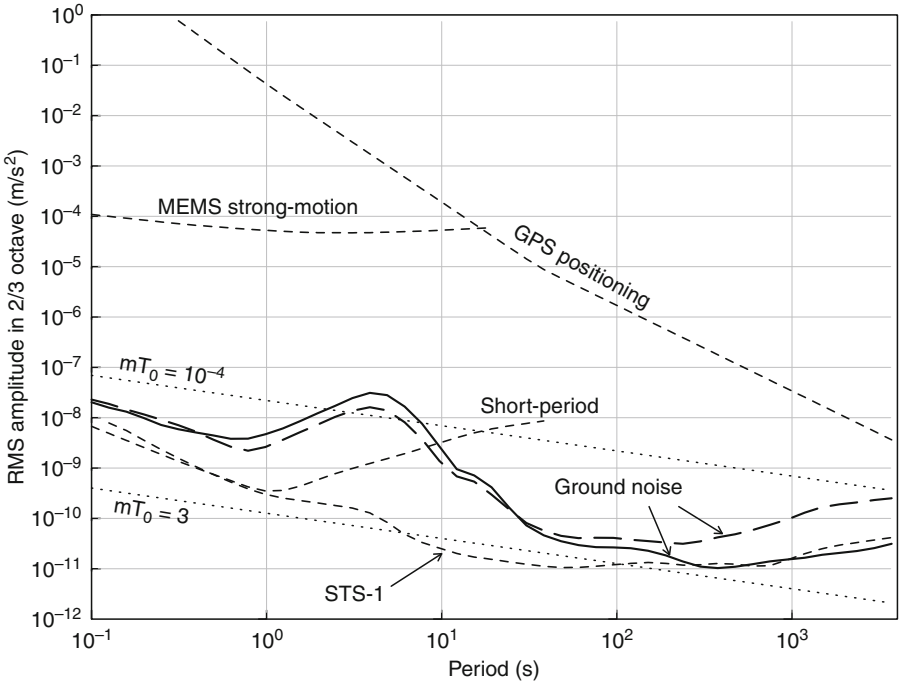
Datalogger. Device for recording electrical signal from a seismometer, usually in digital form and with accurate time information.
Geophone. Another name for seismometer, used in geophysical exploration.
Seismometer. Device for providing a record of ground motion, usually by converting it into an electrical signal.
Strainmeter. Device for continuously measuring the deformation of the Earth, either as extension along a line, or volume change.

Introduction

Seismic instruments give a time record of ground motion caused by elastic waves. The first instruments were built at the end of the nineteenth century; subsequent developments have evolved towards wider coverage of both the frequency and amplitude of the waves that could be recorded. Over most of the twentieth century much of this evolution, and of the diversity of instrument types, was related not to how ground motion was sensed but to how it was recorded; the last 3 decades have seen such rapid development in digital technology, especially in data storage, that recording techniques unique to seismology are no longer needed. Modern seismic systems all consist of a seismometer for converting some aspect of ground motion into an electrical signal, and a datalogger for recording this signal. In exploration geophysics, the name geophone is a common synonym for seismometer; now that data recording is clearly separated from sensing ground motion, the old term seismograph could probably be abandoned. Since dataloggers are not unique to seismology, this article discusses only the general requirements for them, focusing instead on the designs of seismic sensors, or seismometry. Three recent reviews cover seismometry in more detail: Wielandt (2002), Havskov and Alguacil (2004), and Bormann (2009).

Requirements for instruments

The frequencies to be measured by seismometers range from 0.31 mHz (the slowest free oscillation) up to 1 kHz (in near-surface geophysics): roughly 6 orders of magnitude. Figure 1 shows the amplitudes that can occur over most of this frequency range, from the background noise level at quiet sites (Berger et al., 2004) to the large signals that have been observed at moderate distances from very



Seismic Instrumentation, Figure 1 Seismic and instrument noise levels. The solid and long dashed lines show the lowest levels of vertical and horizontal ground motion observed by the stations of the Global Seismographic Network. The dotted lines show the seismometer Brownian noise limit for two different values of the product of the sensor mass and free period. The dashed lines show the noise levels for a high-quality short-period system using a moving-coil transducer (Rodgers, 1994), for a MEMS sensor used for measuring strong motion (Evans et al., 2005), for a sensitive broadband system (Ringler and Hutt, 2009), and for GPS used at seismic frequencies.

large earthquakes, to the even larger, and damaging, strong ground motions close to an earthquake. The large signals are infrequent, but also of considerable scientific importance – but so are very small signals, which can be seen only if the seismometer has lower noise than the ground does.

Since digital dataloggers can record over a much wider range of amplitudes and frequencies than the older analog recorders could, modern seismic systems can be characterized as broadband (covering a wide band of frequencies) and high dynamic range (a wide range of amplitudes). But the range from noise to the largest signal covers over 10 orders of magnitude, too much for any single instrument and datalogger. Since most seismometer designs require practical tradeoffs between low-frequency capability and other performance goals, seismometers are still classified by what frequencies and amplitudes they cover best. Systems for recording the largest signals without distortion are called *strong-motion* seismometers, and have usually been designed to meet the needs of earthquake engineers rather than seismologists, though digital technology is blurring this distinction.

Inertial seismometers: basic principles

Almost all seismometers measure motion of the ground using the inertia of an internal mass, and so are called *inertial* sensors. This mass, of amount m, is subject to five forces: (1) constraints that restrict the mass motion to be in a particular direction (denoted by a unit vector $\mathbf{e}$) making the system one with a single degree of freedom; (2) the gravitational force vector, $\mathbf{g}m$, along that direction, namely, $\mathbf{e} \cdot \mathbf{g}m$; (3) an elastic restoring force (from a spring), $-k(x - x_0)$, where k is the spring constant and x the displacement (in the direction $\mathbf{e}$) away from an equilibrium position x_0; (4) a viscous damping force proportional to velocity $-d\dot{x}$; (5) additional forces F_b that may be applied by a feedback system (discussed below).

The displacement of the mass relative to an inertial frame is $x + \mathbf{e} \cdot \mathbf{u}$, where x is the motion of the mass relative to the Earth, which is what we can measure, and $\mathbf{u}$ is the (vector) motion of the Earth relative to the inertial frame, which is that of the Earth in the absence of seismic waves (we can neglect non-inertial effects from the Earth's rotation). Combining forces (2), (3), and (4), and applying the momentum equation, the acceleration of the mass is given by

$$m(\ddot{x} + \mathbf{e} \cdot \ddot{\mathbf{u}}) = \mathbf{e} \cdot \mathbf{g}m - k(x - x_0) - d\dot{x}$$

where we allow the possibility of variations in both $\mathbf{e}$ and $\mathbf{g}$. With no motion, we assume that the mass position is x_0, so that $\mathbf{e}_0 \cdot \mathbf{g}_0 m = -kx_0$. Then the above equation

becomes, for small variations in **e** and **g**, and after some rearrangement of terms,

$$\frac{k}{m}x + \frac{d}{m}\dot{x} + \ddot{x} = -\mathbf{e}_0 \cdot \ddot{\mathbf{u}} - \mathbf{e}_0 \cdot (\mathbf{g} - \mathbf{g}_0) - (\mathbf{e} - \mathbf{e}_0) \cdot \mathbf{g}_0$$

more usually written as

$$\ddot{x} + 2\gamma\dot{x} + \omega_0^2 x = -\mathbf{e}_0 \cdot \ddot{\mathbf{u}} - \mathbf{e}_0 \cdot (\mathbf{g} - \mathbf{g}_0) - (\mathbf{e} - \mathbf{e}_0) \cdot \mathbf{g}_0 \quad (1)$$

where $\omega_0 = \sqrt{k/M}$ is the natural frequency of the seismometer ($T_0 = 2\pi/\omega_0$ is called the free period); $\gamma = d/2M$ is the damping constant. These names describe the behavior of the seismometer if the right-hand side of the equation is zero and $x = x_0$ at $t = 0$; the subsequent motion of the seismometer mass is then

$$x(t) = x_0 cos\left(\frac{2\pi t}{T_0}\right) e^{-\gamma t},$$

which is a decaying sinusoid with period T; the damping γ needs to be large enough to avoid ringing (a sign of resonance effects), and is usually set to about 0.8.

The right-hand side of Equation 1 shows that an inertial sensor responds to three inputs:

1. The acceleration of the ground, $\ddot{\mathbf{u}}$; this is why inertial sensors are often called accelerometers.
2. Changes in the gravitational vector **g** along the direction of sensitivity; if this is the dominant signal, the sensor is usually referred to as a gravimeter (see *Gravimeters*).
3. Changes in the direction of sensitivity relative to the direction of gravity. If $\mathbf{e}_0$ and $\mathbf{g}_0$ are parallel (the mass moves vertically), then to first order $\mathbf{e} - \mathbf{e}_0$ is perpendicular to $\mathbf{g}_0$, and this contribution is zero. However, if $\mathbf{e}_0$ and $\mathbf{g}_0$ are perpendicular (the mass moves horizontally), then to first order $\mathbf{e} - \mathbf{e}_0$ is parallel to $\mathbf{g}_0$, and this contribution can be significant; if this is the dominant signal, the sensor would be referred to as a tiltmeter, since changes in **e** reflect tilting of the sensor (or the ground it is attached to).

Much confusion has been created by the use of these different terms for what is the same kind of sensor, when the difference actually refers to the type of signal being measured.

Ignoring the tilt and gravity terms, suppose that the ground displacement is purely sinusoidal, $u(t) = Ue^{i\omega t}$ (U being complex, and the real part being taken). Then the mass motion will be $Xe^{i\omega t}$, where

$$\frac{X}{U} = \frac{\omega^2}{\omega_0^2 - \gamma i\omega + \omega^2} \quad (2)$$

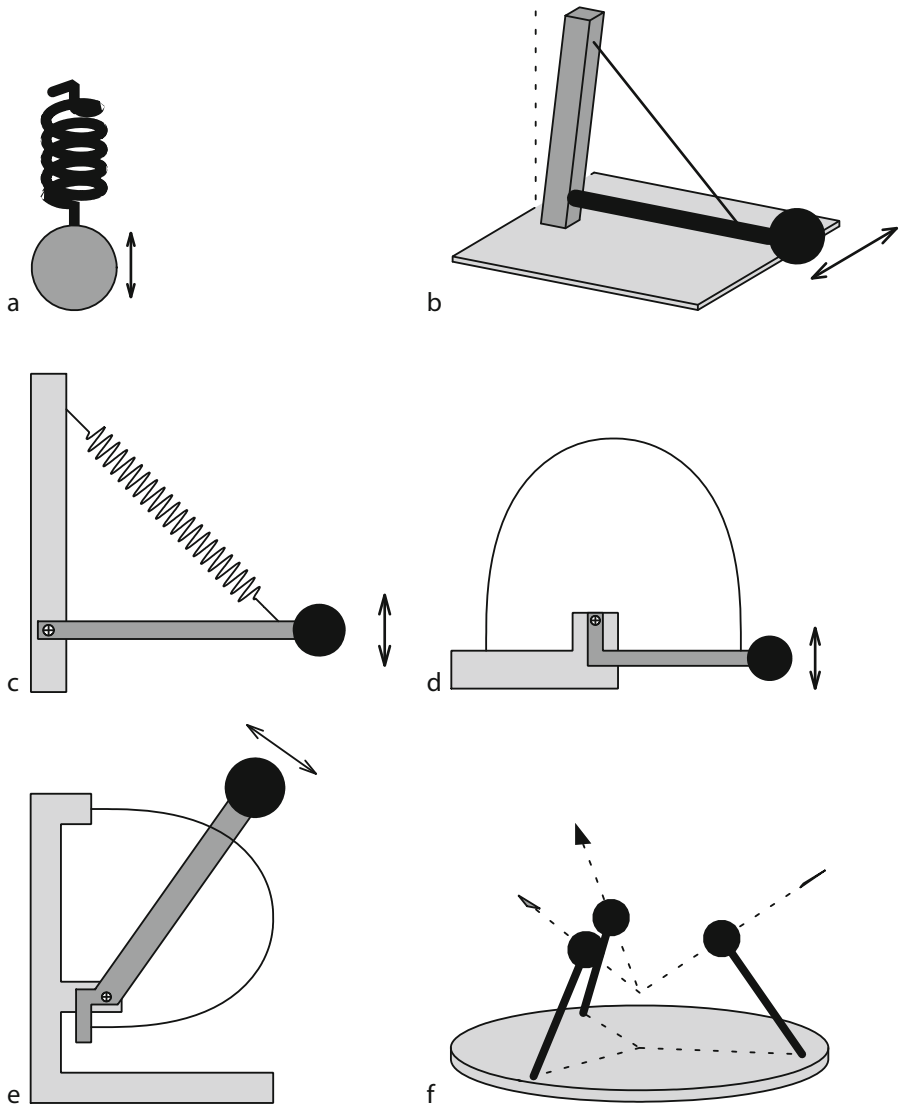
is the frequency response of the seismometer. For ground motion at frequencies much higher than the natural frequency ω_0, $X \approx U$: the mass motion looks like ground displacement. For ground motion at much lower frequencies, $X \approx U\omega^2/\omega_0^2$: the mass motion looks like ground acceleration, scaled by the inverse of the natural frequency of the seismometer.

The minimum noise level of any inertial sensor is set by the thermal (Brownian) motion of the mass in equilibrium with its surroundings. This noise level can be expressed as an equivalent ground acceleration, with a flat power spectral density of $8\omega_0\gamma k_B\theta/m$, where k_B is Boltzmann's constant and θ is the temperature (Aki and Richards, 2002). For a temperature of 310 K and $\gamma = 0.8$, this expression is $1.72 \times 10^{-19}/(mT_0)m^4s^{-3}$, with m in kilograms and T_0 in seconds. The dotted lines in Figure 1 show noise levels for two cases that cover the range of designs: $mT_0 = 10^{-4}$ (1 g mass and free period of 0.1 s), which cannot resolve ground noise at the longer periods, and $mT_0 = 3$ (300 g and 10 s), which shows that a moderately large mass and long period are necessary and sufficient for this thermal noise limit to fall below ground noise at all frequencies.

Mechanical design

In an inertial sensor, the mass needs to be constrained to move in one direction only, with a restoring force that is both exactly linear in the mass displacement and of whatever amount needed for the sensor to have the desired free period. The free period enters into the design partly because of Equation 2: the longer the period, the greater the displacement of the mass for a given acceleration at periods longer than the free period; that is to say, the higher the long-period sensitivity, and the lower the noise. Sensors designed to measure only high frequencies, such as the geophones used in geophysical exploration, can have a short free period, which means that the springs used can be relatively stiff (large k) and rugged. Seismometers for high frequencies use elastic elements both to constrain mass motion to a straight line and to provide a linear restoring force (Figure 2a). The latest innovation for short-period instruments is the MEMS (Micro-Electronic Mechanical Systems) sensor, in which the mass-spring system is machined from monolithic silicon and packaged with integral electronics; these are mass-produced for non-seismological applications and so are relatively inexpensive. While not as sensitive as traditional seismometers, they are already useful for measuring strong ground motions.

The high sensitivity needed for scientific purposes makes the mechanical design more challenging. If the mass moves vertically, it must be stably suspended against gravity while the restoring force is also kept low. In most long-period seismometers, the mass is constrained by hinges and stops to move along a small part of the circumference of a circle, which for most applications is an adequate approximation to a straight line (departures from this cause cross-coupling [Rodgers, 1968; LaCoste, 1967]). For sensing horizontal motion, the circle is slightly tilted from a horizontal plane, in which case the hinged



Seismic Instrumentation, Figure 2 Mechanical designs for seismometers, shown in cartoon form. (**a**) is a simple mass on a spring. (**b**) is the "garden-gate" design used for long-period horizontal instruments; note that the support post is not vertical. (**c**) shows the geometry of a Lacoste suspension; the restoring force of the spring must be equal to its physical length. (**d**) shows a vertical sensor that uses a leaf spring (elastica) to supply the restoring force. (**e**) and (**f**) show the design of a single sensor in a triaxial seismometer, and how three such sensors are arranged to record orthogonal motions.

mass is a horizontal pendulum (also called a garden-gate suspension, shown in Figure 2b). The restoring force then comes from gravity and is easily adjusted by tilting the instrument.

Long-period sensors that measure in other directions use a spring arranged to give a weak, but still linear, restoring force. One common design is the LaCoste suspension (Figure 2c), which uses a *zero-length spring*: a helical spring that exerts a force exactly proportional to its physical length. The geometry of the spring and hinged mass is chosen to make the restoring force nearly independent of mass position, giving a very long period. An alternative method uses a single sheet of elastic material, bent to form a leaf spring (Figure 2d), to create a long-period system with adequate linearity over the actual range of motion.

The traditional arrangement of seismometers is an orthogonal pair for horizontal motion, and a single sensor of vertical motion; if only one sensor is used, it is usually a vertical to give the best signal-to-noise ratio for P waves. Many systems now use a "triaxial" arrangement, with three identical systems measuring motion at an angle of 54.73° to the vertical, and at azimuths 120° apart; this also gives three orthogonal directions, namely those of the edges of a cube balanced on a corner (Figure 2e and f).

Any seismometer requires careful design to ensure that the only vibration that takes place is the mass motion, or at

least that other modes of vibration (called *parasitic resonances*) will not produce any signals if excited by ground motion. Long-period seismometers need special materials that do not creep much under load and whose dimensions and elastic constants are insensitive to temperature changes; even then, isolation and active environmental control may be needed to reduce the effects of changes in air temperature, air pressure (which causes varying buoyancy of the mass), and magnetic fields (which affect spring materials).

Electronic design

Sensing the mass motion in a seismometer is now almost always done electronically, in one of two ways (Agnew, 1986). In the first, electrodynamic method, permanent magnets and coils of wire are arranged so that the motion of the sensor produces an induced voltage $V = G\dot{x}$, where G is the generator constant; this voltage is then amplified and recorded. This method is simple and rugged, and provides damping if some of the induced current flows through a resistor. However, because the voltage produced depends on the velocity of the mass, and because the noise in the amplifiers rises with low frequency, electrodynamic sensing cannot resolve ground noise at periods longer than about 10 s.

The second method is active sensing of the mass motion. An oscillating voltage is input to a variable capacitor or inductor attached to the mass, which produces an output voltage with the same frequency and amplitude proportional to x; this output voltage is then demodulated to produce $V = Sx$, where S is the sensitivity. Up to the point of demodulation, the signal is at a frequency (several kHz) at which electronics noise is relatively low, so such displacement sensors can be made sensitive enough to measure motions less than an atomic diameter.

This output voltage can be recorded directly, but more often is used in a feedback system: the integral of this voltage (over some frequency band) is used to apply an additional force to the mass, usually with a coil and magnet, acting against the mass displacement. Because this force acts against the apparent force of ground acceleration, this is called a force-balance accelerometer. Feedback does not decrease noise, but does have three other merits. First, it can be used to vary the seismometer response much more than mechanical modifications can; for example, a mechanical system with a free period longer than 10 s is difficult to make, but this is easy to create electronically; and the response of the electronic system will be much more stable over time. Second, because the feedback force reduces the mass motion, the mechanical system needs to be linear over a smaller range. Finally, the calibration depends on the relation between voltage and feedback force, and this is very stable over time.

Calibration

Equation 2 is an example of the transfer function of a seismometer: the relation, as a function of frequency, between the ground motion and the output. In that example, the output was assumed to be mass motion x; a more realistic example would include a displacement sensor that generates a voltage $V = Sx$ followed by a datalogger that digitizes the voltage and produces a number $N = AV$. The total transfer function of the system, from input Earth displacement to digital output, would then be

$$\frac{AS\omega^2}{\omega_0^2 - \gamma i\omega + \omega^2}. \tag{3}$$

Seismometer calibration consists of determining all the parameters in expressions like this one. Some components, such as the datalogger, are calibrated by putting a known signal in and recording the output. For the seismic sensor, this requires producing a known ground motion, which can be done with a shake table before the seismometer is installed. After installation, a seismometer calibration can be checked by putting a signal into the seismometer electronics that will apply force to the mass without otherwise disturbing the output.

Calibration information is usually provided in a form like Equation 3: a leading constant, and the (complex) polynomials that give the response as a function of a frequency. Often the polynomials are specified by giving their roots in the complex plane, referred to as poles and zeros. The frequency response can usually be found quite accurately, though the absolute response to ground motion is usually difficult to estimate to better than 1%. Sometimes mundane aspects of the calibration, such as the actual direction of measurement, can be in error by significant amounts.

Installation

An ideal seismometer installation should maximize signals of interest whether distant earthquakes or the shaking of a building, while minimizing noise that might hide these signals. Any actual installation will be a compromise between this ideal and such practical matters as access and cost. For earthquake signals, putting an instrument deeper below the Earth's surface will give better results, but if this requires a deep drill-hole, a shallower installation may be all that can be done. An extreme example of low-quality installations being accepted for reasons of cost is ocean-bottom seismometers, for which the standard installation is the seismometer sitting on the bottom, sometimes on very soft sediments, however it happens to land after being dropped off a ship.

On land, the shallowest installations are for temporary instruments used in geophysical exploration, with seismometers planted on the surface with an attached spike to hold them in place. More permanent installations range from instruments set on surface rock exposures or in shallow holes in soil, to purpose-built deep vaults, caves (if available), and drill-holes. Deeper installations reduce the noise from wind and the range of temperature variation, and even a meter of soil will attenuate daily fluctuations. It is advisable, though not always possible, for instruments to be installed away from noise sources such as roads, machinery, rivers, and trees; in an urban setting borehole

installations at depths of a few hundred meters may be the only way to reduce the noise to reasonable levels.

Displacement seismometry using satellites

Because the mass motion is proportional to acceleration for motions with periods longer than the seismometer free period, data from inertial seismometers has to be integrated twice to find long-period displacements. This integration can introduce significant error, especially if the response is slightly nonlinear or the sensor tilts.

Direct sensing of displacement can now be done using repeated distance measurements between a ground sensor and Earth satellites, since the satellites define an inertial reference frame unaffected by ground motion. Distances between satellites of the Global Positioning System (GPS) and a receiver on the ground can be measured to within less than a millimeter precision, even over short times and in the presence of large accelerations; the actual ground displacement can be found after some processing (Bilich et al., 2008). The accuracy of GPS distance measurements is lower than their precision because of propagation effects, including interference from signals reflected from nearby objects (known as multipath), but it has proved possible to determine horizontal ground displacements to within a few millimeters over time spans up to several minutes and at rates up to 10 Hz. This sensitivity is too low to measure most earthquake signals, but can provide high-quality displacement data for ground motions near large earthquakes; a combination of inertial and GPS sensors is the optimal way to measure strong motion.

Deformation seismometers

A final class of seismometers measures, not the displacement at a point, but quantities related to the displacement gradient. For a vector displacement $\mathbf{u}$, the gradient $\nabla\mathbf{u}$ is a tensor, which can be decomposed into two parts: a symmetric part, the strain tensor, which describes local deformation; and an antisymmetric part, which is equivalent to a vector that describes local rotation, tilting about horizontal axes and rotating about the vertical axis.

Two types of instruments, both known as strainmeters, can measure components of the strain tensor (Agnew, 1986). Extensometers measure relative displacement along a line between two end points; this displacement divided by the baseline (the distance between the points) gives the extensional or compressional strain. Extensometers with baselines as short as a few centimeters are installed in boreholes, and ones with baselines up to tens of meters in tunnels. In both of these, the measurement is made relative to a solid length standard. Much longer instruments use optical methods to measure strain over hundreds of meters. The other class of strainmeter is the volumetric type, in which the change in a volume of fluid is found by sensing the displacement of fluid in and out of a container cemented in a borehole: such an instrument measures the volume strain, or dilatation.

Measuring the rotation vector requires either a stable direction of reference or some way of measuring rotation directly with respect to inertial space. Until recently, no technology existed that do could this with the low noise levels required. A few instruments measure rates of rotation around a vertical axis using Sagnac interferometers, also called laser gyroscopes (Schreiber et al., 2009). These sense the difference in frequency between two light beams propagating in opposite directions around a closed loop.

Seismic data recording

Three specialized requirements for seismological dataloggers are: (1) a high dynamic range, to capture all the signals possible; (2) large amounts of storage, since it may be necessary to record for a long time to capture unpredictable events; and (3) accurate absolute timing, to relate seismic-wave travel times to a common system. In all three areas, the progress of electronics has meant rapidly improving performance at ever-lower cost.

Many seismic dataloggers use specialized systems to provide 24 bits of range (about 5×10^7); this is accomplished by oversampling followed by digital filtering. The amount of storage depends on the sample rate; except in limited situations, a rate of about 200 Hz will capture all seismic data without aliasing. A day of 3-component data with this resolution and sample rate comes to just over 50 Mb, an amount that in the last two decades has gone from requiring specialized and bulky storage to something easily dealt with. Cross-correlation methods can find time delays to a precision of 0.1 of the sample interval, and so the time of each sample point should be known well: for 200 Hz sampling, this requirement would be 500 μs. This too is a level of accuracy that until recently was not easily attained. It can be reached without much difficulty if radio signals from the GPS satellites are available, though providing it over long times in the absence of such a signal still requires expensive equipment. Lower levels of accuracy can be obtained from other radio signals, and, over intervals of a few days, from inexpensive crystal oscillators.

Often, a single system records data telemetered from multiple sensors, forming an array (if the region covered spans only a few wavelengths of the waves being recorded) or network (if larger). The most precise timing then requires corrections for the transmission time of the data (latency) unless a separate datalogger is used for each sensor.

Summary

With over a century of development, inertial seismic sensors are a mature technology, usually capable of recording ground motion much better than it can be modeled. While no single sensor can cover the full range of amplitudes and periods of seismic waves, only a few instruments are needed to provide a faithful record of ground motion. Since other developments in electronics have largely routinized digital recording of seismic data, most users treat seismometers as a "black box" system that can be acquired

and used with little specific expertise: a usually justifiable assumption, though as always it is important to know enough to recognize poor performance when it does occur.

Bibliography

Agnew, D. C., 1986. Strainmeters and tiltmeters. *Reviews of Geophysics*, **24**, 579–624.

Aki, K., and Richards, P. G., 2002. *Quantitative Seismology.* Sausalito: University Science Books.

Berger, J., Davis, P., and Ekström, G., 2004. Ambient Earth noise: a survey of the global seismographic network. *Journal of Geophysical Research*, **109**, B11307, doi:10.1029/2004JB003408.

Bilich, A., Cassidy, J. F., and Larson, K. M., 2008. GPS seismology: application to the 2002 M_W 7.9 Denali Fault earthquake. *Bulletin of the Seismological Society of America*, **98**, 593–606.

Bormann, P. (ed.), 2009. *New Manual of Seismological Observatory Practice.* http://www.iaspei.org/projects/NMSOP.html

Evans, J. R., Hamstra, R. H., Kundig, C., Camina, P., and Rogers, J. A., 2005. TREMOR: a wireless MEMS accelerograph for dense arrays. *Earthquake Spectra*, **21**, 91–124.

Havskov, J., and Alguacil, G., 2004. *Instrumentation in Earthquake Seismology.* Dordrecht: Springer.

LaCoste, L. J. B., 1967. Measurement of gravity at sea and in the air. *Reviews of Geophysics*, **5**, 477–526.

Ringler, A. T., and Hutt, C. R., 2009. Self-noise models of seismic instruments. *EOS Transactions, American Geophysical Union, Fall Meet. Suppl.*, **90**, S23A–1736.

Rodgers, P. W., 1968. The response of the horizontal pendulum seismometer to Rayleigh and Love waves, tilt, and free oscillations of the earth. *Bulletin of the Seismological Society of America*, **58**, 1384–1406.

Rodgers, P. W., 1994. Self-noise spectra for 34 common electromagnetic seismometer/preamplifier pairs. *Bulletin of the Seismological Society of America*, **84**, 222–229.

Schreiber, K. U., Hautmann, J. N., Velikoseltsev, A., Wassermann, J., Igel, H., Otero, J., Vernon, F., and Wells, J. P. R., 2009. Ring laser measurements of ground rotations for seismology. *Bulletin of the Seismological Society of America*, **99**, 1190–1198.

Wielandt, E., 2002. Seismometry (pp. 283–304). In Lee, W. H. K. (ed.), *International Earthquake and Engineering Seismology Part A*. New York: Elsevier Academic Press.

Cross-references

Earthquake, Location Techniques
Earthquakes, Strong-Ground Motion
Free Oscillations of the Earth
Gravimeters
Seismic Noise
Seismogram Interpretation
Seismology, Rotational

SEISMIC MICROZONATION

Fumio Yamazaki, Yoshihisa Maruyama
Graduate School of Engineering, Chiba University, Inage-ku, Chiba, Japan

Definition

Seismic microzonation. The mapping of an area on the basis of various factors that can affect the intensity of ground shaking, such as seismic hazard, geological conditions, and topographical features, so as to account for the effects of local conditions on earthquake-induced damage.

Introduction

Local site conditions affect the intensity of ground shaking, and as a consequence, the extent of earthquake-induced damage. The amplitude, frequency content, and duration of strong ground motion are significantly influenced by local site conditions. A well-known example is the 1985 Mexico City earthquake. Although the fault rupture of the earthquake was about 350 km away from Mexico City, the city sustained catastrophic damage due to the strong amplification of the ground motion by soft soil deposits (Seed et al., 1988). The 1989 Loma Prieta earthquake caused extensive damage in the San Francisco Bay Area. The San Francisco Bay mud significantly influenced the amplitude, frequency content, and duration of ground shaking and resulted in the collapse of the northern portion of the I-880 Cypress Viaduct (Earthquake Engineering Research Institute, 1990; Kramer, 1996). Seismic microzonation provides the basis for site-specific risk analysis, which can assist in the mitigation of earthquake-induced damage.

Methodology

Seismic microzonation typically involves the mapping of predominant periods, soil amplification factors, topographical conditions, liquefaction susceptibility, etc. To draft microzonation maps for a particular region, various data such as existing geological maps, borehole survey data, seismic observation data, and microtremor observation data are collected. Since seismic microzonation entails spatial classification of soil conditions in a small area (e.g., a city), geological data are required for not just a single location, but for many locations. In this regard, geological classification maps are most often used as one of the data sources. However, to classify the target area in a more quantitative manner, actual soil profiles obtained from borehole survey data or seismic observation data are better sources. Unfortunately, in most cases, the borehole survey data and/or seismic observation data available for a small area are insufficient. Thus, microtremor observation data have emerged as a popular source for dense spatial information on site amplification characteristics. Three examples of seismic microzonation are described hereafter.

Example 1. Seismic microzonation based on geomorphological classification maps

Several seismic microzonation studies in Japan have employed geomorphological and geological data from the Digital National Land Information (DNLI), which is a GIS database that covers the whole of Japan with a 1×1 km mesh, to estimate site amplification characteristics (Matsuoka and Midorikawa, 1995; Fukuwa et al., 1998; Yamazaki et al., 2000).

Wakamatsu et al. (2004) drafted the Japan Engineering Geomorphologic Classification Map (JEGM) on the basis of the analysis of local geomorphological features at scales of 1:50,000, and all the attributes were digitized and stored in a GIS database. They recently extended the JEGM to 250 × 250 m grid cells that were categorized into 24 classes on the basis of geomorphological characteristics.

The shear-wave velocity averaged over the upper 30 m (V_s^{30}) is often used as a simplified index of site conditions (Building Seismic Safety Council, 2003). Region-wide site condition maps for California were constructed on the basis of V_s^{30} and the classification of geological units (Wills et al., 2000). The Next Generation of Ground-Motion Attenuation Models (NGA) project was launched in an attempt to collect all publicly available site condition information at strong motion stations. V_s^{30} is used in the absence of site condition information (Chiou et al., 2008). Matsuoka et al. (2006) constructed a nationwide V_s^{30} distribution map using the nationwide shear-wave velocity datasets for Japan, which were obtained from 1,000 K-NET and 500 KiK-net seismic stations and the JEGM.

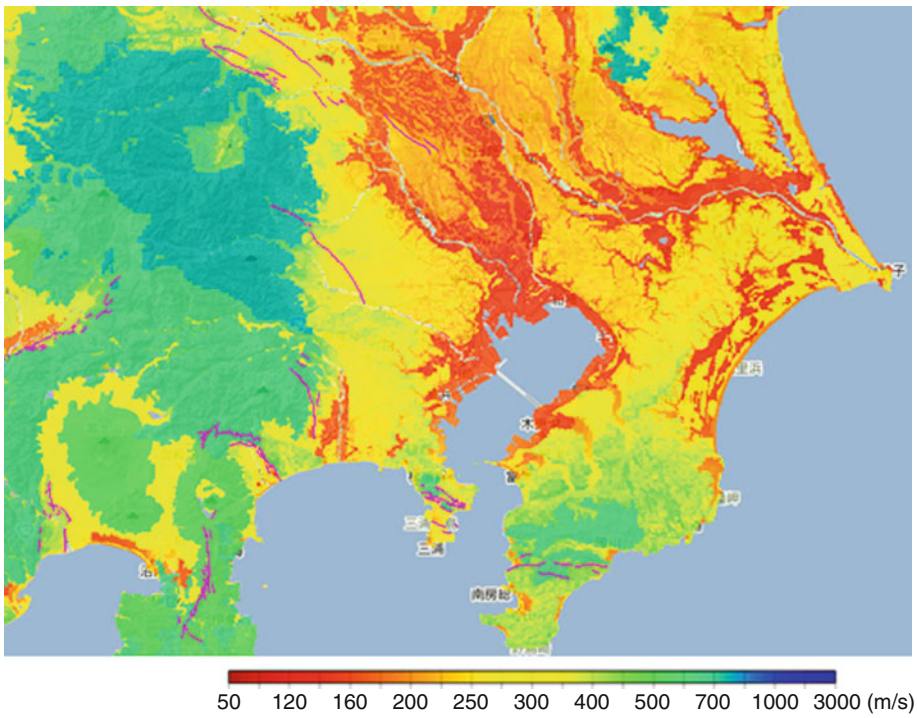
The National Research Institute for Earth Science and Disaster Prevention (NIED), Japan, has developed an open web system that interactively provides seismic hazard maps for Japan; this system is called the Japan Seismic Hazard Information Station (J-SHIS) (Fujiwara et al., 2006). J-SHIS uses the JEGM and V_s^{30} distribution map to draw probabilistic seismic hazard maps for the whole of Japan made by the Headquarters of Earthquake Research Promotion, Japan (Figure 1).

Example 2. Seismic microzonation based on dense borehole data and GIS

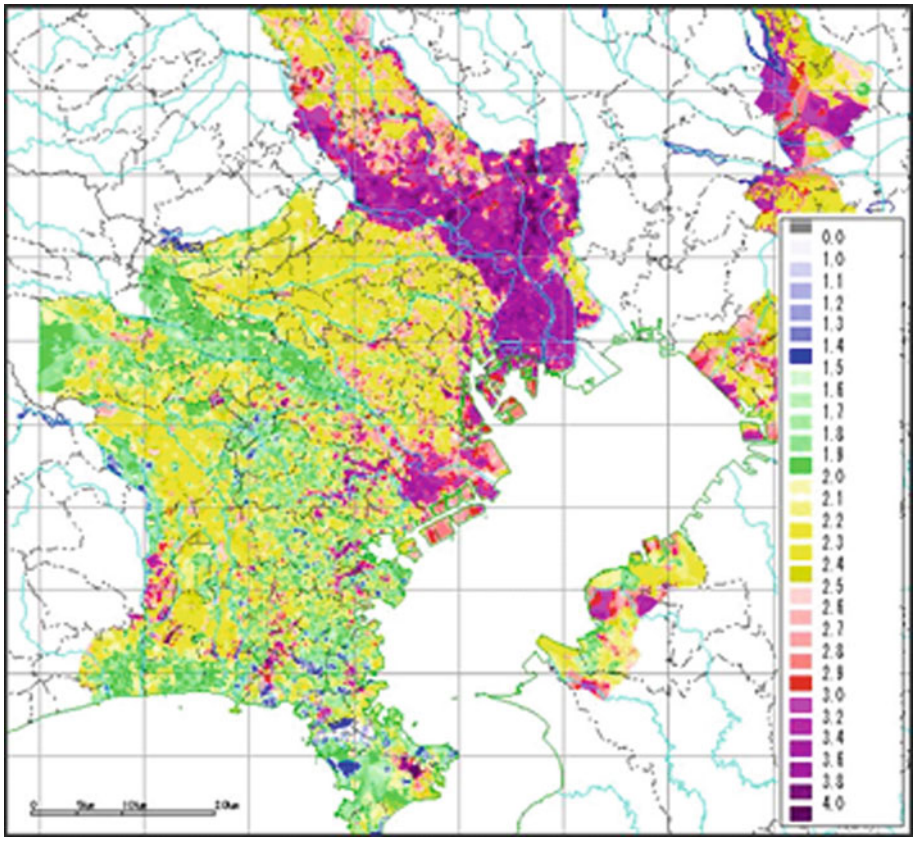
Since 2001, the Tokyo Gas Co., Ltd. has been operating the *Sup*er-Dense *Re*al-time *M*onitoring of *E*arthquakes (SUPREME) system, having about 4,000 seismometers (SI-sensors), in order to control natural-gas supply soon after the occurrence of earthquakes (Shimizu et al., 2006).

This system employs a GIS to interpolate the monitored spectral intensity (SI) values by using subsoil data from 60,000 boreholes. The digitized borehole data specify the location, depths of soil layers, classification of subsurface soil, standard penetration test (SPT) blow counts, surface elevation, and elevation of the ground water table. Thus, microzonation of the area on the basis of individual borehole data is possible. Shear-wave velocities are estimated from an empirical relationship by using the SPT-N values; then, the average shear-wave velocities in the top 20 m of soil at a borehole site are used to estimate the amplification factors of the SI values (Figure 2).

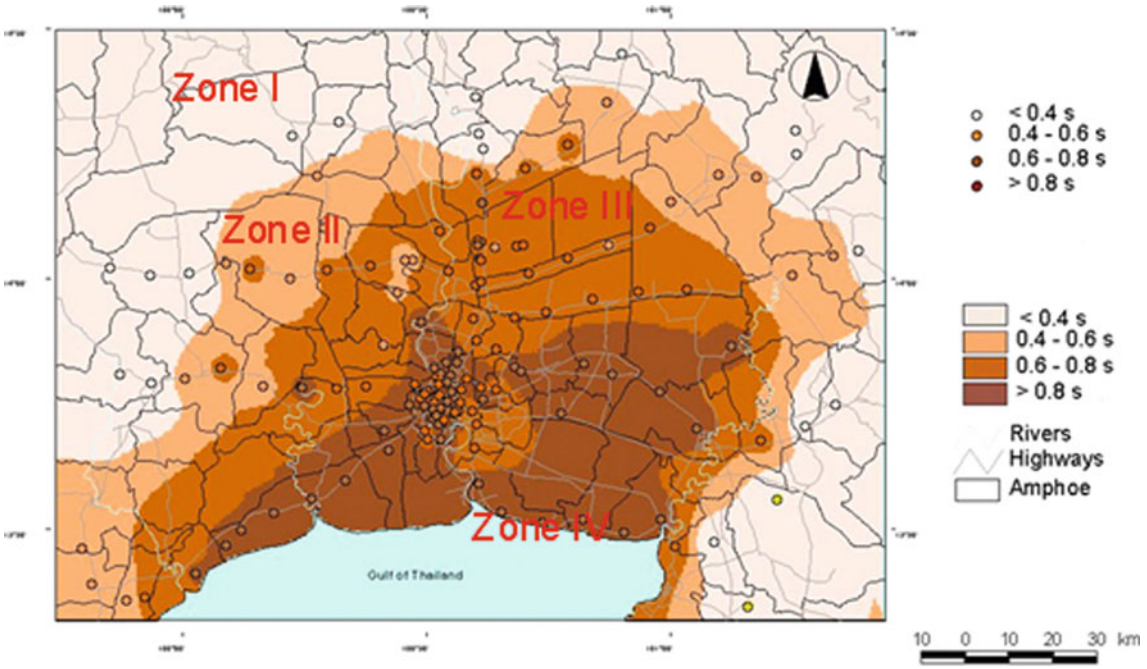
The accuracy of seismic microzonation can be confirmed after several years of operating a dense seismic network by evaluating the seismic records obtained for moderate to small earthquake events occurring in that period.



Seismic Microzonation, Figure 1 V_s^{30} distribution map of Tokyo metropolitan area (http://www.j-shis.bosai.go.jp/).



Seismic Microzonation, Figure 2 Site amplification map of Tokyo and surrounding areas, developed using dense borehole data.



Seismic Microzonation, Figure 3 Microzonation of greater Bangkok area on the basis of variation in predominant period.

Example 3. Seismic microzonation based on microtremor measurements

Microtremor measurements have emerged as a popular tool for determining the dynamic properties of soil layers, and hence, are being widely employed for seismic microzonation. In this method, ambient vibrations (of the order of microns) on the earth's surface are measured. The main sources of these vibrations are traffic and industrial and human activities (Kanai, 1983; Lermo and Chavez-Garcia, 1994). Microtremor measurements can be used to determine the predominant period of vibrations at a site. Nakamura (1989) proposed the horizontal-to-vertical (H/V) spectral ratio method, in which the predominant periods of ground vibrations are determined from the ratio of horizontal and vertical Fourier spectra of the microtremors recorded at a site. Konno and Ohmachi (1998) drafted a map of fundamental periods and amplification factors for the 23 wards of Tokyo on the basis of microtremor measurements carried out at 546 stations.

Tuladhar et al. (2004) drew a seismic microzonation map for the greater Bangkok area, Thailand, on the basis of microtremor observations carried out at 150 sites. The predominant periods of these sites were obtained by using the H/V method. The estimated predominant periods were validated by comparing them with the transfer functions obtained from one-dimensional wave-propagation analysis conducted at eight sites. According to the variation in the predominant period of the ground, the greater Bangkok area was classified into four zones as follows: Zone I (period less than 0.4 s), Zone II (0.4–0.6 s), Zone III (0.6–0.8 s), Zone IV (longer than 0.8 s). Figure 3 illustrates the microzonation of the greater Bangkok area on the basis of variation in the predominant period.

Summary

The objectives and methodologies to perform seismic microzonation are described and some examples are presented. The three major methods introduced to achieve seismic microzonation are the uses of geomorphological classification maps, dense borehole datasets, and microtremor measurements. The results of seismic microzonation are compiled for a GIS to draft microzonation maps and they can be used to predict ground motions during disastrous earthquakes and thus can assist in the mitigation of earthquake-induced damage.

Bibliography

Building Seismic Safety Council (BSSC), 2003. *The 2003 HEHRP Recommended Provisions for the Development of Seismic Regulations for New Buildings and Other Structures*. Washington, DC: FEMA.

Chiou, B., Darragh, R., Gregor, N., and Silve, W., 2008. NGA project strong-motion database. *Earthquake Spectra*, **24**(1), 23–44.

Earthquake Engineering Research Institute, 1990. Loma Prieta earthquake reconnaissance report. *Earthquake Spectra*, **6**.

Fujiwara, H., Kawai, S., Aoi, S., Morikawa, N., Senna, S., Kobayashi, K., Ishii, T., Okumura, T., and Hayakawa, Y., 2006. National seismic hazard maps of Japan. *Bulletin of the Earthquake Research Institute, University of Tokyo*, **81**, 221–231.

Fukuwa, N., Arakawa, M., and Nishizaka, R., 1998. Estimation of site amplification factor using Digital National Land Information. *Journal of the Structural Engineering*, **44**(B), 77–84 (in Japanese).

Kanai, K., 1983. *Engineering Seismology*. Tokyo: University of Tokyo Press, pp. 128–139.

Konno, K., and Ohmachi, T., 1998. Ground-motion characteristics estimated from spectral ratio between horizontal and vertical components of microtremor. *Bulletin of the Seismological Society of America*, **88**(1), 228–241.

Kramer, S. L., 1996. *Geotechnical Earthquake Engineering*. Englewood Cliffs: Prentice Hall.

Lermo, J., and Chavez-Garcia, F. J., 1994. Are microtremors useful in site response evaluation? *Bulletin of the Seismological Society of America*, **84**(5), 1350–1364.

Matsuoka, M., and Midorikawa, S., 1995. GIS-based integrated seismic hazard mapping for a large metropolitan area. In *Proceedings of the Fifth International Conference on Seismic Zonation*, Nice, France, Vol. II, pp. 1334–1341.

Matsuoka, M., Wakamatsu, K., Fujimoto, K., and Midorikawa, S., 2006. Average shear-wave velocity mapping using Japan Engineering Geomorphologic Classification Map. *Journal of Structural Engineering and Earthquake Engineering (Japan Society of Civil Engineers)*, **23**(1), 57s–68s.

Nakamura, Y., 1989. A method for dynamic characteristics estimation of subsurface using microtremor on the ground surface. *Quarterly Reports of the Railway Technical Research Institute*, **30**, 25–33.

Seed, H. B., Romo, M. P., Sun, J. I., Jaime, A., and Lysmer, J., 1988. The Mexico earthquake of September 19, 1985 – relationships between soil conditions and earthquake ground motions. *Earthquake Spectra*, **4**, 687–729.

Shimizu, Y., Yamazaki, F., Yasuda, S., Towhata, I., Suzuki, T., Isoyama, R., Ishida, E., Suetomi, I., Koganemaru, K., and Nakayama, W., 2006. Development of real-time control system for urban gas supply network. *Journal of Geotechnical and Geoenvironmental Engineering, ASCE*, **132**(2), 237–249.

Tuladhar, R., Yamazaki, F., Warnitchai, P., and Saita, J., 2004. Seismic microzonation of the greater Bangkok area using microtremor observations. *Earthquake Engineering and Structural Dynamics*, **33**(2), 211–225.

Wakamatsu, K., Matsuoka, M., Hasegawa, K., Kubo, S., and Sugiura, M., 2004. GIS-based engineering geomorphologic map for nationwide hazard assessment. In *Proceedings of the 11th International Conference on Soil Dynamics & Earthquake Engineering and 3rd International Conference on Earthquake Geotechnical Engineering*, Vol. 1, pp. 879–886.

Wills, C. J., Petersen, M., Bryant, W. A., Reichle, M., Saucedo, G. J., Tan, S., Taylor, G., and Treiman, J., 2000. A site-conditions map for California based on geology and shear-wave velocity. *Bulletin of the Seismological Society of America*, **90** (6B), S187–S208.

Yamazaki, F., Wakamatsu, K., Onishi, J., and Shabestari, K. T., 2000. Relationship between geomorphological land classification and site amplification ratio based on JMA Strong Motion Records. *Soil Dynamics and Earthquake Engineering*, **19**(1), 41–53.

Cross-references

Earthquakes, Intensity
Earthquakes, Strong-Ground Motion
Seismic Hazard
Seismic Zonation
Seismicity, Intraplate
Seismology, Global Earthquake Model

SEISMIC MONITORING OF NUCLEAR EXPLOSIONS

Paul G. Richards[1], Wu Zhongliang[2]
[1]Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, USA
[2]Institute of Geophysics, China Earthquake Administration, Beijing, China

Introduction

The original development of nuclear weapons, and their first use in 1945, was followed by several decades of further weapons development in which more than 2,000 nuclear test explosions were conducted. About 500 of these were carried out in the atmosphere, mostly in the 1950s and 1960s. They generated radioactive fallout that was detected worldwide with some regional concentrations, and aroused widespread public opposition to nuclear testing. A few nuclear tests were carried out underwater and in space. The great majority, about 1,500, were conducted underground in ways that greatly reduced fallout – the first of them in 1957, in Nevada, USA – generating signals that have been intensively studied by seismologists. Hundreds of these individual nuclear tests consisted of multiple nuclear devices and exploded almost simultaneously.

A ban on nuclear testing in the atmosphere, underwater, or in space, was negotiated and went into effect in 1963 between the USA, the USSR, and the UK. Known as the Limited Test Ban Treaty (LTBT), it has since been ratified or acceded to by more than a hundred countries. Though France and China did not sign, and China carried on with nuclear testing in the atmosphere up to 1980, eventually both these countries came to abide by its terms.

The concept of a Comprehensive Test Ban Treaty (CTBT) emerged in the 1950s, intended as a restraint upon nuclear weapons development. It was debated in many forums for more than 40 years, and finalized in terms of specific treaty text in September 1996. But this treaty is not in effect (as of 2010), due to continuing debate in specific countries that have not ratified this treaty, and whose ratification is needed as a condition for the CTBT to enter into force. They include India, North Korea, and Pakistan (not signed or ratified); and China, Israel, and the United States (signed but not ratified). Those countries that have signed the treaty are effectively adhering to a moratorium on nuclear testing. They include the five countries recognized as nuclear weapons states by the Non-Proliferation Treaty of 1968. Listing them in the order in which they acquired nuclear weapons capability, these are the USA, the USSR (whose CTBT obligations have been assumed by Russia), the UK, France, and China. The two countries that by far have conducted the most nuclear test explosions – the USA with 51% of the world total, and the USSR/Russia with 35% – ended nuclear testing in the early 1990s. See Yang et al. (2003) for lists of nuclear explosions conducted in the twentieth century, and Bennett et al. (2010) for a relevant database and seismic waveforms. Since 1996, the only nuclear explosions (as of 2010) have been those conducted by India and Pakistan (in May 1998), and by North Korea (in October 2006, and May 2009).

Seismic monitoring of nuclear explosions has been an important activity ever since the first nuclear test in July 1945 in New Mexico. Such monitoring is driven by two different objectives that have engaged a range of different institutions and organizations. The first objective, which dominated for the early decades of nuclear testing up to the early 1990s when nuclear explosions were being conducted on average about once a week, was to acquire basic information about military weapons being tested, especially if (from the point of view of the monitoring organization) the tests were being carried out by a potential adversary. Relevant questions were: what countries had nuclear weapons programs, developed to the level of carrying out nuclear explosive tests? And how big were these explosions? The second objective, which has become important in recent decades, has been in the context of a major initiative in nuclear arms control, namely, to achieve confidence in the capability to monitor compliance with a CTBT, recognizing that many countries considering whether or not to support such a treaty and to be bound by its terms, would need to have confidence in the monitoring system to some adequate degree. Given that monitoring cannot be done all the way down to zero yield, evaluation of progress toward this second objective entails questions such as: down to what small size can nuclear explosions be detected, and identified, and attributed with high confidence? And what are the specific capabilities of different types of monitoring program, applied to different parts of the world, to catch evidence of a nuclear test, should one occur?

Seismology is the most effective technology for monitoring nuclear tests carried out underground, which is the one environment that was not covered by the LTBT, and which is also the hardest of the environments to monitor. The importance of achieving the two objectives stated above has shaped modern seismology itself, in that much of the funding that has led to the facilities and bodies of knowledge now used widely in seismological research (including studies of seismic hazard), were stimulated by government programs intended to improve capabilities for seismic monitoring of nuclear explosions. These facilities and methods include high-quality Seismic Instrumentation, global networks that monitor for earthquakes as well as explosions, quantitative methods of characterizing seismic sources (various magnitude scales, the moment tensor), theoretical understanding of seismic wave propagation in Earth models of increasing and more realistic complexity, our knowledge of the Earth's internal structure, and methods of seismic signal detection and interpretation.

The technical capability to monitor explosions, or a perceived lack of such capability, has played a role in the development of policy options on weapons testing

and/or arms control and the content of international treaties. A key technical question arising in debates has been: down to what value of yield can monitoring be accomplished – and with what level of confidence? Seismologists claim now that there is no fundamental technical problem with monitoring explosions down to 1 kt, even if determined efforts at evasion must be considered. But there have been assertions that it is possible to muffle and thus hide (or confuse the procedures for identifying) the seismic signal, even from a substantial underground explosion at the level of ten kilotons or more. These latter assertions do not appear plausible after review of the technical difficulties; but, as assertions, one finds that they continue to survive.

Seismic monitoring for underground nuclear explosions must be done with recognition of the great variety and number of earthquakes, chemical explosions, and other nonnuclear phenomena that generate seismic signals every day. Efforts to sort out and identify signals from underground nuclear explosions in the midst of signals from these other phenomena have made great progress since they commenced in the 1950s, and improvements in monitoring capability will surely continue to be made.

Sections below describe basic properties of earthquake and explosion signals, and different steps in seismic monitoring for nuclear explosions. A review is given of methods used for decades in the era when thousands of kilometers separated nuclear weapons testing activity and monitoring stations, when nuclear weapons testing was commonplace and there was little incentive to hide testing activity. Descriptions are then given of modern methods that monitor for very small explosions and the possibility of tests conducted in ways intended to evade discovery. A description is given of so-called "problem events" that were important is developing effective and in some cases new discriminants; and finally a brief summary is given of monitoring capabilities, as of 2010, emphasizing the utility of data and data products from the International Monitoring System and its associated International Data Centre that are operated today by the CTBT Organization, headquartered in Vienna, Austria.

Basic properties of earthquake and explosion signals

Seismic monitoring for underground nuclear explosions has to face the reality of hundreds of earthquakes, chemical explosions, and other nonnuclear phenomena, generating seismic signals daily that will be recorded at multiple stations by any effective monitoring network. But after decades of effort, an extensive infrastructure of national and international agencies now sorts out and identifies the signals from earthquakes, chemical explosions, and the occasional underground nuclear explosion. Modern methods of nuclear explosion monitoring are vastly more capable than they were when this work began in the late 1950s. The improvements have mostly been steady as data quality and quantity from monitoring networks increased, but with occasional jumps in capability as new types of analysis were validated.

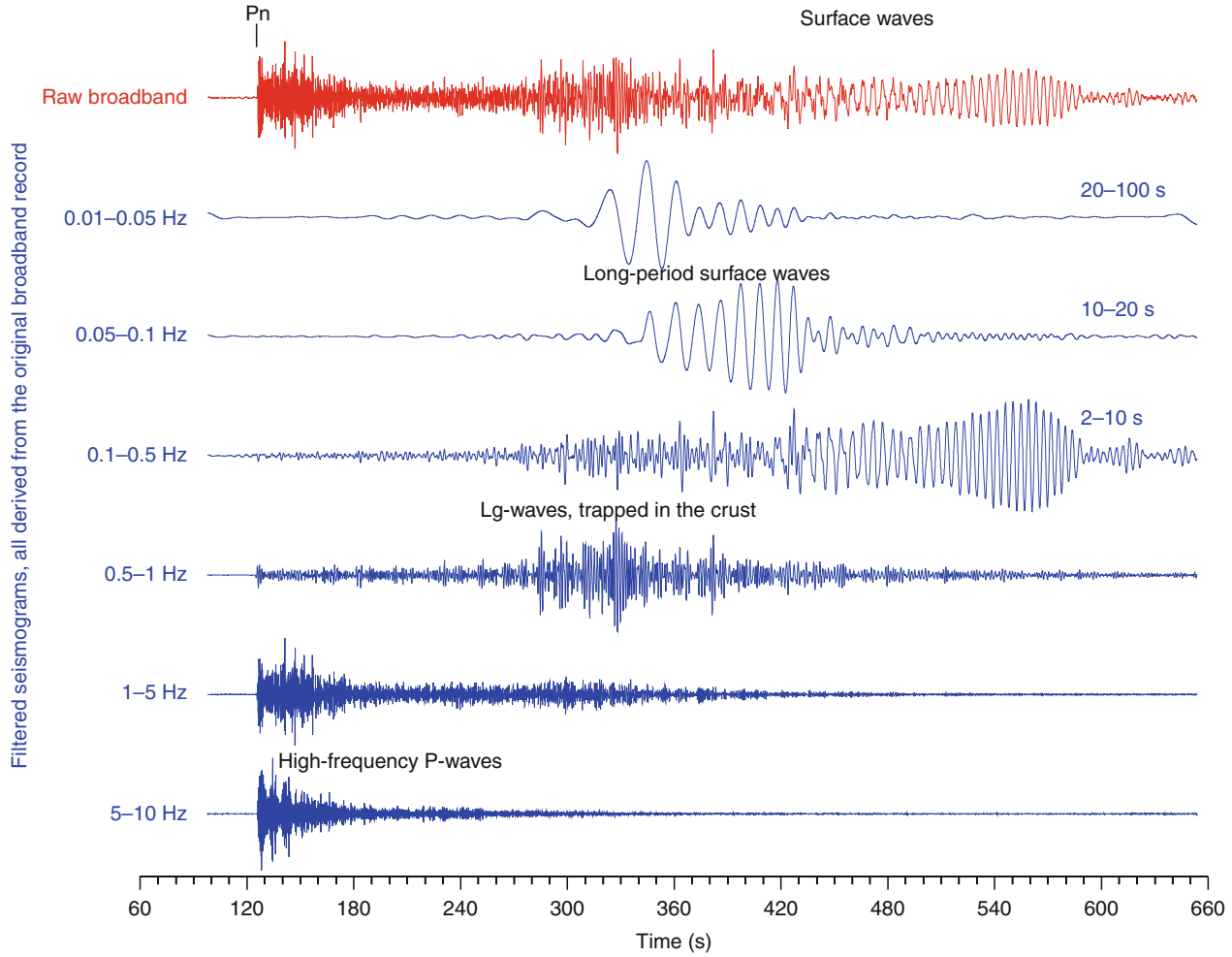
Seismic signals are traditionally grouped into teleseismic waves and regional waves, depending on the distance at which they are observed. Teleseismic waves propagate either as Body Waves through the Earth's deep interior, emerging with periods typically in the range 0.3–5 s at distances greater than about 1,500 km, or as Surface Waves, analogous to the ripples on the surface of a pond, with periods of about 15–100 s.

Teleseismic waves were the basis of most US monitoring of foreign nuclear tests prior to 1987. Teleseismic body waves are further subdivided into *P*-waves and *S*-waves. *P*-waves, which are the fastest-traveling seismic waves and are therefore the first to arrive, are excited efficiently by explosions: earthquakes tend to excite *S*-waves and surface waves more efficiently.

For subkiloton explosions, teleseismic signals can be too weak for detection at distant stations and monitoring then requires regional signals. Regional waves are of several types, including *P*-waves and *S*-waves, all propagating only at shallow depths (less than 100 km below the Earth's surface) with periods as short as 0.05 s (frequencies as high as 20 Hz, i.e., cycles per second). Regional waves reach distances up to 1,000 km and sometimes beyond, depending on source size and whether the propagation path is an attenuating one, or not. They are regional also in the sense that they have speeds and attenuation properties that vary according to details of local structures in the Earth's crust and uppermost mantle, so they can vary from place to place within continents and oceans.

Figure 1 shows a regional seismogram of a Soviet underground nuclear explosion in Kazakhstan recorded in July 1989 at a distance of slightly less than 1,000 km by a high-quality station in northwestern China. The original recording is shown in red. Different signals derived from it are shown in blue, each of them filtered to pass information in a particular band of frequencies.

Seismologists characterize the size of seismic signals by means of logarithmic magnitude scales (see *Earthquake, Magnitude*), with each scale based on a different type of seismic wave. A magnitude scale using teleseismic surface waves was first described in the 1930s based on the logarithm (to the base 10) of amplitude of maximum ground displacement due to surface waves with periods about 20 s. It is known as the M_s scale. Another widely used magnitude scale is that based on the amplitude of teleseismic *P*-waves. Known as m_b, it entails measurement of ground motion at about 1 s period. As part of the assignation of M_s and m_b values, for a particular seismic event as recorded at a particular station, a standard correction is applied to account for the distance between the source and the receiver at which the data was obtained. Magnitudes range from about -3 for the smallest observable microearthquakes, up to above 8 for the largest earthquake. A 1 kt underground explosion has an m_b roughly about 4, and each year there are about 7,500 shallow earthquakes worldwide with $m_b \geq 4$ (Ringdal, 1985). Although use of



Seismic Monitoring of Nuclear Explosions, Figure 1 The seismogram recorded at station WMQ in northwestern China, for an underground nuclear explosion on July 8, 1989 in Kazakhstan at a distance of almost 1,000 km, is shown in red (*top*). Filtered versions of the original trace in different frequency bands are shown in *blue*. Time in seconds at *bottom* is with respect to the time the explosion occurred. Different types of seismic wave propagate at different frequencies, and hence their ground motions show up in different bands. *P*-waves, in this case the regional wave called *Pn* that travels in the uppermost mantle, arrive about 120 s after the explosion at this distance, involving short-period (high frequency) motions. Long-period surface waves can be seen in the top two *blue* traces. Some surface waves arrive up to 600 s after the explosion at this distance and, thus, travel as much as five times slower than *P*-waves. *S*-waves (weak in this example) are shear waves, traveling slower than *P* waves. A high-frequency wave marked as *Lg*, which is often the largest wave at regional distances from an earthquake but is only weakly excited by explosions, is dominated by shearing motions and is largely trapped in the Earth's crust. The amplitude of ground motion in the longest period band is less than 2% the amplitude in the short period band from 1 to 5 Hz. (Adapted from work of W.-Y. Kim.)

seismic moment has superseded use of m_b and M_s in much of modern seismology and magnitude is only an empirical estimator of seismic event size, magnitude scales are still often used in discussion of seismic monitoring because this a practical way to relate that discussion directly to properties of signal strength. For example, monitoring capability is often characterized in terms of contour maps or shaded maps indicating the magnitude levels down to which detection or identification is deemed possible with given resources, such as a particular network. We conclude this article with such a map (see Figure 8). Explosion energy is measured in kilotons. A kiloton is formally defined as a trillion calories, and is roughly the energy released by exploding a thousand tons of TNT.

The different steps in explosion monitoring

Nuclear explosion monitoring entails a series of steps, beginning with *detection* of signals (did a particular station detect anything?) and *association* (can we gather all the different signals, recorded by different stations, that originate from the same "event"?). The next steps involve making a *location* estimate and an *identification* (did it have the characteristics of an earthquake, a mining blast,

a nuclear weapon test?). Then follow the steps of *yield estimation* (how big was it?) and *attribution* (if it was a nuclear test, what country carried it out?).

Detection

Concerning detection, nuclear explosion monitoring is often done with arrays of sensors, deployed as a group spread out over an area about 10 km across (or less), that facilitate methods to enhance signal-to-noise ratios. This is done typically by stacking signals from independent sensors, often with appropriate delays to increase signal strength and reduce noise. Array data can also give estimates of the direction from which signals are arriving.

In the *evaluation* of detection capability, one of the key concepts widely used in seismology is the *magnitude of completeness*, which means that *all* events above this magnitude can be recorded by the monitoring system. Transferring from magnitude to yield, one infers the capability for detecting nuclear tests (NAS, 2002). Practically, however, one of the often-cited expressions of monitoring capability is the *magnitude threshold*, above which 90% of the seismic events can be detected at more than three stations, the least number of stations for routine location.

Association

Association is the effort to identify those sets of signals, from different stations, which all originate from the same seismic event. It is one of the hardest steps in practice, particularly when multiple seismic sources around the world are active at the same time, resulting in signals from different events that are interlaced in the waveforms recorded by each station. In such cases, array data can be helpful in resolving which signals correspond to which event.

Location

To obtain a location estimate, typically the arrival times of various seismic waves are measured from the recorded waveforms such as shown in Figure 1. They are used to find four parameters: latitude, longitude, depth, and origin time. In this work, it is necessary to know the travel time from any hypothesized source location to any particular seismographic station for any type of seismic wave that the station might observe. In practice, locating seismic events accurately on a global basis (say, to within 10 km of their true location) using sparse networks (stations several hundred kilometers apart) requires extensive efforts in station calibration. Thus, it is important to include path-specific travel-time corrections to standard travel-time models to account for lateral variations of Earth structure (Murphy et al., 2005; Myers et al., 2010). Many authors have shown that greatly improved precision of location estimates can be achieved for a given region if seismic events are located in large numbers – preferably thousands of them or more, all at the same time – rather than one at a time (Richards et al., 2006; Waldhauser and Schaff, 2008).

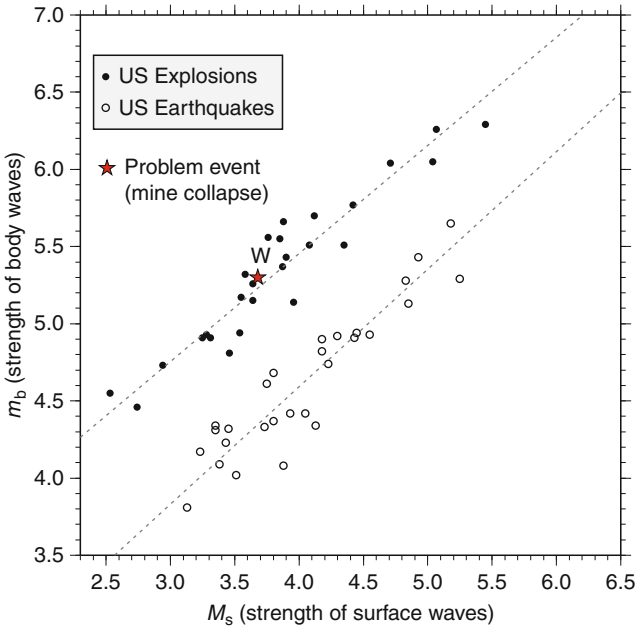
Methods of identification

Identification of the nature of a seismic source on the basis of its seismic signals – that is, making a determination from seismograms as to whether it could be a nuclear explosion, or a natural earthquake, or a mine blast, or something more exotic such as a bolide impacting our planet and exploding in the atmosphere – is a large subject in view of the many possibilities. See for example, Richards (1988), OTA (1988), Dahlman et al. (2009), and Bowers and Selby (2009). Seismic events generate many different types of seismic wave, in various different frequency bands as shown in Figure 1, and different types of seismic source generate a different mix of seismic waves. We can make an analogy here with sound waves, and the capability of the human ear and brain to analyze them. A deep bass voice, a gunshot, a whistle, and rolling thunder, constitute a set of sound sources that are easily distinguished from each other on the basis of their different frequencies, their emergent or impulsive nature, and their duration. It is the mix of information in both the time domain and the frequency domain that is effective.

Seismic methods for discriminating between earthquakes and explosions are based on interpretation of the event location (including its depth); on the relative excitation of a variety of body waves and surface waves; and on properties of the signal spectrum associated with each of these two different types of source. Within these three broad categories, many different methods have been tried, with various degrees of success. As the capabilities of each method are probed, the question of interest is often: "Down to what size of seismic event, does this method of discrimination work?" In some cases, discrimination is unambiguous even at very small event size. (For example, however small an event, it may be presumed to be an earthquake if it is located at a depth greater than 15 km below the Earth's surface. Even a small event will attract attention if it occurs in an area that is geologically stable that for decades has had no seismic activity.)

The most useful methods for discrimination can be listed as follows:

- Interpretation of the location: Is the event in a seismic or an aseismic area? Below the floor of an ocean? At depth below a continent? There is an important role here for common sense: seismic events in Canada tend to attract less attention from western monitoring agencies than such events in North Korea (though a seismic event in the middle of the Canadian Shield would still attract attention and intensive study).
- Relative amplitude of body waves and surface waves. This can be studied by plotting the event of interest on an M_s: m_b diagram, as shown in Figure 2. The surface-wave amplitude is read typically from signals with period about 20 s, and the body-wave amplitude at about 1 s period. (Though effective for large enough



Seismic Monitoring of Nuclear Explosions, Figure 2 An M_s: m_b diagram from Bowers and Walter (2002). It can be seen here that for seismic events of the same M_s value, earthquakes have a significantly smaller m_b value than do the explosions. The offset is about 0.8 m_b units, at $M_s = 5$. Because magnitudes are based on logarithmic scales, and $10^{0.8} \sim 6$, it follows that at frequencies near those at which body wave magnitude is measured (about 1 Hz), the *P*-waves from an underground nuclear explosion are about 6 times larger than such waves from an earthquake having the same strength of surface waves. Also, indicated by the red star are the body-wave and surface-wave magnitudes of an interesting but fortunately rare event, a large mine collapse with *P*-wave magnitude greater than 5. This event, which plots with the explosion population, is discussed further below – see Figures 6 and 7.

events, an explosion with m_b much below 4.5 may not have large enough surface wave signals at teleseismic distances to apply this method dependably.)

- Use of the observed "first motion" of the ground. Is the initial *P*-wave motion of the ground indicative of compression radiated to all directions from the source, leading to upward motions, as would be the case for a simple explosion? Or, are dilatations recorded at some azimuths, leading to downward motions, as would sometimes be expected from earthquakes but not from explosions?

The methods described so far in this section have concerned the use of teleseismic signals, which can be used to monitor effectively for high magnitudes, and on down to somewhere in the magnitude range from 4.0 to 4.5. Since the early 1990s, there has been growing recognition of the merits of regional waves, to monitor down to far lower magnitudes, often well below magnitude 3. The method is based upon the general observation that explosion signals, when compared to earthquakes, have much stronger *P*-waves at high frequency, whereas those from earthquakes have stronger *S*-waves (and surface waves).

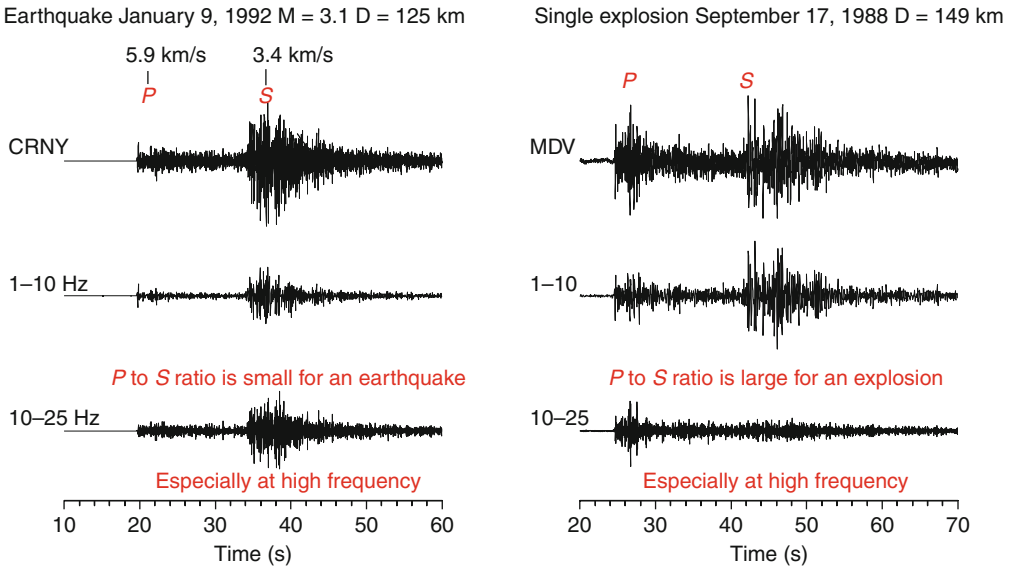
This modern method is being studied with frequencies in the range 0.5–20 Hz. and sometimes even higher. An example is shown in Figure 3 comparing regional signals of a very small earthquake and a small explosion. The method has been demonstrated even down to around m_b 2.

As an important example of this development, Figure 4 shows the results of an analysis of the *P*-wave and *S*-wave spectra, pertinent to identifying the very small underground nuclear explosion conducted by North Korea on October 9, 2006, and the larger test nearly 3 years later on May 25, 2009. The smaller explosion took place at 0135 h (GMT) and by 0706 h the US Geological Survey (USGS) had issued a report based on seismic signals from 20 stations around the world including sites in China, South Korea, Russia, Japan, Kazakhstan, Kyrgyzstan, Alaska, and Nevada. Its magnitude, about 4, indicated a sub-kiloton yield (see Koper et al., 2008, who discuss the uncertainty of estimating yield in view of the variability of seismic signal excitation for shots of different depth). But from such teleseismic signals, the nature of the event was difficult to distinguish from an earthquake. Fortunately, discrimination for events such as this is often very clear, provided high-quality regional data is available.

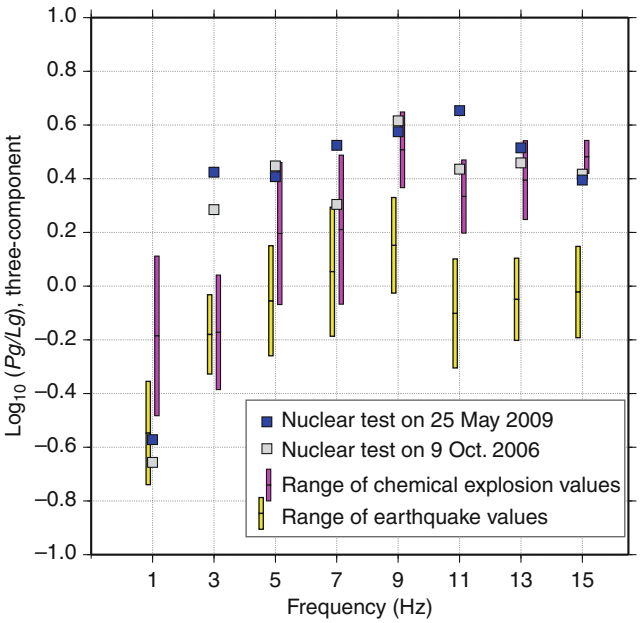
In this analysis, the original seismograms from station MDJ, located in China, are filtered in eight narrow frequency bands as illustrated in blue in Figure 1, but this time with bands centered on each of the frequencies from 1, 3, 5, 7, 9, 11, 13, to 15 Hz as indicated for the horizontal axis in Figure 4. The amplitudes of the *Pg* and *Lg* waves are measured in each narrow band, the amplitude ratio is formed (the "spectral ratio"), and the quantitative comparison can begin. Figure 4 shows how this ratio varies with frequency for the set of eight earthquakes, and for the set of four small chemical explosions. The ratio differs for these two populations as frequency rises, and the separation between them is very clear at high frequencies (from 9 to 15 Hz in this case). It is also clear that the spectral ratios of the signal recorded for the events of 2006 and 2009 are like those of the known chemical explosions.

This successful seismic discriminant based upon regional waves is important in enabling monitoring capability to be extended down to lower magnitudes. In practice, there is often very little difference between the magnitude thresholds for detection (at enough stations to enable a useful location estimate), and identification, since so many regions of the Earth are now monitored to low magnitude for earthquakes as part of investigations into seismic hazard. It may take only one regional seismogram to enable discrimination to be carried out with high confidence (provided the recording is of adequate quality, and is for a station that has an archive of signals from previous known earthquakes and explosions).

Along with the use of regional seismic waves and their spectral ratios at 5 Hz and higher, another discriminant turning out to be successful at distinguishing between



Seismic Monitoring of Nuclear Explosions, Figure 3 Typical vertical-component records from an earthquake and an explosion. Traces plotted are: unfiltered (*top*), low-frequency bandpass filtered (*middle*), and high-frequency bandpass filtered (*bottom*). (From Kim et al., 1993.)



Seismic Monitoring of Nuclear Explosions, Figure 4 Spectral ratios are shown, for the two nuclear explosions carried out by North Korea in 2006 and 2009, as measured from waveforms recorded at station MDJ in China (distance, about 370 km). They are compared with these ratios for a small group of earthquakes, and another group of chemical explosions, all in the vicinity of North Korea's nuclear test site. Colored bars represent ± 1 standard deviation in the ratios for chemical explosions (*magenta*), and small earthquakes (*yellow*). The spectral ratios for events in North Korea on October 9, 2006, and on May 25, 2009, are both explosion-like. (Courtesy of Won-Young Kim.)

earthquakes and explosions is the use of observed seismic waveforms to make estimates of the set of forces that appear to be acting at the seismic source. The set of forces here is quantified by what seismologists call the *moment tensor*. As shown by Ford et al. (2009) from study of numerous earthquakes and underground explosions, seismic events separate into specific populations as determined by the way their moment tensors behave – whether they are more representative of the all-around (isotropic) features of an explosion, or of the type of shearing motions more typical of an earthquake.

In general for underground tests, seismic data alone cannot distinguish between nuclear explosions, and chemical explosions in which all the material making up the explosive is fired within less than about a tenth of a second. But such chemical explosions, if large, are very rare. In the case of the two North Korea tests, both of which were announced as nuclear, objective evidence for the nuclear nature of the 2006 explosion came from several different detections of radionuclides that are diagnostic of a nuclear explosion. Such radionuclides were not detected from the 2009 explosion, which, however, was so large as to be implausible as a chemical explosion, since it would have to have consisted of literally thousands of tons of explosives.

Yield estimation

Yield estimation was of particular importance in the years following 1974 when a bilateral treaty between the USA and the USSR was negotiated, intended to go into effect in 1976. This was the Threshold Test Ban Treaty (TTBT), limiting the size of underground nuclear explosions

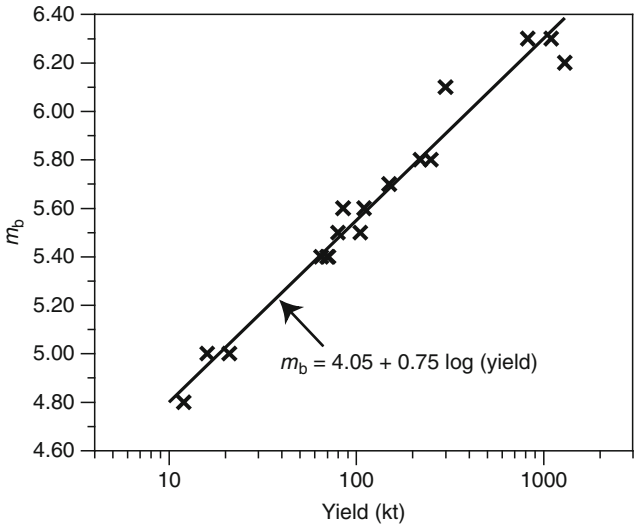
conducted by these two countries to a yield of not more that 150 kt. The TTBT proved contentious, with each side sending the other several inquiries asserting that the agreed-upon limits had possibly been exceeded (Timerbaev, undated). But this treaty was finally ratified in 1990, and has become less important since the CTBT was finalized and a nuclear testing moratorium by the signatory countries began in 1996. Yield estimation is however still important as an exercise in the interpretation of signals from the few underground explosions since that date, specifically those of India and Pakistan in 1998, and of North Korea in 2006 and 2009.

For a few tens of underground nuclear explosions, most of them at the Nevada Test Site, the yield has been announced by the agency conducting the test. It has therefore been possible to calibrate observed seismic magnitudes for these tests against the announced yields, and an example is given in Figure 3 using m_b values and yields reported for Nevada explosions in tuff and rhyolite.

The line $m_b = 4.05 + 0.75$ log (Yield) fits the data well (yield in kilotons). Such a calibration curve can be applied to obtain a seismic yield estimate for Nevada explosions with unannounced yield. But it requires correction, prior to its use in obtaining a seismic yield estimate for an explosion at a different site. This must be done, to allow for physical and geological differences between the sites. For example, in different rock types there can be different efficiencies in the coupling of nuclear yield into seismic energy; and differences in the propagation efficiencies as seismic waves travel out from the source of interest, as compared to seismic signals from a Nevada explosion. In this connection, it is of interest to note m_b and yield for the US nuclear explosion LONGSHOT (conducted in 1965 in the volcanic breccias of an Aleutian island). The m_b value is 5.9, corresponding to a yield of about 300 kt. if the Nevada curve of Figure 5 is applied directly. But the announced yield for LONGSHOT is 80 kt. One way to obtain a calibration curve for the Aleutians is therefore to add a correction of about 0.4 m_b units to the Nevada values of m_b at a given yield, before the curve of Figure 5 is used to supply a seismic yield estimate in this new location. This m_b correction, for a site differing from that where a calibration curve is directly available, is called the *bias*. If the bias correction is not applied, then a Nevada magnitude–yield curve can give too high a seismic yield estimate for a non-Nevada explosion.

Note that the Nevada Test Site is in a region of active tectonics, with significant episodes of volcanism in the last few million years, resulting in high temperatures within the upper mantle, and thus anomalous attenuation of seismic waves propagating through the hot and partially molten upper layers of the Earth, 100 or 200 km in thickness beneath the Nevada Test Site. Such propagation through an attenuating medium is presumed to be a contributing cause of bias.

The existence of m_b bias has long been known in seismology in connection with what is called "station bias." By this term is meant the systematic difference between



Seismic Monitoring of Nuclear Explosions, Figure 5 Seismic magnitude m_b vs. announced yield, for 17 Nevada Test Site nuclear explosions in tuff and rhyolite. The straight line here, which fits the data quite well, can be used to make a yield estimate of other events at this test site, in similar rock, if the seismic magnitude is known. (Data from Nuttli, 1986.)

mean m_b values (obtained for a particular seismic event by averaging reported m_b from seismometers all over the globe), and m_b reported by just one station. For example, the station BMO in Oregon (another region of active tectonism) has reported m_b values that for a given earthquake are typically about 0.3 units below the global average; and station KJN in Finland (in a stable shield region) reports values about 0.15 m_b units higher than the average. Their station bias values are thus −0.3 and +0.15, respectively. Station bias values commonly range over ±0.4 m_b units, so it may be expected that source region bias (which is what must be applied when a standard m_b – yield curve is used for different source regions) will also range over about 0.8 m_b units.

The nuclear weapons test site of the USSR that conducted the most underground nuclear explosions was near the city of Semipalatinsk, in northeastern Kazakhstan. Several multi-megaton underground explosions were conducted on Russia's Novaya Zemlya island test site, far to the north of continental Eurasia (see Khalturin et al., 2005). But these were all prior to the intended date of entry-into-force of the TTBT (March 1976). After that date, the magnitude of the largest underground tests at Semipalatinsk rose higher and higher over several years, with some magnitudes exceeding 6.1. Such magnitudes, according to the Nevada Test Site formula discussed above, $m_b = 4.05 + 0.75$ log (Yield), implied yields great than 500 kt, far in excess of the TTBT limit (150 kt). Intensive discussion in political and technical arenas ensued with stronger and stronger evidence accumulating to indicate a substantial test site bias between the Nevada and Semipalatinsk test Sites. For example, it

was of great interest that teleseismic signals from the largest underground explosions from these two tests, if recorded at the same station in a shield region, looked significantly different. The teleseismic *P*-wave from a large underground explosion at the site in Kazakhstan would routinely have frequency content at the 5 Hz level and sometimes higher (Der et al., 1985). The signal from Nevada would not contain such high frequencies. It was as if the signal from Nevada had passed through some type of filter, which of course would reduce its amplitude. Correcting for that effect would mean that the appropriate relation between magnitude and yield for an underground nuclear explosion at Semipalatinsk had the form

$$m_b = 4.05 + \text{bias} + 0.75 \log (\text{Yield}),$$

and Ringdal et al. (1992) and Murphy (1996) among many others concluded that the appropriate formula relating teleseismic *P*-wave magnitude and yield at Semipalatinsk should be this equation with a bias of 0.4. Support for this conclusion came from many arguments (see Richards, 1988 for a review). But in the political realm, the most persuasive was the very practical one associated with a Joint Verification Experiment of September 14, 1988, in which a team from the USA at the Semipalatinsk Test Site was allowed to make close-in measurements (within a few tens of meters) of a large Soviet underground nuclear explosion, in particular of the speed and extent of the shock wave it sent out into rock near the source at that test site. From such shock measurements, a reliable non-seismic method provided an accurate yield estimate (it was in the range 100–150 kt). Stations around the world provided measurements teleseismically, giving a seismic magnitude around 6.1 – comparable with the largest magnitudes of Semipalatinsk explosions since 1976, indicating that they too had been conducted in a way that respected the 150 kt limit of the TTBT. A reciprocal Joint Verification Experiment had been conducted at the Nevada Test Site, on August 17, 1988 with a Russian team making its own close-in measurements of the shock wave from a large US underground nuclear test intended to be in the range 100–150 kt. According to many news reports, the yield of this explosion slightly exceeded 150 kt. Timerbaev (undated) and news reports give it as 180 kt.

Problem events

The work of monitoring – for both earthquakes and explosions – is done in practice by hundreds of professionals who process the vast majority of seismic events routinely, and who also look out for the occasional events that, in the context of monitoring for the possibility of underground nuclear explosions, exhibit interesting characteristics, and which may then become the subject of special study.

These special events have stimulated the development of effective new discrimination techniques and a better appreciation of overall monitoring capability. Examples include a mine collapse in 1989 in Germany and two such collapses in 1995, in the Urals (Russia) and in Wyoming (USA); a small earthquake of magnitude 3.5 and its smaller aftershock in 1997 beneath the Kara Sea near Russia's former nuclear test site on Novaya Zemlya; and two underwater explosions in 2000 associated with the loss of a Russian submarine in the Barents Sea; the series of nuclear explosions carried out by India and Pakistan in 1998; and the nuclear tests conducted by North Korea in 2006 and 2009.

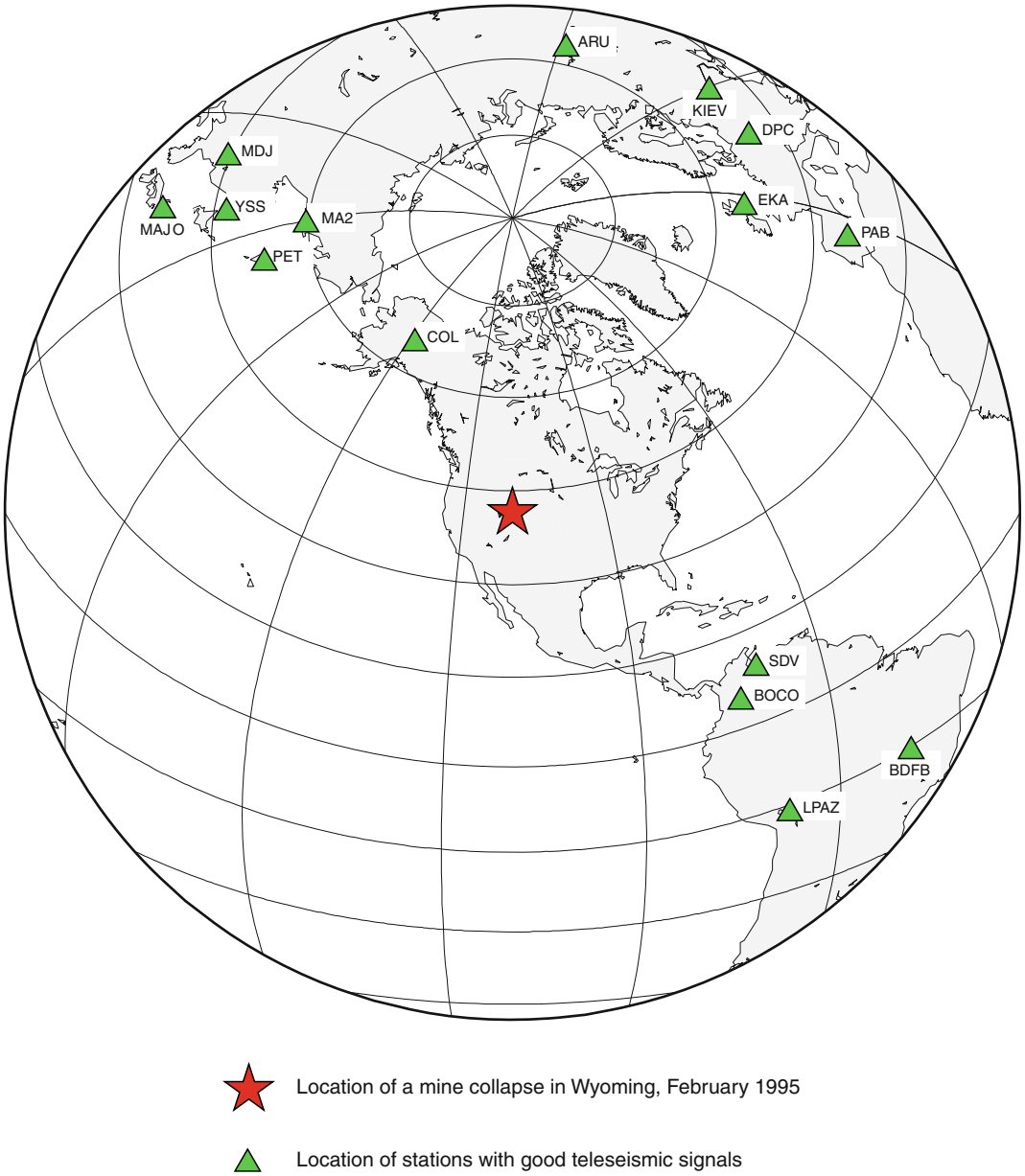
The mining collapses were seismically detected all over the world. For example, stations that detected the Wyoming event of 1995 are indicated in Figure 6. Mining collapses such as these have caused concern because their mix of surface waves and body waves as recorded teleseismically can appear explosion like using the classical M_s: m_b discriminant, as shown in Figure 2. But a careful analysis of regional and teleseismic waves from these events has showed that although the surface waves were quite weak, and in this respect seemed explosion like, they had the wrong sign. Therefore the motion at the source was *implosive* (the ground had moved inward toward the source) rather than *explosive*. Indeed, mining collapses are an implosion phenomenon, and it was important to learn that their implosive nature could be reliably determined from seismic recordings. Teleseismic waveforms from the Wyoming mine collapse are shown in Figure 7. This is an example of the use of what seismologists call the "first motion" of the *P*-wave, which is clearly downward in these data.

The Kara Sea earthquake was too small to apply the M_s: m_b discriminant (the surface waves were too small to measure reliably). This event showed the importance of accurate locations, and of using spectral ratios of regionally recorded *P*-waves and *S*-waves to discriminate small events (Richards and Kim, 1997).

As we have discussed earlier, the North Korea nuclear test of 2006 was of interest as an example of a nuclear explosion that was promptly detected globally, though its yield has been estimated at less than 1 kt. This event required regional seismic data in order to determine that indeed an explosion had been carried out and that the signals were not from an earthquake. Subsequently, xenon radionuclides were detected that decisively identified the explosion as nuclear.

Evasion

Several methods have been proposed, by which underground explosions might be concealed. One method is simply to make them small enough; but then there would be relatively little to learn, from the point of view of a weapons designer. The more important methods are those which combine as many features as possible, designed to reduce seismic signal-to-noise ratios at all relevant monitoring stations. Proposed methods include: emplacement of the nuclear device in material such as dry alluvium, to reduce the coupling of explosion energy into seismic signal (but that method is likely to result in leakage of detectable radioactivity); waiting until a sufficiently large natural earthquake occurs fairly near

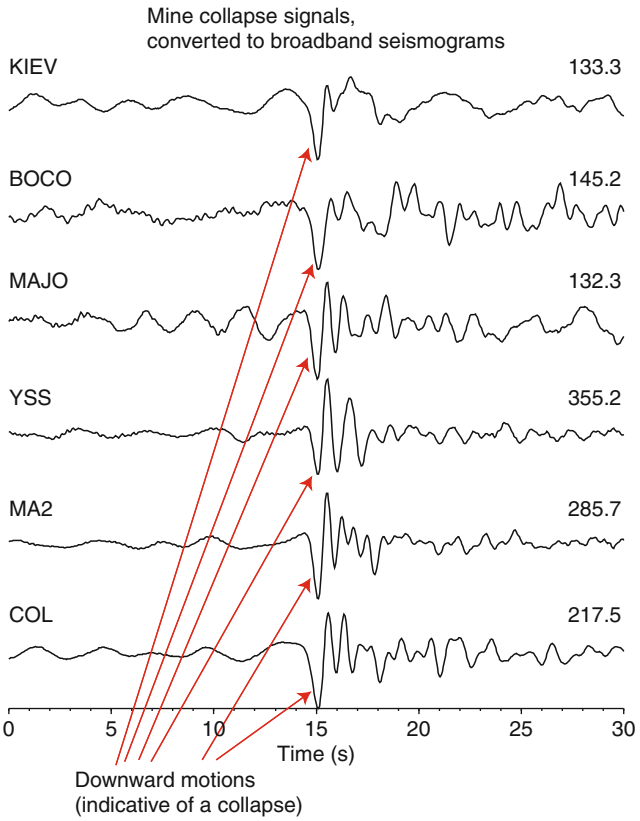


Seismic Monitoring of Nuclear Explosions, Figure 6 A global map showing stations recording teleseismic *P*-waves from a mine collapse in Wyoming. Its body wave magnitude was 5.3, and surface wave magnitude was 3.7. This combination is explosion like, as shown in Figure 2. (From Bowers and Walter, 2002.)

a test site (which presents the formidable challenge of identifying the event within a few minutes of its occurrence as large enough, and then within a couple of minutes executing the weapons test so that its seismic signals would hopefully be swamped by the large and prolonged signals from the earthquake); and setting off a sequence of several explosions that are designed to simulate a natural earthquake signal.

Careful study of each of these methods indicates that they are relatively ineffective in comparison with the methods known as cavity decoupling and mine masking, which we next discuss, and which are widely regarded as setting the practical levels down to which seismic monitoring of nuclear explosions is possible.

When an underground explosive device is tightly packed into its hole ("tamped" or "fully coupled"), and is detonated at sufficient depth to contain all radioactive products, a shock wave travels some distance from the shot-point out into the surrounding rock at speeds that exceed the normal *P*-wave speed. This nonlinear phenomenon reduces at sufficient distance from the shot-point, and thereafter the wave propagation can be regarded as



Seismic Monitoring of Nuclear Explosions, Figure 7 The *P*-wave signals of the Wyoming mine collapse are shown at six teleseismic stations, processed to bring out the fact that the first motion of the ground at these stations is downward, indicative of an implosion rather than explosion (for which the first motion would be upward). (From Bowers and Walter, 2002.)

elastic. The so-called "elastic radius" for a tamped explosion, i.e., the radius beyond which wave propagation is linear, is roughly 100 meters times the cube root of the yield (in kilotons).

If the explosion is set off inside a large underground cavity instead of being tamped, then the shock wave set up in the rock can be weakened or even eliminated, in which case only elastic waves are radiated. The explosion is said to be fully decoupled if only elastic waves result, and theoretical work reported in 1959 has addressed the question of how much weaker the seismic signal might be made. Theoretical work has indicated that signals could thereby be reduced by factors in the range 50–100, compared to a tamped explosion. The cavity radius itself is the "elastic radius" for a fully decoupled shot. For salt, the cavity radius required for full decoupling has been estimated at about 25 m times the cube root of the yield (in kilotons). For hard rock, the cavity size for full decoupling is comparable; for weak salt it is somewhat greater. See Sykes (1996) for further discussion, and Denny and Goodman (1990) for estimates of the decoupling factor derived from the practical experience in 1966 of carrying out a small nuclear explosion (about 0.38 kt) in the cavity produced by a tamped shot of 5.3 kt conducted 2 years earlier in a Mississippi salt dome. They conclude that the amplitude reduction is about 70, at low frequencies, for salt. At frequencies that have conventionally been used for seismic monitoring, the seismic signal strength is proportional (very roughly) to the volume within the elastic radius. This volume is substantially reduced by fully decoupling, which is the reason why cavity decoupling has been proposed as offering the technical possibility of a clandestine program of nuclear testing. However, the signal strength is not nearly so strongly reduced, by decoupling, at frequencies above that associated with resonances of the internal surface at the elastic radius. In practice, the frequency above which decoupling is likely to be substantially less effective is around 10–20 Hz, divided by the cube root of the yield (in kilotons). The overall effect on the seismic signals from a fully decoupled shot of yield Y, given the results of Denny and Goodman, is to make these signals look like those from a tamped shot of yield $Y/70$.

A thorough discussion of decoupling as an evasion scenario would have to include several non-seismological considerations. These include: the military significance of being able to carry out nuclear tests up to various different yield levels (e.g., 0.1, 1, or 10 kt.); and the political consequences if a clandestine test program were uncovered. Technical considerations include methods of (clandestine) cavity construction, and the capabilities of non-seismological surveillance techniques. Leakage of radioactivity from an underground cavity would be a challenge, given that much of the energy of a decoupled explosion goes into pumping up the pressure in the cavity. While some assert that clandestine use of cavity decoupling would be so difficult to execute that it belongs to the realm of fantasy, others have been persuaded that the risk might indeed be manageable, and that estimates of concealable yields, under this evasion scenario, must be made. The NAS (2002) report describes ten "layers of difficulty" with successfully hiding an underground nuclear explosion via cavity decoupling, concluding that even a nation committing significant resources to this work could not have confidence in being able to get away with tests above 1 or 2 kt.

The evasion scenario known as mine masking hypothesizes the execution of an underground nuclear weapon test explosion in a mining region, concurrently with a large mine blast. Such blasts in a big commercial operation consist of hundreds of separate charges, fired in sequence to break and/or move as much rock as possible, in a procedure known as ripple-firing (Khalturin et al., 1998). Regardless of the logistical difficulties of such a scenario, estimates of the possibilities of concealment via this approach can come from taking examples of signals from large mine blasts, and signals from small underground nuclear explosions, then adding them together before subjecting them to the methods used to discriminate between various types of seismic events. What is

typically found, is that the maximum size of the identifiable waves (e.g., the *P*-waves) from the mine blast is about that expected from individual sub-blasts (commonly called "delays"), and these amplitudes are spread out over a longer time in seismograms.

A study of mine masking possibilities by Smith (1993) used several different examples of mine-blast seismograms together with single-fired explosion records, and found a number of features that could be used to identify a simultaneous shot within a ripple-fired blast. He concluded that to conceal a single-fired deep detonation (depth is required for containment of radionuclides), the single explosive shot should not exceed 10% of the total explosive.

The conclusion here is that mine blasts are not effective for concealing large releases of energy at the level associated with kiloton-scale nuclear weapons tests, unless the nuclear explosion were subject to efforts at decoupling. Again non-seismic considerations arise, including an assessment of the plausibility of carrying out a complicated decoupled and masked nuclear explosion at the same time and location as a large mine blast that would itself attract some level of monitoring attention – particularly if the seismic signals seemed unusual in comparison with those from prior blasting in the region.

Event detection capability of the international monitoring system

In 1976, a group of international scientists was established at the Conference on Disarmament in Geneva, for the study of monitoring technologies and data analysis methods in the context of supporting a future test ban treaty. This group of scientific experts (GSE) played an essential role in laying the scientific groundwork for the final stage of CTBT negotiations conducted from 1994 to 1996. Prior to the negotiation, GSE organized a series of technical tests – GSETT-1 in 1984, GSETT-2 in 1991, and GSETT-3 in 1995. These tests contributed significantly to the development of the international system being built today to support treaty verification.

The finalized sections of the CTBT include an extensive description of networks to monitor treaty compliance using hydroacoustic, infrasound, and radionuclide technologies as well as seismological methods. The CTBT Organization (CTBTO) operates an International Monitoring System specified in treaty text, as well as an International Data Centre to analyze signals sent via satellite to headquarters in Vienna. Extensive descriptive material on these networks is available online (see http://www.ctbto.org).

To implement the CTBT seismic monitoring system, a sequential four-step process is needed to build each station (CTBTO PrepComm, 2009): (1) Site survey, (2) Installation, (3) Certification, and (4) Operation. It must be demonstrated for IMS stations that data received at the International Data Centre (IDC) are *authentic*. This is achieved through a special digital "signature" embedded in the data flow from each station. The IMS station must be certified to ensure that all of its equipment, infrastructure, and settings meet the technical specifications set by the CTBTO, and to also ensure that all data are transmitted to the IDC through the Global Communication Infrastructure (GCI) in a timely manner.

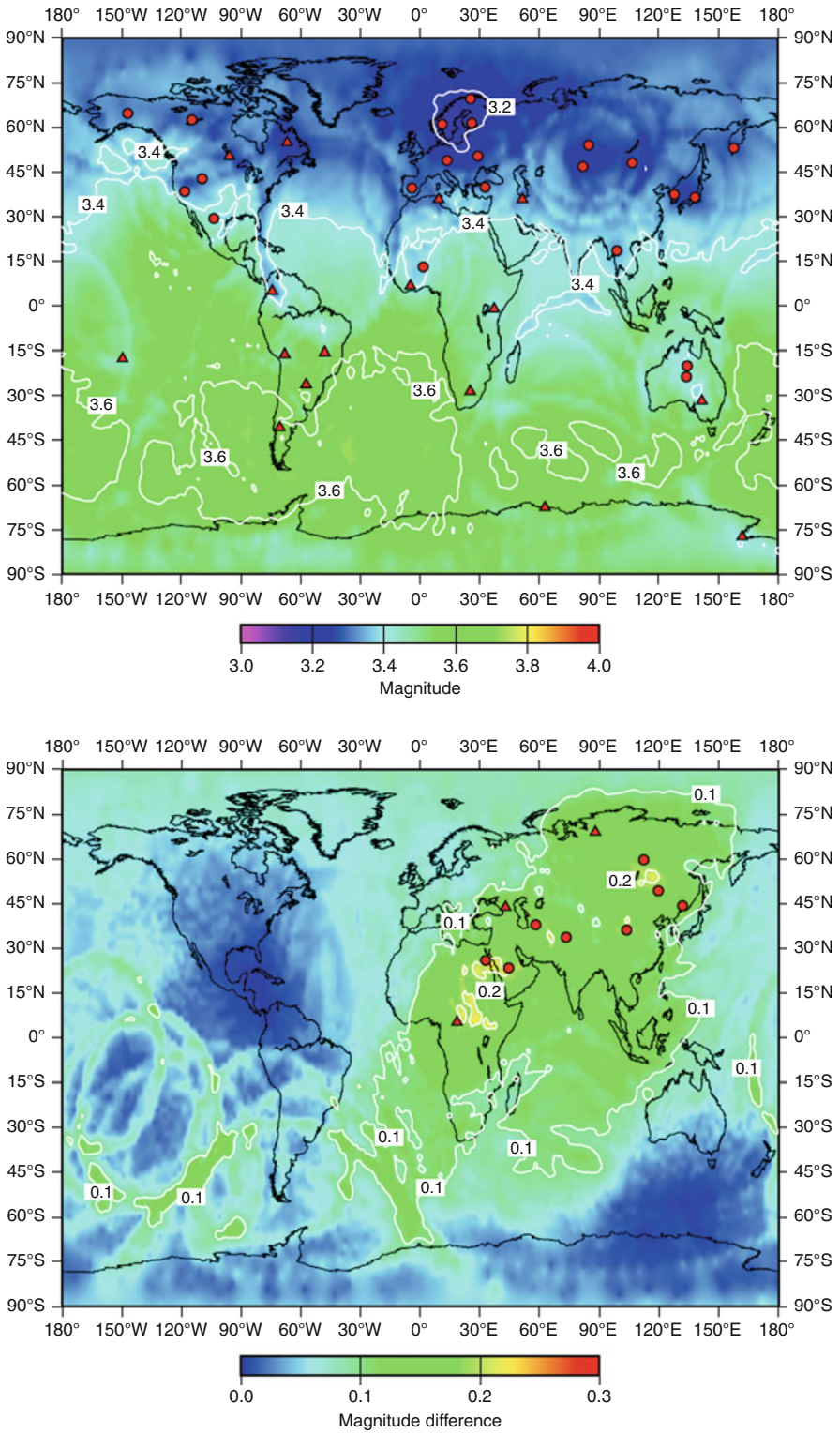
Here, we note that the primary seismographic network is to consist of 50 stations, many of them arrays; and that location estimates are based upon detection of signal at 3 stations or more. An auxiliary network of 120 continuously operating stations is to provide seismic waveform data, again via satellite, in order to help characterize the events detected by the primary network. Although these two networks are not completely built (as of 2010), there are enough stations operating to provide good indications of what the detection capability will be when all stations are installed and providing data.

Figure 8 shows maps of the detection capability of the primary seismic network of the IMS. The upper figure shows the actual capability of 38 operating stations based upon experience in the year 2007. The lower figure shows how much this capability is expected to improve when 11 additional stations are operational, most of them in Eurasia. Capability is expressed in terms of magnitude thresholds, above which 90% of the seismic events are expected to be detected at enough stations to provide a location estimate. The work of identifying events is left to member states. This work is not just a technical matter since it is a political act for one country to make an allegation that another country has committed a treaty violation. The evidence in support of such an allegation can come from the IMS and IDC, as well as from the National Technical Means of member states, and/or from a subset of the thousands of seismographic stations operated around the world for purposes not directly related to monitoring for nuclear explosions.

Summary

We have described the basic steps in monitoring nuclear explosions, and have emphasized the seismic monitoring system specified by the Comprehensive Nuclear Test Ban Treaty of 1996.

When the treaty was being negotiated, the goal for the International Monitoring System was that it be capable of detecting and identifying treaty violations – nuclear explosive tests – at the 1 kt level and higher, if they were not evasively tested. Recognizing that a 1 kt underground nuclear explosion has a magnitude in the range about 4–4.5, if it is conducted in the way that almost all the more than 1,500 prior underground nuclear explosions were carried out (i.e., well tamped and not with intent to reduce the signals picked up by monitoring networks), the evidence from Figure 8 is that this design capability has been significantly exceeded. For almost the entire northern hemisphere, including Eurasia and North America, capability is good down to about magnitude 3.3. This corresponds to a yield of less than 100 t (0.1 kt) for a well-tamped explosion in hard rock. Only time will tell whether this capability, combined with other monitoring assets, is deemed adequate to support entry into force of the CTBT.



Seismic Monitoring of Nuclear Explosions, Figure 8 Maps showing the detection capability of the IMS primary seismographic network. The upper figure shows the capability of the network in late 2007, with 38 stations sending data to the IDC. The capability is represented by the magnitude of the smallest seismic event that would be detected with a 90% probability by three stations or more. The lower figure shows the estimated improvement over this capability that could be achieved by bringing 11 of the remaining 12 primary seismic stations into operation. (From Kværna and Ringdal, 2009.)

Acronyms

CTBT–Comprehensive Test Ban Treaty or Comprehensive Nuclear-Test-Ban Treaty (its formal name)
CTBTO–CTBT Organization
IDC–International Data Centre (of the CTBTO)
IMS–International Monitoring System (of the CTBTO)
LTBT–Limited Test Ban Treaty
TTBT–Threshold Test Ban Treaty

Bibliography

Bennett, T. J., Oancea, V., Barker, B. W., Kung, Y.-L., Bahavar, M., Kohl, B. C., Murphy, J. R., and Bondár, I. K., 2010. The Nuclear Explosion Database (NEDB): a new database and web site for accessing nuclear explosion source information and waveforms. *Seismological Research Letters*, **81**, 12–25.

Bowers, D., and Walter, W. R., 2002. Discriminating between large mine collapses and explosions using teleseismic P Waves. *Pure and Applied Geophysics*, **159**, 803–830.

Bowers, D., and Selby, N. D., 2009. Forensic Seismology and the Comprehensive Nuclear-Test-Ban Treaty. *Annual Reviews of Earth and Planetary Sciences*, **37**, 209–236.

CTBTO PrepComm, 2009. Verification regime (available online at http://www.ctbto.org/verification-regime/).

Dahlman, O., Mykkeltveit, S., and Haak, H., 2009. *Nuclear Test Ban: Converting Political Visions to Reality*. Berlin: Springer.

Denny, M. D., and Goodman, M. D., 1990. A case study of the seismic source function: Salmon and Sterling reevaluated. *Journal of Geophysical Research*, **95**, 19705–19723.

Der, Z., Mcelfresh, T., Wagner, R., and Burnetti, J., 1985. Spectral characteristics of *P* waves from nuclear explosions and yield estimation. *Bulletin of the Seismological Society of America*, **75**, 379–390 (also erratum, **75**, 1222–1223).

Ford, S. R., Dreger, D. S., and Walter, W. R., 2009. Identifying isotropic events using a regional moment tensor inversion. *Journal of Geophysical Research*, **114**, B01306, doi:10.1029/2008JB005743.

Khalturin, V. I., Rautian, T. G., and Richards, P. G., 1998. The seismic signal strength of chemical explosions. *Bulletin of the Seismological Society of America*, **88**, 1511–1524.

Khalturin, V. I., Rautian, T. G., Richards, P. G., and Leith, W. S., 2005. A review of nuclear testing by the Soviet Union at Novaya Zemlya, 1955-1990. *Science & Global Security*, **13**, 1–42.

Kim, W.-Y., Simpson, D. W., and Richards, P. G., 1993. Discrimination of earthquakes and explosions in the Eastern United States using regional high-frequency data. *Geophysical Research Letters*, **20**, 1507–1510.

Koper, K. D., Herrmann, R. B., and Benz, H. M., 2008. Overview of open seismic data from the North Korean event of 9 October 2006. *Seismological Review Letters*, **79**, 178–185.

Kværna, T., and Ringdal, F., 2009. *Analysis of the IDC Reviewed Event Bulletin for Detection Capability Estimation of the IMS Primary and Auxiliary Seismic Stations*. Poster presented at the International Scientific Studies conference, June 10–12, 2009, Vienna, Austria.

Murphy, J. R., 1996. Types of seismic events and their source descriptions. In Husebye, E. S., and Dainty, A. M. (eds.), *Monitoring a Comprehensive Nuclear Test Ban Treaty. NATO ASI Series E*. Dordrecht: Kluwer Academic, Vol. 303, pp. 225–245.

Murphy, J. R., Rodi, W., Johnson, M., Sultanov, D. D., Bennett, T. J., Toksöz, M. N., Ovtchinnikov, V., Barker, V. W., Reiter, D. T., Rosca, A. C., and Shchukin, Y., 2005. Calibration of International Monitoring System (IMS) stations in Central and Eastern Asia for improved seismic event location. *Bulletin of the Seismological Society of America*, **95**, 1535–1560.

Myers, S. C., Begnaud, M. L., Ballard, S., Pasyanos, M. E., Phillips, W. S., Ramirez, A. L., Antolik, M. S., Hutchenson, K. D., Dwyer, J. J., Rowe, C. A., and Wagner, G. S., 2010. A crust and upper mantle model of Eurasia and North Africa for *Pn* travel time calculation. *Bulletin of the Seismological Society of America*, **100**, 640–656.

NAS, 2002. National Academy of Sciences report. *Technical Issues Related to the Comprehensive Nuclear-Test-Ban Treaty*, Washington, DC: National Academy Press.

Nuttli, O. W., 1986. Yield estimates of Nevada Test Site Explosions obtained from Seismic *Lg* Waves. *Journal of Geophysical Research*, **91**, 2137–2151.

OTA, 1988. Office of Technology Assessment, Congress of the United States, *Seismic Verification of Nuclear Testing Treaties*, OTA-ISC-361. Washington: U.S. Government Printing Office.

Richards, P. G., 1988. Seismic methods for verifying test ban treaties. In Schroeer, D., and Hafemeister, D. (eds.) Chapter 4 of *Nuclear Arms Technologies in the 1990's*. *AIP Conference Proceedings*. New York: American Institute of Physics, Vol. 178, pp. 54–108.

Richards, P. G., and Kim, W.-Y., 1997. Testing the nuclear test-ban treaty. *Nature*, **389**, 781–782.

Richards, P. G., Waldhauser, F., Schaff, D., and Kim, W.-Y., 2006. The applicability of modern methods of earthquake location. *Pure and Applied Geophysics*, **163**, 351–372.

Ringdal, F., 1985. Study of magnitudes, seismicity and earthquake detectability using a global network. In Kerr, A. U. (ed.) *The VELA Program: A Twenty-Five Year Review of Basic Research*. Defense Advanced Research Projects Agency, pp. 611–624.

Ringdal, F., Marshall, P. D., and Alewine, R. W., 1992. Seismic yield determination of soviet underground nuclear explosion at the shagan river test site. *Geophysical Journal International*, **109**, 65–77.

Smith, A. T., 1993. Discrimination of explosions from simultaneous mining blasts. *Bulletin of the Seismological Society of America*, **83**, 160–179.

Sykes, L. R., 1996. Dealing with decoupled nuclear explosions under a comprehensive test ban treaty. In Husebye, E. S., and Dainty, A. M. (eds.), *Monitoring a comprehensive nuclear test ban treaty*. Dordrecht: Kluwer Academic. NATO ASI series E, Vol. 303, pp. 247–293.

Timerbaev, R., undated. *On the "Threshold" Test Ban Treaties of 1974–76*. Paper available from the PIR Center as http://www.pircenter.org/kosdata/page_doc/p1650_1.pdf

Waldhauser, F., and Schaff, D. P., 2008. Large-scale relocation of two decades of Northern California seismicity using cross-correlation and double-difference methods. *Journal of Geophysical research*, **113**, doi:10.1029/2007JB005479.

Yang, X., North, R., Romney, C., and Richards, P. G., 2003. Worldwide nuclear explosions. In chapter 84 of *International Handbook of Earthquake and Engineering Seismology*, Part B, edited by Lee, W. H. K., Kanamori, H., Jennings, P., and Kisslinger, C. on behalf of the International Association of Seismology and Physics of the Earth's Interior, Academic, pp. 1595–1599.

Cross-references

Body Waves
Earthquake, Magnitude
Seismic Instrumentation
Seismological Networks
Seismology, Monitoring of CTBT
Surface Waves

SEISMIC NOISE

Dhananjay Kumar[1], Imtiaz Ahmed[2]
[1]Chevron, Houston, TX, USA
[2]BP, Houston, TX, USA

Definition

Seismic noise. Noise is the undesirable part of seismic data that is not signal, and signal is what fits our conceptual model.
Multiple. A seismic event that experiences more than one reflection in the subsurface.
SNR. Signal-to-noise ratio – is a measure of strength of signal compared to noise, and it is a measure of seismic data quality.
Active seismic. Seismic data recordings of artificial man-made sources such as dynamite and vibroseis in land surveys and air gun in marine surveys. Example: exploration and engineering seismology.
Passive seismic. Seismic data recordings of natural sources such as earthquake, solar waves, and ocean waves. Example: earthquake seismology.

Introduction

Seismic noise comprises all of the unwanted recorded energy that contaminates seismic data. A part of the seismic energy is considered noise if it does not fit the conceptual model of seismic signal. Seismic noise can be random or coherent. The identification of seismic noise depends on the type of data analysis and the type of data available – a part of data treated as noise in one application can be signal in another application. For example, the S-wave energy is generally considered noise in a P-wave processing project; vice versa, the P-wave energy is considered noise in an S-wave processing project.

Historically, only the traveltime information of the seismic data was used to study the subsurface. For example, the knowledge about the earth's deep interior was primarily derived from seismic traveltime information recorded during deep earthquakes. Also, only the traveltime information was used to derive the structural image of subsurface for exploration projects. As seismic technology has advanced and the appetite for understanding complicated geological features has increased, there has been a push toward technologies using the complete waveforms (amplitude and traveltime) in seismic analysis. Full waveform inversion technology (Tarantola, 1986) is one such example that uses the complete waveform to estimate properties of the subsurface from the seismic data. The success of such technologies is understandably very dependent on clean, noise-free seismic data.

It is essential that seismic data are carefully processed to derive high quality seismic images (see Yilmaz, 2001 for seismic data processing). One of the major challenges in seismic data processing is to separate noise from signal or to attenuate noise. In practice, noise cannot be completely attenuated and occasionally it is not even desired to attenuate noise but to use it as signal. The objective of the noise attenuation or noise separation process in seismic data processing is to enhance signal-to-noise ratio (SNR). There have been significant progress in data processing to improve SNR; advances have been made in random noise attenuation (Yilmaz, 2001) and coherent noise attenuation (see Weglein and Dragoset, 2005 for multiple attenuation methods).

Recently, there have been various efforts to use seismic noise as signal, for example: (1) using multiple reflected energy (multiples) in seismic migration and inversion to image subsurface; and, (2) using very low and very high frequency passive seismic signal for reservoir monitoring. In the following sections, we write brief descriptions about the types of seismic noise, the noise attenuation techniques, and how seismic noise can be useful.

Types of seismic noise

There are two types of seismic noise: random noise and coherent noise. In a multichannel seismic dataset, random noise does not correlate either with the neighboring channels (i.e., no spatial correlation) or along the same channel (i.e., no temporal correlation). Coherent noise, however, is part of the data that correlates spatially and/or temporally. Random noise is easier to attenuate during seismic data processing. Coherent noise is difficult to attenuate in processing; therefore, residual coherent noise can interfere with real signal and be misinterpreted as signal. The possible sources of these seismic noises can be placed under four categories: (1) ambient sources, (2) wave propagation related noise, (3) data acquisition related noise, and (4) data processing artifacts. The severity and types of noise can differ between marine and land acquisition environment.

Ambient noise is the noise from unwanted sources like wind, swell, power line, activities on nearby road, surface facility, marine activities like ships and marine animals, and other cultural noise. Ambient noise can be present in various forms on seismic data, such as linear features, localized very high amplitude response, and mono-frequency events. Ambient noise can be random noise and coherent noise.

Wave propagation related noise includes the surface waves, multiples, and geologic noise. Seismic response for an active source survey include primary reflection event (e.g., incident P-wave reflected back as P-wave), refraction, ground roll, mode converted event (e.g., incident P-wave reflected as S-wave), and several events reflected multiple times in the subsurface (multiples). In reflection seismology, we are interested in P-wave (and/or S-wave) primary reflections. All the other coherent energies recorded are considered noise. Multiples are considered the major source of coherent noise in the seismic experiment and are really difficult to attenuate. Wave propagation related noise also includes seismic response from unwanted geology (complicated stratigraphy,

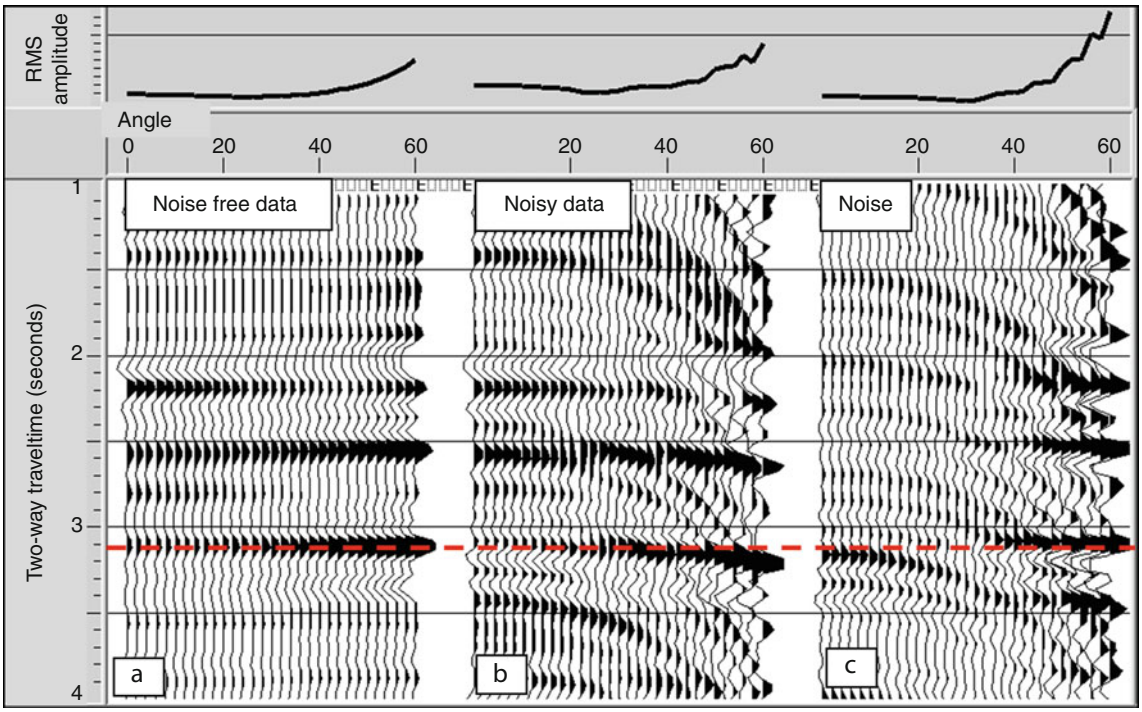
shallow gas, and faults) not considered in seismic analysis; they are known as geologic noise.

Data acquisition related noise is due to poor source and receiver coupling to the earth, source and recording instruments-generated noise, and acquisition footprint related to the acquisition design. Noise from poor coupling and noise related to instruments can be easily identified and attenuated. Acquisition footprint related to the acquisition design, observed as linear trend in seismic amplitude map view, is common in seismic data and can be suppressed in processing.

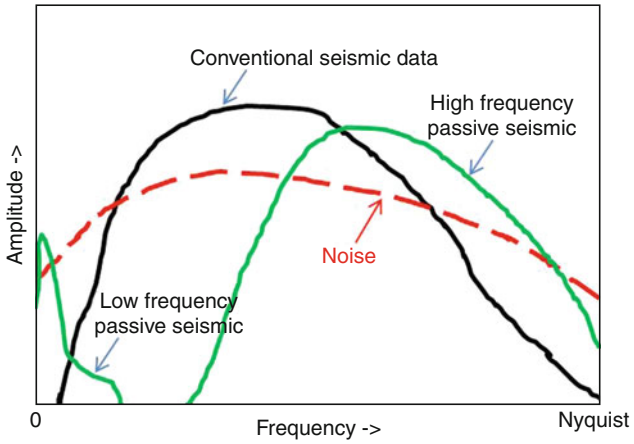
Seismic data processing is another source of noise in processed seismic data. There are various factors affecting seismic reflection amplitude that do not contain subsurface information (Sheriff, 1975), but it is impossible to correct for all the factors affecting amplitudes. Many approximations and assumptions are made in seismic data processing depending on computer resources availability, project timeline, understanding of the physics of wave propagation, and the type of seismic data available. Some of these factors may alter the data and may introduce noise in the recorded data. Some examples are: (1) noise introduced in seismic data due to poor multiple attenuation and poor normal move-out (NMO) correction (Figure 1); (2) noisy subsurface image due to imperfect velocity model used and/or approximate physics used for migration; (3) noise introduced due to inaccurate amplitude calibration of the raw amplitudes from the processed seismic data with synthetics seismic amplitude for quantitative AVO/AVA (amplitude variation with offset/angle) analysis; (4) artifacts introduced from the process of frequency enhancement to broaden frequency bandwidth for better depth resolution; and (5) artifacts introduced from the process of interpolation and regularization to compensate for the irregular acquisition geometry.

Enhancing signal over noise

There is always some noise present in the seismic data. Figure 2 schematically shows amplitude spectra of various seismic signals and noise, and noise is present in the whole signal bandwidth. Thus, the objective is to enhance SNR by noise attenuation or separation so that data can be effectively used for analysis. An important step in noise



Seismic Noise, Figure 1 Example of seismic noise. Shown in this figure is normal move-out (NMO) corrected synthetic seismic angle gather for AVA analysis with angle (in degrees) as *x*-axis and two-way traveltime (in seconds) as *y*-axis for: noise-free case (**a**), with noise (**b**), and the noise present in data (**c**). The response in the third panel (**c**) is the difference between response in the first panel (**a**) and the second panel (**b**). On the top of angle gather is the root mean square (RMS) amplitude value as a function of angle; the RMS value is computed in entire time gate shown here. In (c), the dipping coherent noises are multiples and the strong amplitudes at far angles are residual data after NMO. Thus, part of the data became residual data (un-flattened data after NMO in (b) is noise in AVA analysis) due to imperfect seismic velocity and/or seismic modeling method used in NMO (see red lines marked at 3,100 ms across three gathers for residual data).



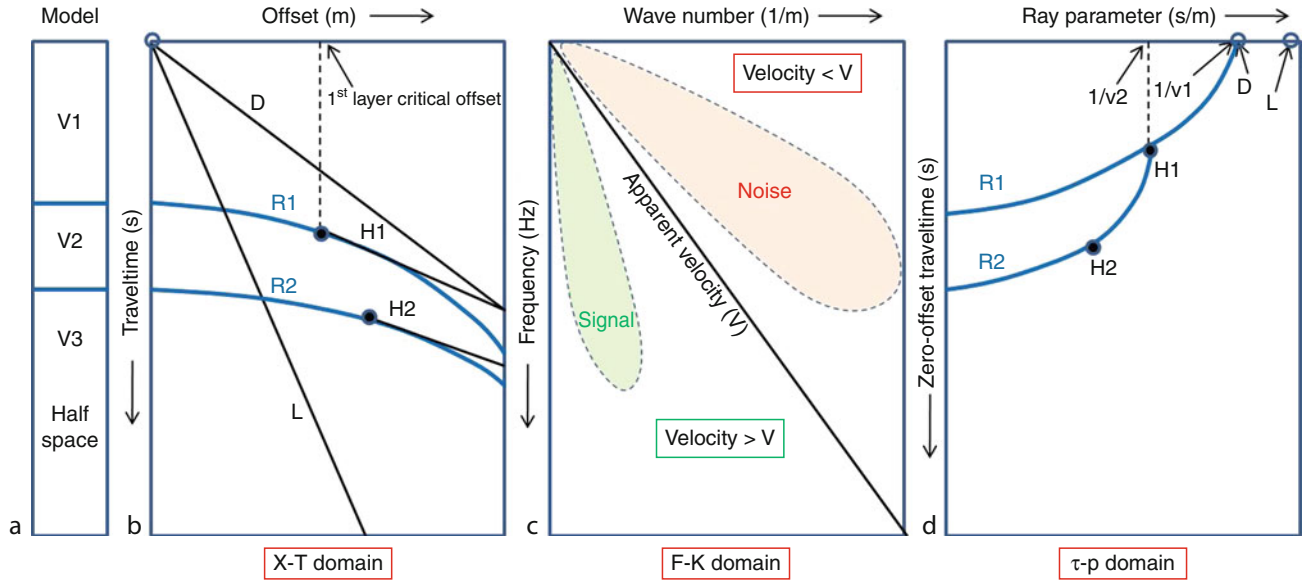
Seismic Noise, Figure 2 Schematic amplitude spectrum: for conventional broadband seismic data (active seismic for exploration range 10–60 Hz, and passive seismic for earthquake range from 10–100 Hz), low frequency passive seismic data (typically less than 10 Hz), high frequency passive seismic (microseismic) data (typically 30–300 Hz), and noise. Because noise is present at all frequencies, there is a need to do noise attenuation so that we get signal-to-noise ratio greater than 1. Spectrum is plotted for near zero frequency to Nyquist frequency. Nyquist frequency is the highest frequency without aliasing; beyond this frequency signal cannot be reliably estimated. Note that passive low and high frequency data can be treated as noise in broadband seismic data.

attenuation/separation is to identify signal from noise. White (1984) discusses spectral methods for signal and noise estimation on seismic reflection data. Kanasewich (1990) and Chapter 6 of Yilmaz (2001) have good discussion on noise and multiple attenuation methods with data examples. Noise level in data compared to signal can be estimated as the ratio of autocorrelation of data to a cross correlation of data. This is because autocorrelation of data represents both signal and noise but cross correlation of data with a representative signal trace (can be local partial stack of data) represents signal in data. Signal and noise separation process can be broadly divided into two methods: prediction based methods and separation based methods. By performing seismic simulation over a known (conceptual) model and comparing synthetics to the field seismic data, signal and noise can be identified on data, and therefore noise can be attenuated. Also, signal and noise representation differs in different data domains. Therefore, a suitable data domain can be identified for the optimum separation of the two. Figure 3 schematically shows various seismic events for a three-layer model in native acquisition (X-T) domain and transformed domains: (1) X-T domain (represents spatial-time domain), (2) F-K domain (represents frequency – wave number domain), and (3) τ-p domain (representing zero-offset traveltime – ray parameter domain). The key to successful noise attenuation is large separation between signal and noise in a specific data domain. Different types of noise may require different domains for better separation between signal and noise. For example, a linear event on a shot gather (X-T domain) maps to a radial line in the F-K domain and it maps onto a point in the τ-p domain, and thus can be rejected by F-K filtering and τ-p filtering, respectively. Caution should be observed while performing the forward and inverse transforms, as some implementations may not be completely reversible and may also introduce artifacts.

Random noise is not correlated and can be attenuated easily during data processing. One of the most robust methods to attenuate random noise on multichannel data by a factor of square root of N is by stacking (compositing) data from N channels. If prestack data is needed, however, F-X deconvolution (predictive deconvolution) is effective in random noise attenuation, where F-X corresponds to frequency – space domain. Deconvolution operation (Yilmaz, 2001) is designed in space domain to predict coherent signals, and then coherent signals can be subtracted from the seismic record.

Coherent noise is relatively more difficult to attenuate. An appropriate data domain (Figure 3) can be selected to distinguish signal from noise using the noise characteristics, like velocity and frequency. Radon filtering in τ-p domain and wavelet transform filtering in F-K domain is very effective method for coherent noise attenuation. In many field data cases, signal and noise are not well separated in any data domain and therefore it is very difficult to attenuate the noise. Recently a version of F-K filtering called Curvelet transform filtering has shown better results for both random and coherent noise attenuation (Neelamani et al., 2008). Multiples are example of coherent noise present in the data. There has been extensive research in the field of multiples suppression. The surface related multiple elimination (SRME) has been heavily relied upon to suppress multiples (Berkhout and Verschuur, 1997). This technique is based on the stationary phase theory of the seismic signal and the multiples are predicted by convolving traces with one another and then stacking these convolved traces.

Another way to have better SNR in seismic data is via improved seismic acquisition techniques. There have been various developments in seismic acquisition on land and in marine environment. For example, over/under marine seismic acquisition promises better SNR in both low and high frequency range (Moldoveanu et al., 2007). In over/under towed-streamer acquisition pairs of streamer are deployed at two different depths in the same vertical plane. Other advances in seismic acquisition promise cleaner and better seismic images, for example, Multicomponent Ocean Bottom recording, Wide Azimuth recording, and Simultaneous Sources recording. The multiples suppression is strongly dependent on the acquisition effort. For example, the suppression of multiples in the case of Wide Azimuth acquisition is the result of a natural weighting of the traces going into the stack



Seismic Noise, Figure 3 Seismic responses in various data domains for a three-layer model. The time domain earth model is shown in left (**a**) with seismic velocity for three layers as V1, V2, and V3. The second from left is the seismic response for a three-layer model in X-T domain (**b**), where D represents direct arrival, L represents linear noise, such as ground rolls, R1 and R2 are primary reflection events from 1st and 2nd layer interfaces, and H1 and H2 are refraction events. The third from left shows the seismic response in F-K domain (**c**), where noise (D, L events) is on the upper right corner with apparent velocity less than the reference apparent velocity (V) and signal (R1, R2, H1, H2 events) are on the lower left corner corresponding to lower wave numbers. The right most plot (**d**) shows the seismic response in the τ-p domain, where the linear event in X-T domain plots on a point in the τ-p domain (Diebold and Stoffa, 1981).

because of the areal nature of the acquisition (VerWest and Lin, 2007).

Some noise will always remain in the seismic data even after careful seismic processing. Also, there are some noises that are still not understandable and difficult to attenuate. Therefore, it is important to incorporate noise in the seismic data analysis – for example, in seismic inversion the data should not be over fitted as it might be fitting the noise.

Use of noise as signal

There have been significant efforts recently to use all of the recorded energy in the seismic data. Traditionally, only the primary reflection wavefield is used in reflection seismic imaging, while the multiple/ghost is discarded. However, the ghost/multiples can be used as signal, for example, as an input to mirror migration (Verm, 1987) to produce superior shallow images for the sparse receiver ocean bottom data where illumination holes deteriorate the primary images (Clarke et al., 2006; Dash et al., 2009). Note that multiples are also generated from the same source as primary, they travel longer paths and contain more information than primary, and therefore in some circumstances multiples can be more useful than primaries. Full waveform inversion is another technique that uses both the primary and the multiples to invert for the subsurface parameters more effectively.

Multiples have also been used to interpolate for missing near offsets in seismic recording using the technique of interferometry (Schuster, 2009). Seismic interferometry can also be used to extract primary signal from background noise and multiples by simply cross-correlating recorded data at two different stations (Sabra et al., 2005; Curtis et al., 2006; Draganov et al., 2007).

Some new seismic acquisition methods have been useful in using conventional multiples to improve SNR and/or extract primary. For example, over/under towed-streamer acquisition technology uses multiples to improve seismic data quality (Moldoveanu et al., 2007), and there is a possibility to estimate primaries from surface multiples from data recorded with simultaneous sources also called blended data (van Groenestijn and Verschuur, 2010). In the simultaneous source acquisition (Berkhout, 2008; Hampson et al., 2008) multiple sources are fired in a short time interval to speed up the acquisition.

Converted S-wave data, regarded as noise in reflected P-wave data imaging, has been quite successful in imaging gas reservoirs and areas where we have shallow gas anomalies (Tatham and McCormack, 1991). The property that S-wave does not get as attenuated as P-wave when the wavefield travels through these porous medium helps create better images through pure S-wave or converted S-wave imaging.

Low frequency earth's ambient passive response caused by natural phenomenon (such as wind, ocean waves, and human-made noise) and high frequency passive seismic response due to small earthquakes (microseisms) caused by induced fractures in a petroleum reservoirs are considered noise in a broadband active seismic data, but they can be effectively used to study subsurface. Low frequency earth's ambient noise can be used to extract signal (Draganov et al., 2007), and the low frequency passive seismic anomaly can be correlated with the presence of hydrocarbon (Saenger et al., 2009); however the research is still in its early stages. High frequency passive seismic data (also called microseismic data) are effectively used in hydraulic fracture monitoring (Warpinski, 2009) by locating microseisms induced by hydrocarbon production related activities.

Summary

Seismic noise is an integral part of the seismic record and is defined as all unwanted seismic energy on data. It can be divided into two categories: random and coherent noises. Random noise is not correlated among traces and is easier to attenuate compared to coherent noise that is spatially and/or temporally correlated. Multiples and geologic noise that are coherent noise are more difficult to attenuate and often interfere with seismic signal and makes seismic analysis challenging. Strategies for seismic noise attenuation are needed to preserve the seismic signal of interest and to improve signal-to-noise ratio (SNR). The success in noise attenuation lies in identification and then separation or prediction of signal and noise. Transformation of data to different data domains (X-T, F-K, τ-p, curvelet, wavelet domains) have helped in better separating noise from signal. Wave-equation extrapolations, inverse scattering methods, surface related multiple elimination, deconvolution, etc., model the noise and/or data in the process of noise attenuation. Advances in seismic data acquisition and processing have been made to improve SNR. Recently various efforts have been made to use noise as signal and are an active topic of research. This includes using multiples as well as primaries in seismic migration and inversion, and using low and high frequency passive seismic data in imaging subsurface.

Bibliography

Berkhout, A. J., 2008. Changing the mindset in seismic acquisition. *The Leading Edge*, **27**, 924–938.

Berkhout, A. J., and Verschuur, D. J., 1997. Estimation of multiple scattering by iterative inversion, Part I: theoretical considerations. *Geophysics*, **62**, 1586–1595.

Clarke, R., Xia, G., Kabir, N., Sirgue, L., and Mitchell, S., 2006. Case study: a large 3D wide azimuth ocean bottom node survey in deepwater GOM. In *76th Annual International Meeting, Society of Exploration Geophysicists*, Expanded Abstracts, 1128–1132.

Curtis, A., Gerstoft, P., Sato, H., Snieder, R., and Wapenaar, K., 2006. Seismic interferometry – turning noise into signal. *The Leading Edge*, **25**, 1082–1092.

Dash, R., Spence, G., Hyndman, R., Grion, S., Wang, Y., and Ronen, S., 2009. Wide-area imaging from OBS multiples. *Geophysics*, **74**(4), Q41–Q47.

Diebold, J., and Stoffa, P. L., 1981. The traveltime equation tau-p mapping and inversion of common midpoint data. *Geophysics*, **46**, 238–254.

Draganov, D., Wapenaar, K., Mulder, W., Singer, J., and Verdel, A., 2007. Retrieval of reflections from seismic background-noise measurements. *Geophysical Research Letters*, **34**, L04305, doi:10.1029/2006GL028735.

Hampson, G., Stefani, J., and Herkenhoff, F., 2008. Acquisition using simultaneous sources. *The Leading Edge*, **27**, 918–923.

Kanasewich, E. R., 1990. *Seismic Noise Attenuation. Handbook of Geophysical Exploration, Seismic Exploration*. New York: Pergamon, Vol. 7.

Moldoveanu, N., Combee, L., Egan, M., Hampson, G., Sudora, L., and Abriel, W., 2007. Over/under towed-streamer acquisition: a method to extend bandwidth to both higher and lower frequencies. *The Leading Edge*, **26**, 41–58.

Neelamani, R., Baumstein, A. I., Gillard, D. G., Hadidi, M. T., and Soroka, W., 2008. Coherent and random noise attenuation using the curvelet transformation. *The Leading Edge*, **27**, 240–248.

Sabra, K. G., Gerstoft, P., Roux, P., and Kuperman, W. A., 2005. Extracting time-domain Green's function estimates from ambient seismic noise. *Geophysical Research Letters*, **32**, L03310, doi:10.1029/2004GL021862.

Saenger, E. H., Schmalholz, S. M., Lambert, Marc-A, Nguyen, T. T., Torres, A., Metzger, S., Habiger, R. M., Muller, T., Rentsch, S., and Mendez-Hernandez, E., 2009. A passive seismic survey over a gas field: analysis of low-frequency anomalies. *Geophysics*, **74**(2), 029–040.

Schuster, G. T., 2009. *Seismic Interferometry.* UK: Cambridge University Press.

Sheriff, R. E., 1975. Factors affecting seismic amplitudes. *Geophysical Prospecting*, **23**, 125–138.

Tarantola, A., 1986. A strategy for nonlinear elastic inversion of seismic reflection data. *Geophysics*, **51**, 1893–1903.

Tatham, R. H., and McCormack, M. D., 1991. *Multicomponent seismology in petroleum exploration*. Tulsa: Society of Exploration Geophysicists.

van Groenestijn, G. J. A., and Verschuur, D. J., 2010. Using surface multiples to estimate primaries by sparse inversion from blended data. *Geophysical Prospecting*, doi:10.1111/j.1365-2478.2010.00894.x.

Verm, R., 1987. Imaging VSP's 3 kilometers beyond the borehole receiver. In *Offshore Technology Conference Proceedings*, Paper 5570.

VerWest, B. J., and Lin, D., 2007. Modeling the impact of wide-azimuth acquisition on subsalt imaging. *Geophysics*, **72**, 241–250.

Warpinski, N., 2009. Microseismic monitoring: inside and out. *Journal of Petroleum Technology*, **61**, 80–85.

Weglein, A. B., and Dragoset, W. H. (eds.), 2005. Multiple attenuation. *Geophysics Reprint Series No. 23, Society of Exploration Geophysicists*, Tulsa.

White, R. E., 1984. Signal and noise estimation from seismic reflection data using spectral coherence methods. *Proceedings of the IEEE*, **72**(10), 1340–1356.

Yilmaz, O., 2001. *Seismic data analysis: processing, inversion, and interpretation of seismic data*. Tulsa: Society of Exploration Geophysicists.

Cross-references

Earthquakes, Source Theory
Seismic Data Acquisition and Processing
Single and Multichannel Seismics

SEISMIC PHASE NAMES: IASPEI STANDARD

Dmitry A. Storchak[1], Johannes Schweitzer[2], Peter Bormann[3]
[1]International Seismological Centre (ISC), Thatcham, Berkshire, UK
[2]NORSAR, Kjeller, Norway
[3]Department 2: Physics of the Earth, GFZ German Research Center for Geosciences, Potsdam, Germany

Introduction

The working group (hereinafter WG) on the Standard Seismic Phase Names was set up by the IASPEI Commission on Seismological Observation and Interpretation (CoSOI) in 2001. The WG was chaired by D. A. Storchak, and included R. D. Adams, P. Bormann, E. R. Engdahl, J. Havskov, B. L. N. Kennett, and J. Schweitzer. The WG put together a modified standard nomenclature of seismic phases that was meant to be concise, consistent, and self-explanatory on the basis of agreed rules. The list was not meant to satisfy specific requirements of seismologists to name various phases used in a particular type of research. Instead, it was hoped that the new list would ensure an expanded standardized data reporting and exchange by data analysts and other users. After numerous consultations with the seismological community, the Standard Seismic Phase List was finalized and adopted by the CoSOI/IASPEI at its meeting in Sapporo on July 04, 2003.

The original list of standard seismic phase names was first published as part of the New Manual of Seismological Observatory Practice (Storchak et al., 2002) and then the version formally approved by the IASPEI was published in the Seismological Research Letters (Storchak et al., 2003). Various updates to the list were required due to progress in observational seismology and relevant changes in other observational standards. This article accommodates the advances made in the nomenclature since its last publication.

The new nomenclature partially modified and complemented an earlier one published in the old edition of the Manual of Seismological Observatory Practice (Willmore, 1979). It is more in tune with modern Earth and travel-time models. As opposed to former practice, the WG tried to make sure that the phase name generally reflects the type of wave and the path it has traveled. Accordingly, symbols for characterizing onset quality, polarity, etc., are no longer part of the phase name. The WG was also aware that seismic phases exist that are common in some regions but are only rarely or not found in other regions, such as Pb, PnPn, PbPb, etc.

The extended list of phase names as presented below reflects significantly increased detection capabilities of modern seismic sensors and sensor arrays, even of rather weak phases, which were rarely observed on the classical analog records. It also accounts for improved possibilities of proper phase identification by means of digital multi-channel data processing such as frequency-wave number (f-k) analysis and polarization filtering, by modeling the observations with synthetic seismograms or by showing on the records the theoretically predicted onset times of phases. Finally, the newly adopted IASPEI Seismic Format (ISF) (www.isc.ac.uk/doc/code/isf/isf.pdf) is much more flexible then the older formats previously used by the ISC, the NEIC, and other data centers. It also allows reporting, computer parsing, and archiving of phases with long or previously uncommon names. ISF also accepts complementary parameters such as onset quality, measured back azimuth and slowness, amplitudes and periods of other phases in addition to P and surface waves, for components other than vertical ones, and for instruments with nonstandard response characteristics.

This increased flexibility of the parameter-reporting format requires improved standardization, which limits the uncontrolled growth of incompatible and ambiguous parameter data. Therefore, the WG agreed on certain rules. They are outlined below prior to the listing of the standardized phase names. To facilitate the understanding of the phase names, ray diagrams are presented below. They have been calculated for local seismic sources on the basis of an average one-dimensional two-layer crustal model and for regional and teleseismic sources using the global 1D Earth model AK135 (Kennett et al., 1995).

Before assigning abbreviated shortcut seismic phase names, one should agree first on the language to be used and its rules. As in any other language, we need a suitable alphabet (here plain Latin letters), numbers (here Arabic numbers and +/− signs), an orthography, which regulates, for example, the use of capital and lower case letters, and a syntax, that describes the rules of correct order and mutual relationship of the language elements. One should be aware, however, that like any historically developed language, the seismological nomenclature will inevitably develop exceptions to the rules and depend on the context in which it is used. Although not fully documented below, some exceptions will be mentioned. Note that our efforts are mainly aimed at standardized names to be used in international data exchange so as to build up unique, unambiguous global databases for research. Many of the exceptions to the rules are related to specialized, mostly local research applications. The identification of related seismic phases often requires specialized procedures of data acquisition and processing that are not part of seismological routine data analysis. Also, many of these exceptional phases are rarely or never used in seismic event location, magnitude determination, source mechanism calculations, etc., which are the main tasks of international data centers. We focus, therefore, on phases that are particularly important for seismological data centers as well as for the refinement of regional and global Earth models on the basis of widely exchanged and accumulated parameter readings. In addition, we

added references to the first definition of some wave types and phase names.

Standard letters, signs, and syntax used for describing seismic phases

Capital letters

Individual capital letters that stand for primary types of seismic body waves include:

P:	Longitudinal wave that has traveled through the Earth's crust and mantle, from *undae primae* (Latin) = first waves (Borne, 1904)
K:	Longitudinal wave that has traveled through the Earth's outer core, K, from Kern (German) = core (Sohon, 1932; Bastings, 1934)
I:	Longitudinal wave that has traveled through the Earth's inner core (Jeffreys and Bullen, 1940)
S:	Transverse wave that has traveled through the Earth's crust and mantle, from *undae secundae* (Latin) = second waves (Borne, 1904)
T:	Wave that has partly traveled as sound wave in the sea, from *undae tertiae* (Latin) = third waves (Linehan, 1940)
J:	Transverse wave that has traveled through the Earth's inner core (Bullen, 1946)

Exceptions

- A capital letter N used in the nomenclature does not stand for a phase name but rather for the number of legs traveled (or N-1 reflections made) before reaching the station. N should usually follow the phase symbol to which it applies. For examples see syntax below.
- The lowercase letters p and s may stand, in the case of seismic events below the Earth's surface, for the relatively short *upgoing leg* of P or S waves, which continue, after reflection and possible conversion at the free surface, as downgoing P or S wave. Thus seismic depth phases (e.g., pP, sP, sS, pPP, sPP, pPKP, etc.) are uniquely defined. The identification and reporting of such phases is of utmost importance for source depth determination (Scrase, 1931; Stechschulte, 1932; Gutenberg et al., 1933; Macelwane et al., 1933).
- Many researchers working on detailed investigations of crustal and upper-mantle discontinuities denote both the up- and downgoing short legs of converted or multiply reflected P and S phases as lowercase letters p and s, respectively.

Individual or double capital letters that stand for surface waves include:

L:	(Relatively) long-period surface wave, unspecified, from *undae longae* (Latin) = long waves (Borne, 1904)
R:	Rayleigh waves (short- to very long-period waves in crust and upper mantle) (Angenheister, 1921)
Q:	Love waves, from Querwellen (German) = transverse waves (Angenheister, 1921)
G:	(Very long-period) global (mantle) Love waves, firstly observed and reported by Gutenberg and Richter (1934); in honor of Gutenberg, Byerly proposed the usage of G for these waves (Richter, 1958)
LR:	Long-period Rayleigh waves, usually relating to the Airy phase maximum in the surface wave train
LQ:	Long-period love waves

Lowercase letters and signs

Single lowercase letters generally specify the part of Earth's crust or upper mantle in which a phase has its turning point or at which discontinuity it has been reflected and/or eventually converted:

g:	Following the phase name characterizes waves "bottoming" (i.e., having their turning point in case of P or S body waves) or just travel (surface waves) within the upper ("granitic") Earth's crust (e.g., Pg, Sg; Rg), (Jeffreys, 1926)
b:	Following the phase name characterizes body waves bottoming in the lower ("basaltic") Earth's crust (Jeffreys, 1926) (e.g., Pb, Sb; alternative names for these phases are P*, S*, (Conrad, 1925))
n:	Following the phase name characterizes a P or S wave that is bottoming or traveling as head wave in the Earth's uppermost mantle (e.g., Pn, Sn), introduced after Andrija Mohorovičić discovered the Earth's crust and separated the crustal travel-time curve from the normal (=n) mantle phase (Mohorovičić, 1910)
m:	(Upward) reflections from the outer side of the Mohorovičić (Moho) discontinuity (e.g., PmP, SmS)
c:	Reflections from the outer side of the core-mantle boundary (CMB), usage proposed by James B. Macelwane (see Gutenberg, 1925)
i:	Reflections from the outer side of the inner core boundary (ICB)
z:	Reflections from a discontinuity (other than free surface, CMB or ICB) at depth z (measured in km). Upward reflections from the outer side of the discontinuity may additionally be complemented by a + sign (e.g., P410 + P; this, however, is not compulsory) while downward reflections from the inner side of the discontinuity must be complemented by a − sign (e.g., P660−P)

An exception from these rules is the use of lowercase p or s to indicate arrivals of longitudinal or transverse waves that were first radiated to go up toward the free surface to be reflected/converted back into the Earth as normal P or S waves (see near source surface reflections and conversions section of the phase list below).

Double lowercase letters following a capital letter phase name indicate the travel-time branch to which this phase belongs. Due to the geometry and velocity structure of the

Earth, the same type of seismic wave may develop a triplication of its travel-time curve with different, in some cases, well-separated, branches. Thus, it is customary to differentiate between different branches of core phases and their multiple reflections at the free surface or the CMB. Examples are PKPab, PKPbc, PKPdf, SKSac, SKKSac, etc. The separation of the different PKP branches with letters ab, bc, and df was introduced by Jeffreys and Bullen (1940).

Three lower case letters may follow a capital letter phase name to specify its character, e.g., as a forerunner (pre) to the main phase, caused by scattering (e.g., PKPpre) or as a diffracted wave extending the travel-time branch of the main phase into the outer core shadow (e.g., Pdif in the outer core shadow for P).

Syntax of generating complex phase names

Due to refraction, reflection, and conversion in the Earth, most phases have a complex path history before they reach the station. Accordingly, most phases cannot be described by a single capital letter code in a self-explanatory way. By combining the capital and lower case letters as mentioned above, one can describe the character of even rather complex refracted, reflected, or converted phases. The order of symbols (syntax) regulates the sequence of phase legs due to refraction, reflection, and conversion events in time (from left to right) and in space.

Examples for creating complex standard phase names

Traditional examples of complex phase names are as follows.

Refracted and converted refracted waves

- PKP is a pure refracted longitudinal wave. It has traveled the first part of its path as P through crust and mantle, the second through the outer core, and the third again as P through mantle and crust. An alternative name for PKP is P' (Angenheister, 1921), which should be read as "P prime."
- PKIKP (alternative to PKPdf) is also a pure refracted longitudinal wave. It has traveled the first part of its path as P through crust and mantle, the second through the outer core, the third through the inner core, and the fourth and fifth parts back again through outer core and mantle/crust.
- SKS is a converted refracted wave. It has traveled as a shear wave through crust and mantle, being converted into a longitudinal P wave when refracted into the outer core and converted back again into an S wave when entering the mantle.
- SKP or PKS are converted refracted waves with only one conversion from S to P when entering the core or from P to S when leaving the core, respectively.

Pure reflected waves

- In the case of (downward only) reflections at the free surface or from the inner side of the CMB, the phase symbol is just repeated, e.g., PP, SS (Geiger, 1909), PPP, SSS, KK, KKK, etc.
- In the case of (upward) reflections from the outer side of the Moho, the CMB, or the ICB, this is indicated by inserting symbols m, c, or i, respectively, between the phase symbols, e.g., PmP, PcP, ScS, PKiKP.
- Reflections from any other discontinuity in mantle or crust at depth z may be from the inner side (−; i.e., downward back into the mantle) or from the outer side (+; i.e., back toward the surface). To differentiate between these two possibilities, the sign has to follow z (or the respective number in km); for example, P410 + P or P660 − P.
- To abbreviate names of multi-leg phases due to repeated reflections, one can also write Phasename N. This type of abbreviation is customary in case of multiple phases with long phase names such as PmP2 for PmPPmP (free-surface reflection of PmP), SKS2 for SKSSKS (the alternative name for S'2, the free-surface reflection of SKS), PKP3 for PKPPKPPKP (double free-surface reflection of PKP; alternative name to P'3) or P4KP for PKKKKP (triple reflection of P at the inner side of the CMB).

Two additional notes are to be mentioned. First, PKP2 = PKPPKP are now alternative names for P'2 or P'P', respectively. This should not be mistaken for the old usage of PKP2 for PKPab. Secondly, in the case of multiple reflections from the inner side of the CMB, the WG followed the established tradition of placing the number N not after but in front of the related phase symbol K.

Reflected waves with conversion at the reflection point

In the case that a phase changes its character from P to S, or vice versa, one writes:

- PS (first leg P, second leg S) or SP (first leg P, second leg S) in the case of reflection/conversion from the free surface downward into the mantle (Geiger and Gutenberg, 1912a, 1912b).
- PmS or SmP, respectively, for reflection/conversion from the outer side of the Moho.
- PcS or ScP for reflection/conversion from the outer side of the CMB.
- Pz + S or Sz−P for reflection/conversion from the outer (+) side or inner (−) side, respectively, of a discontinuity at depth z. Note that the − is compulsory, the + is not.
- pS or sP reflection/conversion at the free surface for body waves with a direct upgoing first leg.

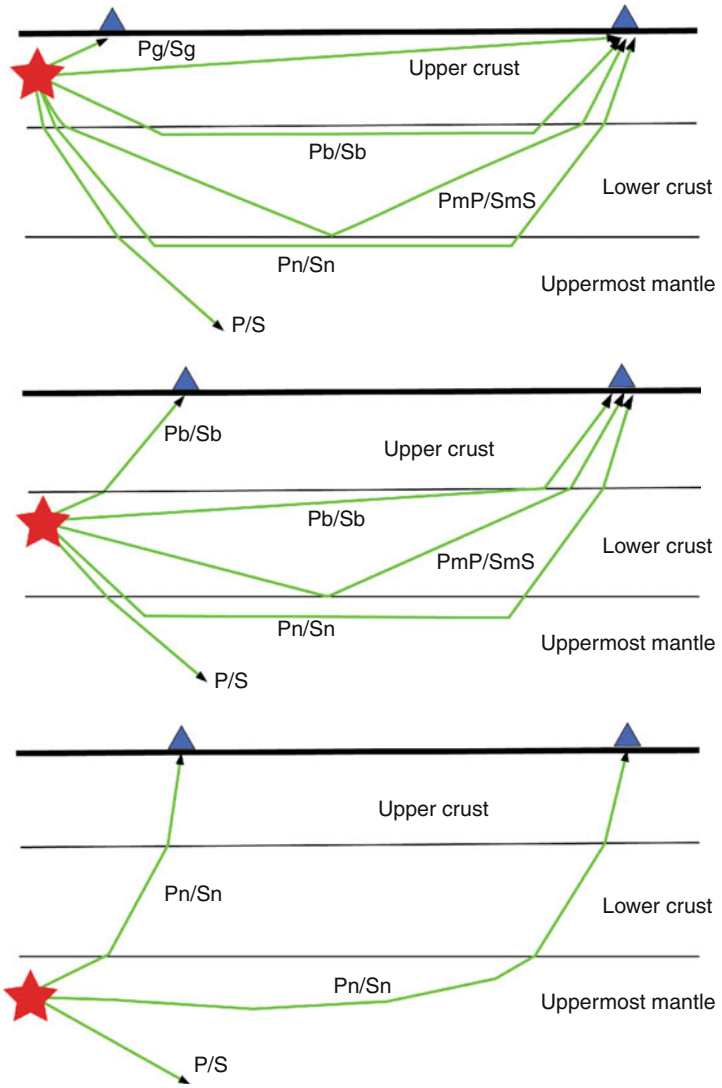
In this context, it is worth mentioning that mode conversion is impossible for reflections from the inner side of the CMB back into the outer core because the liquid outer core does not allow the propagation of S waves.

The WG determined the new IASPEI standard phase names along these lines and rules. Where these deviate from other traditionally used names, the latter are given

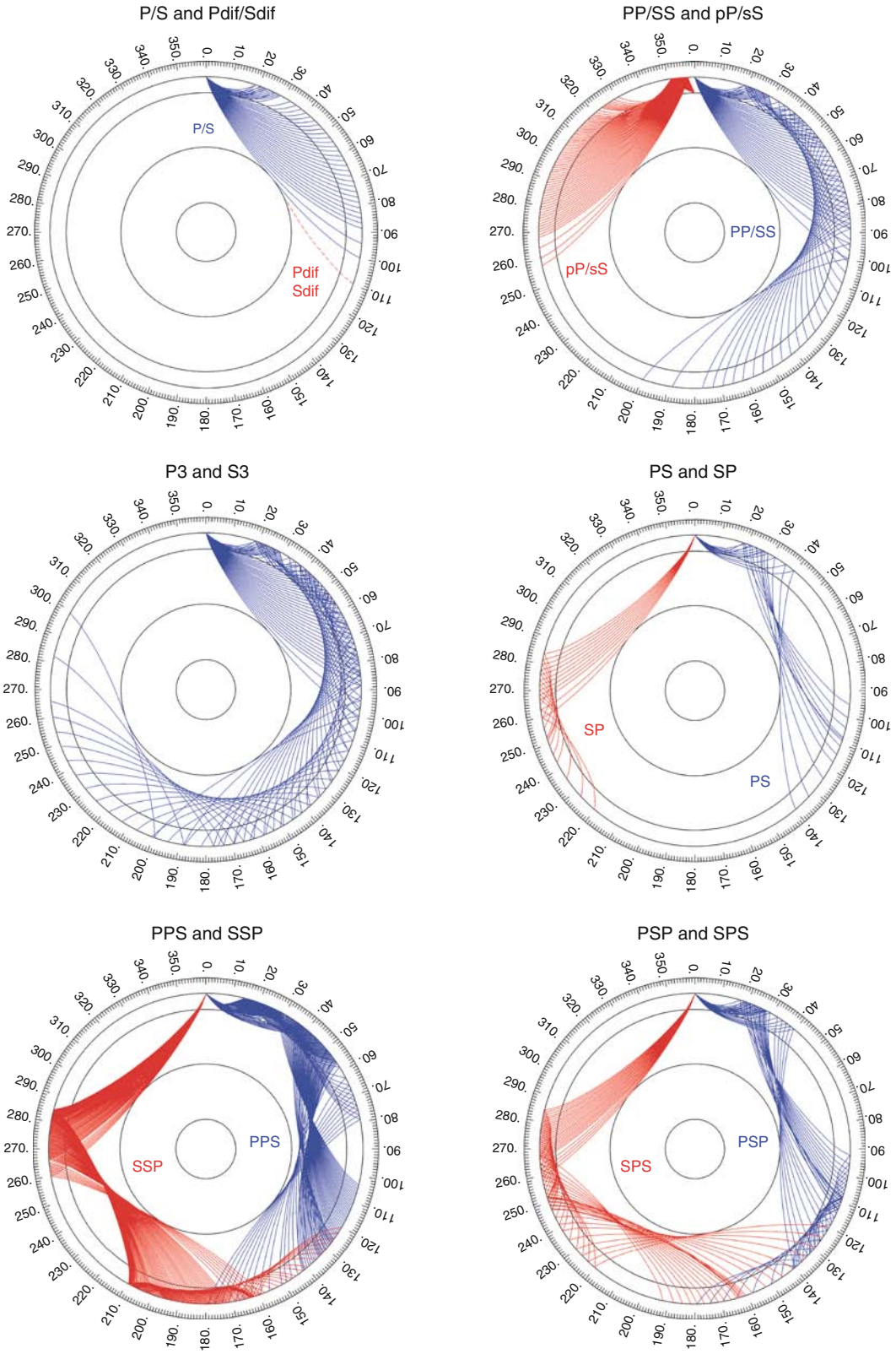
as well. Either the traditional names are still acceptable alternatives (alt) or they are old names (old), which should no longer be used.

Ray-path diagrams for some of the IASPEI standard phases

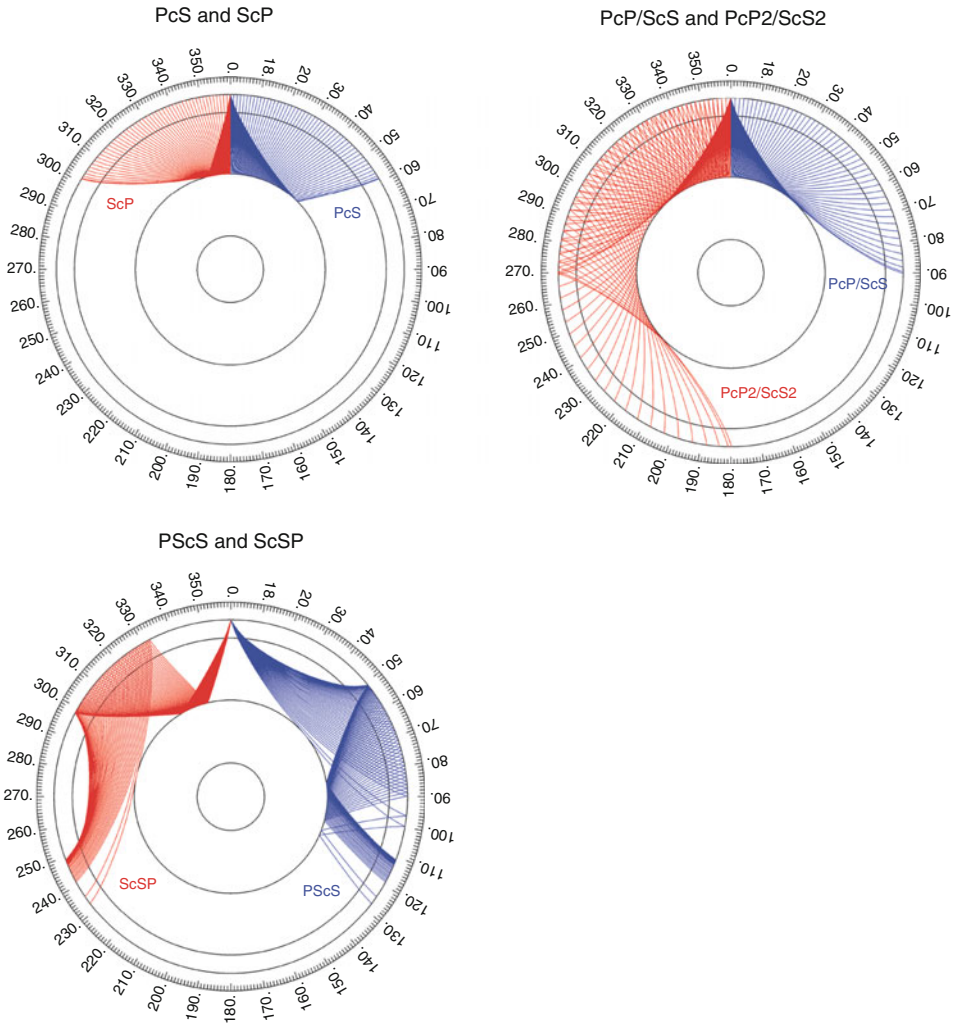
We show ray paths through the Earth for many of the mentioned phases. The three diagrams for crustal phases are sketches illustrating the principal ray paths in a two-layer crust (Figure 1). The rays in all other figures (Figures 2–6) were calculated by using the ray picture part of the WKBJ3 code (Chapman, 1978; Dey-Sarkar and Chapman, 1978); as the velocity model, we chose the standard Earth model AK135 (Kennett et al., 1995). For some types of P and S phases, the ray paths through the Earth are very similar because the velocity ratio V_P/V_S does not change enough to give very different ray pictures. In these cases, we calculated only the ray paths for the P-type ray (i.e., P, Pdif, pP, PP, P660P, P660−P, PcP, PcP2 and PcP4) and assume that the corresponding ray paths of the respective S-type phases are very similar. To show the different ray paths for phases with similar phase names, we show on many figures rays leaving the source once to the left and once to the right in different colors. The three most important discontinuities inside the Earth are indicated as black circles (i.e., the border between upper and lower mantle, the CMB, and the ICB).



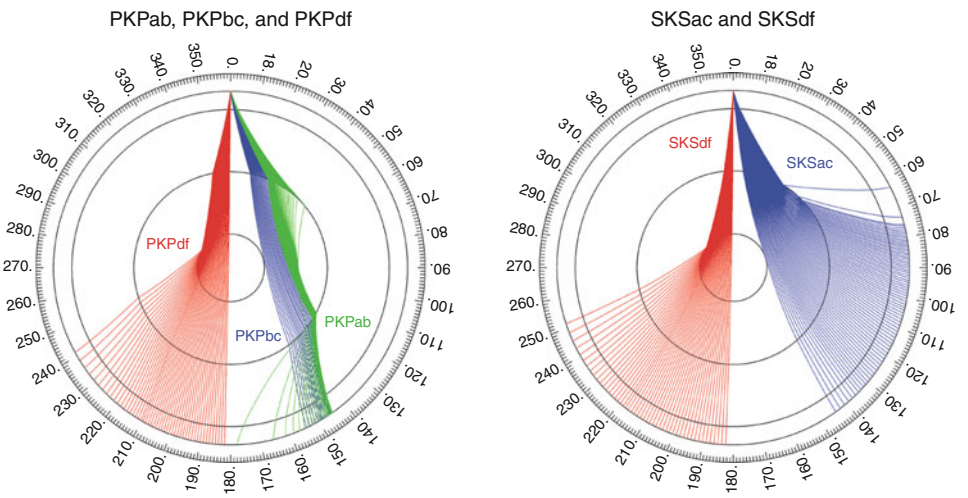
Seismic Phase Names: IASPEI Standard, Figure 1 Seismic "crustal phases" observed in the case of a two-layer crust in local and regional distance ranges ($0^\circ < D <$ approximately 20°) from the seismic source in the: (**a**) upper crust; (**b**) lower crust; and (**c**) uppermost mantle.



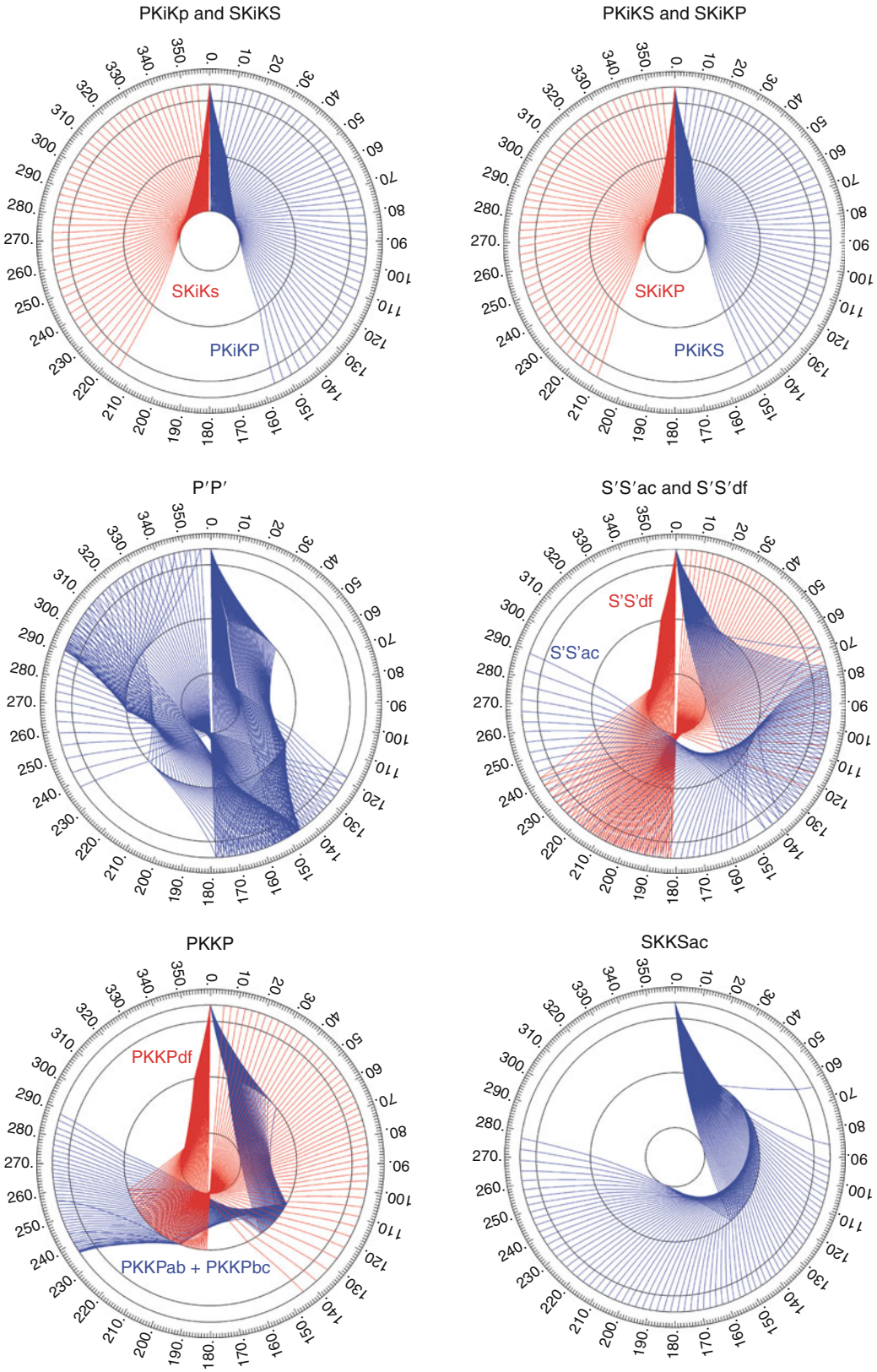
Seismic Phase Names: IASPEI Standard, Figure 2 Mantle phases observed at the teleseismic distances range (D > approximately 20°).



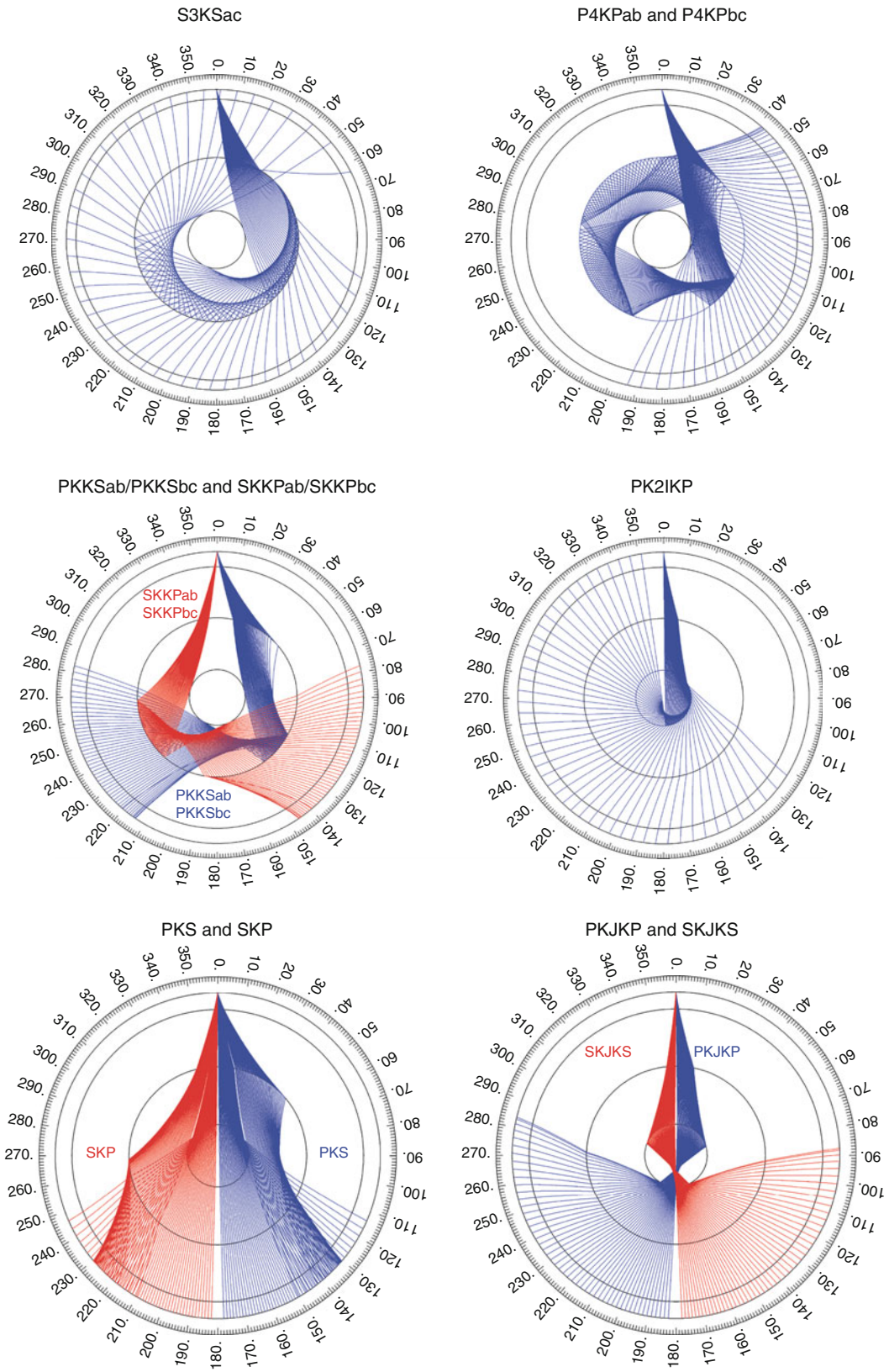
Seismic Phase Names: IASPEI Standard, Figure 3 Reflections from the Earth's core.



Seismic Phase Names: IASPEI Standard, Figure 4 Seismic rays of direct core phases.



Seismic Phase Names: IASPEI Standard, Figure 5 Seismic rays of single-reflected core phases.



Seismic Phase Names: IASPEI Standard, Figure 6 Seismic rays of multiple-reflected and converted core phases.

IASPEI standard seismic phase list

Crustal phases

Pg	At short distances, either an upgoing P wave from a source in the upper crust or a P wave bottoming in the upper crust. At larger distances also, arrivals caused by multiple P-wave reverberations inside the whole crust with a group velocity around 5.8 km/s
Pb	(alt:P*) Either an upgoing P wave from a source in the lower crust or a P wave bottoming in the lower crust
Pn	Any P wave bottoming in the uppermost mantle or an upgoing P wave from a source in the uppermost mantle
PnPn	Pn free-surface reflection
PgPg	Pg free-surface reflection
PmP	P reflection from the outer side of the Moho
PmPN	PmP multiple free-surface reflection; N is a positive integer. For example, PmP2 is PmPPmP
PmS	P to S reflection/conversion from the outer side of the Moho
Sg	At short distances, either an upgoing S wave from a source in the upper crust or an S wave bottoming in the upper crust. At larger distances also, arrivals caused by superposition of multiple S-wave reverberations and SV to P and/or P to SV conversions inside the whole crust
Sb	(alt:S*) Either an upgoing S wave from a source in the lower crust or an S wave bottoming in the lower crust
Sn	Any S wave bottoming in the uppermost mantle or an upgoing S wave from a source in the uppermost mantle
SnSn	Sn free-surface reflection
SgSg	Sg free-surface reflection
SmS	S reflection from the outer side of the Moho
SmSN	SmS multiple free-surface reflection; N is a positive integer. For example, SmS2 is SmSSmS
SmP	S to P reflection/conversion from the outer side of the Moho
Lg	A wave group observed at larger regional distances and caused by superposition of multiple S-wave reverberations and SV to P and/or P to SV conversions inside the whole crust. The maximum energy travels with a group velocity of approximately 3.5 km/s
Rg	Short-period crustal Rayleigh wave

Mantle phases

P	A longitudinal wave, bottoming below the uppermost mantle; also an upgoing longitudinal wave from a source below the uppermost mantle
PP	Free-surface reflection of a P wave leaving a source downward
PS	P, leaving a source downward, reflected as an S at the free surface. At shorter distances, the first leg is represented by a crustal P wave
PPP	analogous to PP
PPS	PP which is converted to S at the second reflection point on the free surface; travel time matches that of PSP
PSS	PS reflected at the free surface
PcP	P reflection from the core-mantle boundary (CMB)
PcS	P converted to S when reflected from the CMB
PcPN	PcP reflected from the free surface N − 1 times; N is a positive integer. For example PcP2 is PcPPcP
Pz + P	(alt:PzP) P reflection from outer side of a discontinuity at depth z; z may be a positive numerical value in km. For example, P660 + P is a P reflection from the top of the 660 km discontinuity
Pz−P	P reflection from inner side of a discontinuity at depth z. For example, P660 − P is a P reflection from below the 660 km discontinuity, which means it is precursory to PP
Pz + S	(alt:PzS) P converted to S when reflected from outer side of discontinuity at depth z
Pz−S	P converted to S when reflected from inner side of a discontinuity at depth z
PScS	P (leaving a source downward) to ScS reflection at the free surface
Pdif	(old:Pdiff) P diffracted along the CMB in the mantle
S	Shear wave, bottoming below the uppermost mantle; also an upgoing shear wave from a source below the uppermost mantle
SS	Free-surface reflection of an S wave leaving a source downward
SP	S, leaving a source downward, reflected as P at the free surface. At shorter distances, the second leg is represented by a crustal P wave
SSS	Analogous to SS
SSP	SS converted to P when reflected from the free surface; travel time matches that of SPS
SPP	SP reflected at the free surface
ScS	S reflection from the CMB
ScP	S converted to P when reflected from the CMB
ScSN	ScS multiple free-surface reflection; N is a positive integer. For example ScS2 is ScSScS
Sz + S	(alt:SzS) S reflection from outer side of a discontinuity at depth z; z may be a positive numerical value in km. For example S660 + S is an S reflection from the top of the 660 km discontinuity
Sz−S	S reflection from inner side of discontinuity at depth z. For example, S660 − S is an S reflection from below the 660 km discontinuity, which means it is precursory to SS
Sz + P	(alt:SzP) S converted to P when reflected from outer side of a discontinuity at depth z
Sz−P	S converted to P when reflected from inner side of a discontinuity at depth z
ScSP	ScS to P reflection at the free surface
Sdif	(old:Sdiff) S diffracted along the CMB in the mantle

Core phases

PKP	(alt:P') unspecified P wave bottoming in the core
PKPab	(old:PKP2) P wave bottoming in the upper outer core; ab indicates the retrograde branch of the PKP caustic
PKPbc	(old:PKP1) P wave bottoming in the lower outer core; bc indicates the prograde branch of the PKP caustic

PKPdf	(alt:PKIKP) P wave bottoming in the inner core
PKPpre	(old:PKhKP) a precursor to PKPdf due to scattering near or at the CMB
PKPdif	P wave diffracted at the inner core boundary (ICB) in the outer core
PKS	Unspecified P wave bottoming in the core and converting to S at the CMB
PKSab	PKS bottoming in the upper outer core
PKSbc	PKS bottoming in the lower outer core
PKSdf	PKS bottoming in the inner core
P'P'	(alt:PKPPKP) Free-surface reflection of PKP
P'N	(alt:PKPN) PKP reflected at the free surface N − 1 times; N is a positive integer. For example, P'3 is P'P'P'
P'z−P'	PKP reflected from inner side of a discontinuity at depth z outside the core, which means it is precursory to P'P'; z may be a positive numerical value in km
P'S'	(alt:PKPSKS) PKP converted to SKS when reflected from the free surface; other examples are P'PKS, P'SKP
PS'	(alt:PSKS) P (leaving a source downward) to SKS reflection at the free surface
PKKP	Unspecified P wave reflected once from the inner side of the CMB
PKKPab	PKKP bottoming in the upper outer core
PKKPbc	PKKP bottoming in the lower outer core
PKKPdf	PKKP bottoming in the inner core
PNKP	P wave reflected N − 1 times from inner side of the CMB; N is a positive integer
PKKPpre	A precursor to PKKPdf due to scattering near the CMB
PKiKP	P wave reflected from the inner core boundary (ICB)
PKNIKP	P wave reflected N − 1 times from the inner side of the ICB
PKJKP	P wave traversing the outer core as P and the inner core as S
PKKS	P wave reflected once from the inner side of the CMB and converted to S at the CMB
PKKSab	PKKS bottoming in the upper outer core
PKKSbc	PKKS bottoming in the lower outer core
PKKSdf	PKKS bottoming in the inner core
PcPP'	(alt:PcPPKP) PcP to PKP reflection at the free surface; other examples are PcPS', PcSP', PcSS', PcPSKP, PcSSKP
SKS	(alt:S') unspecified S wave traversing the core as P
SKSac	SKS bottoming in the outer core
SKSdf	(alt:SKIKS) SKS bottoming in the inner core
SPdifKS	(alt:SKPdifS) SKS wave with a segment of mantle-side Pdif at the source and/or the receiver side of the ray path
SKP	Unspecified S wave traversing the core and then the mantle as P
SKPab	SKP bottoming in the upper outer core
SKPbc	SKP bottoming in the lower outer core
SKPdf	SKP bottoming in the inner core
S'S'	(alt:SKSSKS) Free-surface reflection of SKS
S'N	SKS reflected at the free surface N − 1 times; N is a positive integer
S'z−S'	SKS reflected from inner side of discontinuity at depth z outside the core, which means it is precursory to S'S'; z may be a positive numerical value in km
S'P'	(alt:SKSPKP) SKS converted to PKP when reflected from the free surface; other examples are S'SKP, S'PKS
S'P	(alt:SKSP) SKS to P reflection at the free surface
SKKS	Unspecified S wave reflected once from inner side of the CMB
SKKSac	SKKS bottoming in the outer core
SKKSdf	SKKS bottoming in the inner core
SNKS	S wave reflected N − 1 times from inner side of the CMB; N is a positive integer
SKiKS	S wave traversing the outer core as P and reflected from the ICB
SKJKS	S wave traversing the outer core as P and the inner core as S
SKKP	S wave traversing the core as P with one reflection from the inner side of the CMB and then continuing as P in the mantle
SKKPab	SKKP bottoming in the upper outer core
SKKPbc	SKKP bottoming in the lower outer core
SKKPdf	SKKP bottoming in the inner core
ScSS'	(alt:ScSSKS) ScS to SKS reflection at the free surface; other examples are ScPS', ScSP', ScPP', ScSSKP, ScPSKP

Near source surface reflections and conversions (depth phases)

pPy	All P-type onsets (Py) as defined above, which resulted from reflection of an upgoing P wave at the free surface or an ocean bottom; WARNING: The character "y" is only a wild card for any seismic phase, which could be generated at the free surface. Examples are pP, pPKP, pPP, pPcP, etc.
sPy	All Py resulting from reflection of an upgoing S wave at the free surface or an ocean bottom; for example, sP, sPKP, sPP, sPcP, etc.
pSy	All S-type onsets (Sy) as defined above, which resulted from reflection of an upgoing P wave at the free surface or an ocean bottom. For example, pS, pSKS, pSS, pScP, etc.
sSy	All Sy resulting from reflection of an upgoing S wave at the free surface or an ocean bottom. For example, sSn, sSS, sScS, sSdif, etc.
pwPy	All Py resulting from reflection of an upgoing P wave at the ocean's free surface
pmPy	All Py resulting from reflection of an upgoing P wave from the inner side of the Moho

Surface waves

L	Unspecified long-period surface wave
LQ	Love wave
LR	Rayleigh wave
G	Mantle wave of Love type
GN	Mantle wave of Love type; N is integer and indicates wave packets traveling along the minor arcs (odd numbers) or major arc (even numbers) of the great circle
R	Mantle wave of Rayleigh type
RN	Mantle wave of Rayleigh type; N is integer and indicates wave packets traveling along the minor arcs (odd numbers) or major arc (even numbers) of the great circle

PL	Fundamental leaking mode following P onsets generated by coupling of P energy into the waveguide formed by the crust and upper mantle
SPL	S wave coupling into the PL waveguide; other examples are SSPL, SSSPL

Acoustic phases

H	A hydroacoustic wave from a source in the water, which couples in the ground
HPg	H phase converted to Pg at the receiver side
HSg	H phase converted to Sg at the receiver side
HRg	H phase converted to Rg at the receiver side
I	An atmospheric sound arrival, which couples in the ground
IPg	I phase converted to Pg at the receiver side
ISg	I phase converted to Sg at the receiver side
IRg	I phase converted to Rg at the receiver side
T	A tertiary wave. This is an acoustic wave from a source in the solid Earth, usually trapped in a low velocity oceanic water layer called the SOFAR channel (SOund Fixing And Ranging)
TPg	T phase converted to Pg at the receiver side
TSg	T phase converted to Sg at the receiver side
TRg	T phase converted to Rg at the receiver side

Amplitude measurements

The following set of amplitude measurement names refers to the IASPEI Magnitude Standard (see www.iaspei.org/commissions/CSOI/Summary_of_WG_recommendations.pdf), compliance to which is indicated by the presence of leading letter I. The absence of leading letter I indicates that a measurement is nonstandard. Letter A indicates a measurement in nm made on a displacement seismogram, whereas letter V indicates a measurement in nm/s made on a velocity seismogram.

IAML	Displacement amplitude measured according to the IASPEI standard for local magnitude M_L
IAMs_20	Displacement amplitude measured according to IASPEI standard for surface-wave magnitude $M_S(20)$
IVMs_BB	Velocity amplitude measured according to IASPEI standard for broadband surface-wave magnitude $M_S(BB)$
IAmb	Displacement amplitude measured according to IASPEI standard for short-period teleseismic body-wave magnitude m_b
IVmB_BB	Velocity amplitude measured according to IASPEI standard for broadband teleseismic body-wave magnitude $m_B(BB)$
AX_IN	Displacement amplitude of phase of type X (e.g., PP, S, etc.), measured on an instrument of type IN (e.g., SP, short-period; LP, long-period; BB, broadband)
VX_IN	Velocity amplitude of phase of type X and instrument of type IN (as above)
A	Unspecified displacement amplitude measurement
V	Unspecified velocity amplitude measurement
AML	Displacement amplitude measurement for nonstandard local magnitude
AMs	Displacement amplitude measurement for nonstandard surface-wave magnitude
Amb	Displacement amplitude measurement for nonstandard short-period body-wave magnitude
AmB	Displacement amplitude measurement for nonstandard medium to long-period body-wave magnitude
END	Time of visible end of record for duration magnitude

Unidentified arrivals

x	(old: i, e, NULL) unidentified arrival
rx	(old: i, e, NULL) unidentified regional arrival
tx	(old: i, e, NULL) unidentified teleseismic arrival
Px	(old: i, e, NULL, (P), P?) unidentified arrival of P-type
Sx	(old: i, e, NULL, (S), S?) unidentified arrival of S-type

Bibliography

Angenheister, G. H., 1921. Beobachtungen an pazifischen Beben, *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-physikalische Klasse,* 113–146.

Bastings, L., 1934. Shear waves through the Earth's core. *Nature*, **134**, 216–217.

Borne, G. von dem, 1904. Seismische Registrierungen in Göttingen, Juli bis Dezember 1903, *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-physikalische Klasse,* 440–464.

Bullen, K. E., 1946. A hypothesis on compressibility at pressures of the order of a million atmospheres. *Nature*, **157**, 405.

Chapman, C. H., 1978. A new method for computing synthetic seismograms. *Geophysical Journal of the Royal Astronomical Society*, **54**, 481–518.

Conrad, V., 1925. Laufzeitkurven des Tauernbebens vom 28. November, 1923, *Mitteilungen der Erdbeben-Kommission der Akademie der Wissenschaften in Wien*, Neue Folge. **59**, 23.

Dey-Sarkar, S. K., and Chapman, C. H., 1978. A simple method for the computation of body wave seismograms. *Bulletin of the Seismological Society of America*, **68**, 1577–1593.

Geiger, L., 1909. Seismische Registrierungen in Göttingen im Jahre 1907 mit einem Vorwort über die Bearbeitung der Erdbebendiagramme, *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-physikalische Klasse*, 107–151.

Geiger, L., and Gutenberg, B., 1912a. Konstitution des Erdinnern, erschlossen aus der Intensität longitudinaler und transversaler Erdbebenwellen, und einige Beobachtungen an den Vorläufern. *Physikalische Zeitschrift*, **13**, 115–118.

Geiger, L., and Gutenberg, B., 1912b. Ueber Erdbebenwellen. VI. Konstitution des Erdinnern, erschlossen aus der Intensität longitudinaler und transversaler Erdbebenwellen, und einige Beobachtungen an den Vorläufern, *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen. Mathematisch-physikalische Klasse*, 623–675.

Gutenberg, B., 1925. Bearbeitung von Aufzeichnungen einiger Weltbeben. *Abhandlungen der Senckenbergischen Naturforschenden Gesellschaft*, **40**, 57–88.

Gutenberg, B., and Richter, C. F., 1934. On seismic waves (first paper). *Gerlands Beitrage zur Geophysik*, **43**, 56–133.
Gutenberg, B., Wood, H. O., and Richter, C. F., 1933. Re suggestion by Dr. Harold Jeffreys regarding $\bar{P}$ and Pg. *Gerlands Beitrage zur Geophysik*, **40**, 97–98.
Jeffreys, H., 1926. On near earthquakes. *Monthly Notices of the Royal Astronomical Society. Geophysical Supplement*, **1**, 385–402.
Jeffreys, H., and Bullen, K. E., 1940. *Seismological Tables*. London: British Association for the Advancement of Science, Gray Milne Trust, p. 50.
Kennett, B. L. N., Engdahl, E. R., and Buland, R., 1995. Constraints on seismic velocities in the Earth from traveltimes. *Geophysical Journal International*, **122**, 108–124.
Linehan, D., 1940. Earthquakes in the West Indian region. *Transactions of the American Geophysical Union*, **30**, 229–232.
Macelwane J. B., Brunner, G. J., and Joliat, J. S., 1933. Re suggestion by Doctor Harold Jeffreys and others regarding $\bar{P}$ and Pg, *Gerlands Beiträge zur Geophysik*, **40**, 98.
Mohorovičić, A., 1910. Potres od 8. X 1909, *God. Izvjeste Zagr. met. Ops. Zag. 1909*, Zagreb. (Das Beben vom 8. X 1909, *Jahrbuch des meteorologischen Observatoriums in Zagreb für das Jahr 1909*, **9**, 4, 1–63).
Richter, C. F., 1958. *Elementary Seismology*. San Francisco and London: W. H. Freeman, p. 768.
Scrase, F. J., 1931. The reflected waves from deep focus earthquakes. *Proceedings of the Royal Society of London. Series A*, **A-132**, 213–235.
Sohon, F. W., 1932. *Seismometry*, Part II of Macelwane, J. B., and Sohon, F. W., 1932. *Introduction to Theoretical Seismology*, New York, 149 pp.
Stechschulte, V. C., 1932. The Japanese earthquake of March 29, 1928. *Bulletin of the Seismological Society of America*, **22**, 81–137.
Storchak, D. A., Bormann, P., and Schweitzer, J., 2002. Standard nomenclature of seismic phases. In: Bormann, P. (eds.), *New Manual of Seismological Observatory Practice,* GeoForschungsZentrum, Potsdam, Vol. 2, IS2.1, 1–18.
Storchak, D. A., Schweitzer, J., and Bormann, P., 2003. The IASPEI standard seismic phase list. *Seismological Research Letters*, **74**, 761–772.
Willmore, P. L., 1979. *Manual of Seismological Observatory Practice*, World Data Center A for Solid Earth Geophysics, Report SE-20, September 1979, Boulder, Colorado, 165 pp.

Cross-references

Seismogram Interpretation

SEISMIC PROPERTIES OF ROCKS

Nikolas I. Christensen
Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, BC, Canada

Synonyms

Rock P and S velocities

Definition

Compressional (*P*) *waves.* Seismic waves in which the rock particles vibrate parallel to the direction of wave propagation.

Shear (*S*) *waves.* Seismic waves in which the rock particles vibrate perpendicular to the direction of wave propagation.

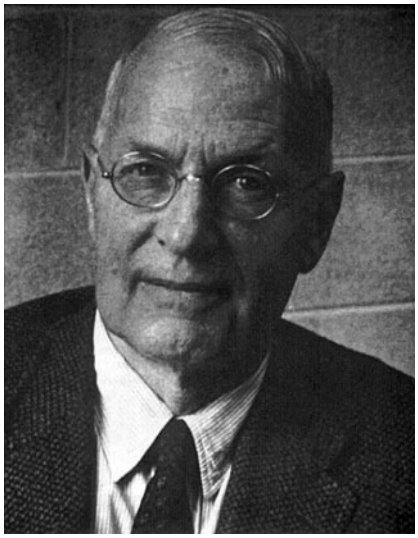
Poisson's ratio (σ). The ratio of the lateral unit strain to the longitudinal unit strain in a body that has been stressed longitudinally within its elastic limit ($2\sigma = (R^2 - 2)/(R^2 - 1)$ where $R = Vp/Vs$).

Transversely isotropic. Anisotropic solids with a single symmetry axis. Rock symmetry axes are usually normal to foliation, cleavage, or layering.

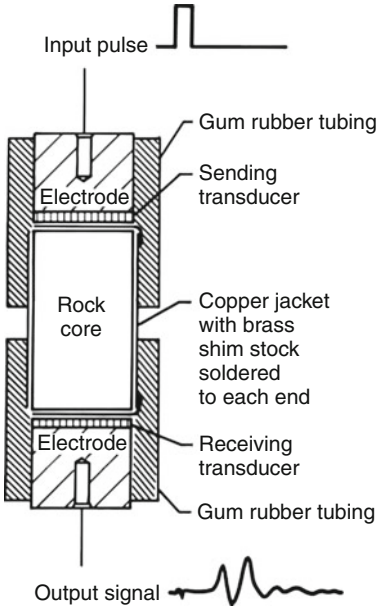
Introduction

Although many disciplines have contributed significantly to our knowledge of the Earth's interior, none has a resolution comparable to seismology. For nearly 6 decades seismic studies have provided geophysicists with worldwide information on crustal and upper mantle compressional (P) and shear (S) wave velocities. Significant data have recently become available on velocity gradients, velocity reversals, compressional and shear wave velocity ratios, and anisotropy in the form of azimuthal variations of compressional wave velocities, as well as shear wave splitting. Reflections within the crust and mantle originate from contrasts of acoustic impedances, defined as products of velocity and density. The interpretation of this seismic data requires detailed knowledge of rock velocities provided by laboratory techniques to a precision at least comparable with that of seismic measurements. In particular, to infer composition of the Earth's upper 30–50 km, the "crust," requires studies of the elasticity of rocks at conditions approaching those that exist at these depths. Of fundamental importance is the presence of mean compressive stress and temperature increasing with depth and on the average reaching about 1 GPa and 500 °C at the base of the crust. Because of this, the most relevant velocity measurements for identifying probable rock types within the crust have been measurements at elevated pressures and temperatures. These measurements often allow the seismologist to infer mineralogy, porosity, the nature of fluids occupying pore spaces, temperature at depth, and present or paleolithospheric stress based on mineral and crack orientations.

Francis Birch (Figure 1) was the pioneer in the study of rock velocities. In addition to his laboratory work on physical properties of rocks and minerals at high pressures and temperatures, he was well known for his studies of heat flow and theoretical work on the composition of the Earth's interior. Two of his benchmark papers on compressional wave velocities in rocks (Birch, 1960, 1961) set the stage for modern experimental studies of rock elasticity and have been frequently cited during the past 5 decades. These papers for the first time provided information on compressional wave velocities for many common rock types, as well as major findings on their anisotropies and relations to density. It is interesting to note that these measurements were carried out to pressures of 1 GPa, a pressure at which even today only a limited number of



Seismic Properties of Rocks, Figure 1 Francis Birch (1903–1992), a pioneer in rock physics research.



Seismic Properties of Rocks, Figure 2 Transducer and rock core assembly for velocity measurements at elevated pressures.

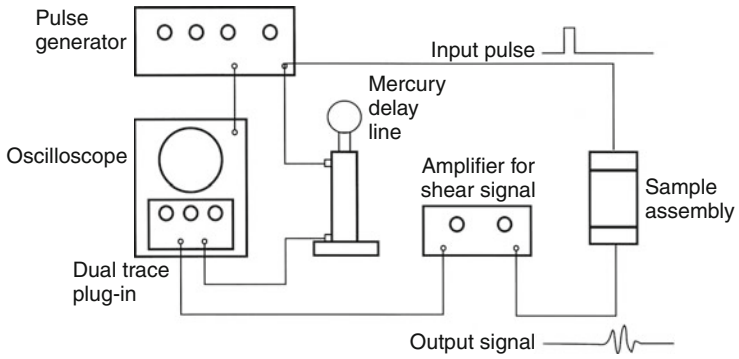
laboratories have been able to generate for modern rock seismic velocity measurements.

Measurement techniques

Rock velocities are usually measured in the laboratory using the pulse transmission technique. The transit time of either a compressional or shear wave is measured along the axis of a cylindrical rock specimen of known length. The cores are usually taken from rock samples using a 2.54 cm inner diameter diamond coring bit. The cores are trimmed and ground flat and parallel on a diamond grinding disk. The volume of each core is obtained from the length and diameter. The cores are weighed and densities are calculated from their masses and dimensions. The cores are then fitted with a copper jacket to prevent penetration of high-pressure oil into the rock samples. For measurements at high temperatures, where gas is the pressure medium, the samples are usually encased in stainless steel. Transducers are placed on the ends of the rock core (Figure 2). Compressional and shear waves are often generated by means of lead zirconate titanate (PZT) and AC cut quartz transducers with resonant frequencies of 1 MHz. The sending transducer converts the input, an electrical pulse of 50–500 V and 0.1–10 μs width, to a mechanical signal, which is transmitted through the rock. The receiving transducer changes the wave to an electrical pulse, which is amplified and displayed on an oscilloscope screen (Figure 3). Once the system is calibrated for time delays, the travel time through the specimen is determined directly by a computer or with the use of a mercury delay line. The major advantage of the delay line is that it increases the precision, especially for signals with slow rise times, because the gradual onset of the first arrival from the sample is approximated by the delay line. The velocity is the ratio of the length of the specimen to the travel time of the compressional or shear wave. The total error limits for Vp and Vs are estimated to be less than 1%. Interfacing the pressure system with a computer for data acquisition and storage permits automatic calculations of velocities as successive readings are taken. Using a least-squares routine, the computer fits a curve to the data points and calculates velocities for selected pressures. A velocity versus pressure curve is plotted along with recorded data points. Sample length, density, measured pressure velocity pairs, traces of the waveforms at selected pressures, the curve fit equations, and calculated pressure velocity pairs are recorded and stored digitally.

Hydrostatic pressure generating systems capable of producing true hydrostatic pressures as high as 3 GPa, equivalent to a depth of approximately 100 km, have been used for rock velocity measurements. Low viscosity synthetic petroleum and argon for high temperature measurements are frequently used as pressure media. An alternate technique for obtaining velocities under quasi-hydrostatic conditions has used a triaxial press with cubic samples. Transducers are placed on the six pistons and corrections are made for travel times through the pistons. Rock velocities obtained using this technique have provided valuable information on the effect of temperature on velocity, but have been limited to pressures of 0.6 GPa.

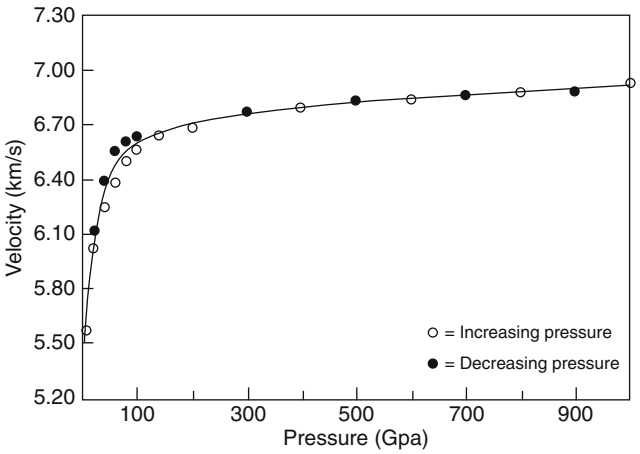
The behavior of a rocks velocity as a function of pressure is primarily dependent upon mineralogy and porosity. Many igneous and metamorphic rocks have porosities of the order of a few tenths of 1%, which are present as thin openings between grain boundaries. As pressure is applied to the rock, the cracks close and velocities increase. Once the cracks close any increase in velocity with increasing pressure is related to the intrinsic effects of pressure on



Seismic Properties of Rocks, Figure 3 Electronics for velocity measurements using a mercury delay line.

the mineral velocities. This is illustrated in Figure 4 for a garnet granulite. Velocities first increase rapidly over the first 100 MPa as cracks close and then increase slowly as pressure is increased. Also the velocity determined at a given pressure depends upon whether the pressure is approached from lower or higher pressure (Figure 4). This hysteresis is usually quite small if sufficient time is taken for measurements between pressure increments.

A considerable number of investigations have also focused on the influence of temperature on rock velocities. These studies have used either resonance techniques or more frequently the pulse transmission method. Early studies demonstrated that the application of temperature to rock at atmospheric pressure results in the creation of cracks that often permanently damage the rock and dramatically lower velocities. Thus reliable measurements of the temperature derivatives of velocities are obtained only at confining pressures high enough to prevent crack formation. At elevated confining pressures $\delta Vp/\delta T$ for common rocks often ranges from -0.3×10^{-3} to -0.6×10^{-3} km/s/°C and $\delta Vs/\delta T$ varies between -0.2×10^{-3} and -0.4×10^{-3} km/s/°C.



Seismic Properties of Rocks, Figure 4 Compressional wave velocity measurements as a function of confining pressure for a mafic granulite.

Rock velocities

Seismic velocities have been measured for practically all igneous and metamorphic rock types believed to be important constituents of the lithosphere. Because rock classification schemes allow for considerable variations in mineralogy for a given rock type, many rocks have wide ranges in elastic properties. However, some lithologies, such as the monomineralic rocks hornblendite and dunite with little or no alteration have fairly well-defined velocities. For detailed summaries of rock seismic properties the reader is referred to the compilations of Birch (1960), Christensen (1982, 1996), Gerbande (1982), Holbrook et al. (1992), Rudnick and Fountain (1995), and Mavko et al. (1998).

Table 1 contains average velocities, in the order of increasing compressional wave velocity, for several common igneous and metamorphic rocks. Volcanic rocks usually have lower velocities than their plutonic equivalents. This is due to the presence of glass, abundant alteration products and vesicles in volcanic rocks, all of which have lower velocities. In general, for a given composition, velocity increases with increasing metamorphic grade. For example, mica and quartz bearing schists have higher velocities than slates and phyllites. Low-grade metamorphosed basalts have lower velocities than higher grade amphibolite and mafic granulite. Eclogites have the highest velocity of mafic rocks. Note that shear velocities are relatively high in quartzites and low in serpentinites.

Early attempts to infer crustal composition by comparing laboratory and field derived velocities relied primarily on compressional wave velocities. However correlations between compressional wave velocity and composition are limited due to the similar velocities of many common crustal rock types. Because of this nonuniqueness of compressional wave velocity laboratory and field data comparisons, many recent studies have focused on investigations of crustal composition using both compressional and shear

wave velocities. In these studies the ratio Vp/Vs or Poisson's ratio (σ) calculated from Vp/Vs have resolved some of the ambiguities.

The values of Vp/Vs and σ, assuming isotropic elasticity, are given in Table 1 for several common igneous and metamorphic rocks at 1 GPa. This high-pressure eliminates cracks so the values only reflect mineralogy. The relatively low σ for quartzites (0.10) agrees well with isotropic aggregate calculations based on the elastic constants of single crystal quartz. Anorthosites, on the other hand, have relatively high Poisson's ratios (~0.31). As expected, values for the granites and granitic gneisses, consisting primarily of quartz and feldspar, fall in between those of quartzites and anorthosites and are relatively low. Thus, crustal regions where field measured values of $\sigma \leq 0.25$ are observed, are likely quartz-rich. Serpentinites containing lizardite, the variety of serpentine stable at crustal PT conditions, have extremely high values of Poisson's ratio (0.36), whereas unaltered dunites and peridotites have Poisson's ratios in the range of 0.25–0.26. Partially serpentinized dunites and peridotites have Poisson's ratios that fall between these limiting values. Laboratory measurements have established a well-defined relationship between Poisson's ratio, percent serpentinization and density. Changes in Poisson's ratio with progressive metamorphism of mafic igneous and pelitic rocks are considerably more complicated than the above examples (Christensen, 1996).

Seismic Properties of Rocks, Table 1 Average compressional (Vp) and shear (Vs) wave velocities, velocity ratios (Vp/Vs), and Poisson's ratios (σ) at 1 GPa for common rock types Christensen (1996)

Rock	Vp (km/s)	Vs (km/s)	Vp/Vs	σ
Serpentinite	5.607	2.606	2.152	0.36
Andesite	5.940	3.177	1.870	0.30
Quartzite	6.091	4.054	1.502	0.10
Basalt	6.118	3.291	1.859	0.30
Granitic gneiss	6.271	3.627	1.729	0.25
Granite-Granodiorite	6.372	3.726	1.710	0.24
Tonalite gneiss	6.366	3.636	1.751	0.26
Slate	6.379	3.432	1.858	0.30
Phyllite	6.398	3.608	1.774	0.27
Mica quartz schist	6.523	3.654	1.785	0.27
Zeolite facies basalt	6.530	3.493	1.869	0.30
Diorite	6.675	3.756	1.777	0.27
Diabase	6.814	3.766	1.809	0.28
Greenschist facies basalt	6.983	3.955	1.766	0.26
Marble	6.985	3.794	1.841	0.29
Mafic granulite	7.000	3.849	1.818	0.28
Amphibolite	7.046	3.987	1.767	0.26
Anorthosite	7.124	3.717	1.917	0.31
Gabbro	7.299	3.929	1.858	0.30
Pyroxenite	7.935	4.519	1.756	0.26
Eclogite	8.198	4.594	1.785	0.27
Dunite	8.399	4.783	1.756	0.26

Velocity anisotropy

Most crustal and upper mantle rocks show some degree of velocity anisotropy, which can originate from several processes. Laminar flow within magma and lava will orient elongate crystals such as feldspars along flow directions. Tabular sediments may settle preferentially and anisotropy may be enhanced by sediment compaction. Plastic flow and recrystallization during metamorphism often produce strong mineral orientations parallel to foliation and banding. In shallow crustal rocks oriented cracks producing anisotropy often originate from differential principle stresses. Anisotropy observed in laboratory measurements at pressures above approximately 100 MPa, where cracks are closed, often originate from preferred orientations of highly anisotropic minerals such as micas, amphiboles, pyroxenes, and olivine.

In general, for a given propagation direction in an anisotropic rock there are three waves, one compressional and two shear. Their vibration directions form an orthogonal set, which usually are not parallel or perpendicular to their propagation direction. Compressional and shear wave velocities vary with propagation direction, and two shear waves travel in a given direction through the rock with different velocities. This latter property of anisotropic rocks, termed shear wave splitting, was first recognized in laboratory studies and has been observed by field studies in several crustal and upper mantle regions.

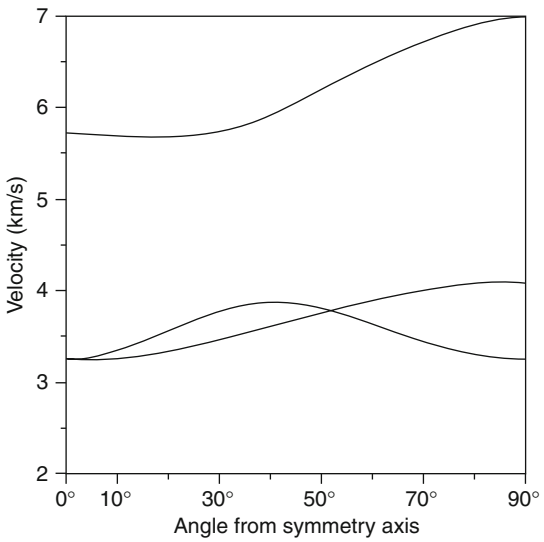
Anisotropy is usually studied in the laboratory by taking cores from a rock in different directions. Following the early investigations of Birch (1960, 1961) it is common practice to take three mutually perpendicular cores for velocity measurements. If the rock has a planar structure, such as a foliation, cleavage, or bedding, two cores are oriented with their axes within the planar structure. One of these cores is oriented parallel to a lineation, if present. In general, three velocities are measured per core: the compressional wave velocity, the velocity of the shear wave vibrating parallel to the layering, and the velocity of the shear wave vibrating in a plane perpendicular to layering. For cores taken perpendicular to layering, shear wave velocities are measured with vibration directions parallel to the axes of the cores taken in the layering.

It has been demonstrated that the above procedure provides information on maximum compressional wave anisotropy and maximum shear wave splitting. In general, the highest compressional wave velocities propagate in the plane of the foliation and parallel to lineations. Maximum shear wave splitting is observed for propagation within the planar structures. At near- normal incidence there is often minimal shear wave splitting.

Velocity measurements using multiple cores have provided detailed information on anisotropic velocities for non-axial propagation (e.g., Johnston and Christensen, 1995; Christensen and Okaya, 2007). The number of measurements necessary to completely describe wave propagation depends on the symmetry of the rock fabric. In the most general anisotropic elastic solid, 21 independent

constants are required to describe the equations of motion. Examples of materials with this type of behavior are minerals possessing triclinic symmetry such as kyanite and some feldspars. For many rocks, the existence of symmetry elements in the elastic properties leads to the vanishing of some elastic constants along with simple algebraic relations between others (e.g., Auld, 1990). Some crustal and upper mantle metamorphic rocks behave as elastic solids with orthorhombic symmetry, which require nine independent constants to describe the elastic tensors. Shales and many low to medium grade metamorphic rocks often have well-developed bedding or foliation and behave as transversely isotropic elastic solids (hexagonal symmetry with the symmetry axis normal to bedding or foliation). Transversely isotropic solids have five independent elastic constants, which can be calculated from five independent velocity measurements (two compressional wave velocities, one quasi-compressional wave velocity, and two shear wave velocities) and density. In an isotropic solid, only two independent constants are required for a complete description of elastic behavior and wave velocities are independent of propagation direction.

To describe three-dimensional wave propagation in anisotropic rocks, phase velocity surfaces can be calculated using the Kelvin-Christoffel equations (e.g., Auld, 1990) and elastic constants can be determined from velocity and density measurements. For transversely isotropic rocks, these surfaces describe variations in velocity as a function of angle to the bedding or foliation (Figure 5). Three velocity surfaces are calculated, one for the quasi-compressional wave, one for the shear wave vibrating parallel to the planar structure, and one for the quasi-shear wave vibrating in a plane perpendicular to the foliation or bedding. For propagation parallel and perpendicular to the foliation all wave modes are pure. The velocities shown in Figure 5 were calculated from velocity measurements at 600 MPa for a transversely isotropic quartz -mica schist from South Island, New Zealand. They show several important features about elastic wave propagation in this rock, which are typical of many foliated rocks. First, compressional wave velocities do not increase significantly until propagation directions greater than about 45° from foliation normal are reached. At larger angles, compressional wave velocity increases rapidly and reaches a maximum for propagation parallel to the foliation. Shear wave singularities (directions in which two shear waves have equal velocities) occur for propagation parallel to and at approximately 50° to the symmetry axis. Shear wave splitting occurs for all other propagation directions and reaches to maximum of 90° from the normal to the foliation.



Seismic Properties of Rocks, Figure 5 Compressional wave anisotropy (*upper curve*) and shear wave splitting (*lower curves*) for a transversely isotropic quartz-mica schist.

Summary

Beginning with the compressional wave velocity measurements of Birch (1960, 1961), much has been learned about elastic wave propagation in rocks. Compressional wave velocities are now available for most common rock types at pressures existing in the continental crust and uppermost mantle. Additional important contributions include laboratory measurements of shear wave velocities, velocity ratios, and the influence of temperature on velocities. Estimates of crustal composition from compressional wave velocities are nonunique, but this ambiguity is often reduced by complimentary shear wave velocity (and Poisson's ratio) observations. Recent field studies have found that seismic anisotropy is an important crustal feature. Thus systematic laboratory studies of compressional wave anisotropy and shear wave splitting will be critical in understanding crustal composition and deformation, just as they have been in investigations of the upper mantle.

Bibliography

Auld, B. A., 1990. *Acoustic Fields and Waves in Solids*. Malabar: Robert E. Krieger, Vol. 1.

Birch, F., 1960. The velocity of compressional waves in rocks to 10 kilobars, 1. *Journal of Geophysical Research*, **65**, 1083–1102.

Birch, F., 1961. The velocity of compressional waves in rocks to 10 kilobars, 2. *Journal of Geophysical Research*, **66**, 2199–2224.

Christensen, N. I., 1982. Seismic velocities. In Carmichael, R. S. (ed.), *Handbook of Physical Properties of Rocks*, 2nd edn. Boca Raton: CRC, pp. 1–228.

Christensen, N. I., 1996. Poisson's ratio and crustal seismology. *Journal of Geophysical Research*, **101**, 3139–3156.

Christensen, N. I., and Okaya, D. A., 2007. Compressional and shear wave velocities in South Island, New Zealand rocks and their application to the interpretation of seismological models of the New Zealand crust. In Okaya, D., Stern, T., and Davey, F. (eds.), *A Continental Plate Boundary: Tectonics at South Island, New Zealand*. Washington, DC: American Geophysical Union. American Geophysical Union geophysical monograph 175, pp. 125–155.

Gerbande, H., 1982. Elastic wave velocities and constants of elasticity of rocks and rock forming minerals. In Angenheister, G. (ed.), *Physical Properties of Rocks*. Springer: Landolt-Bornstein, pp. 1–140.
Holbrook, W. S., Mooney, W. D., and Christensen, N. I., 1992. Seismic velocity structure of the deep continental crust. In Fountain, D. M., Arculus, R., and Kay, R. (eds.), *Lower Continental Crust*. New York: Elsevier, pp. 1–43.
Johnston, J. E., and Christensen, N. I., 1995. Seismic anisotropy of shales. *Journal of Geophysical Research*, **100**, 5991–6003.
Mavko, G., Mukerji, T., and Dvorkin, J., 1998. *The Rock Physics Handbook*. Cambridge, MA: Cambridge University Press, pp. 289–303.
Rudnick, R. L., and Fountain, D. M., 1995. Nature and composition of the continental crust; a lower crustal perspective. *Reviews of Geophysics*, **33**, 267–309.

Cross-references

Deep Seismic Reflection and Refraction Profiling
Seismic Anisotropy
Seismic Imaging, Overview
Seismic, Reflectivity Method
Traveltime Tomography Using Controlled-Source Seismic Data
Vertical Seismic Profiling

SEISMIC QUIESCENCE AND ACTIVATION

Gennady Sobolev
Institute of Physics of the Earth, Russian Academy of Sciences, Moscow, Russia

Definition

Seismic quiescence. Relative decrease in number of earthquakes or energy in any area of a seismic active region within a certain time interval in comparison with long-term observations in the same region.
Seismic activation. Relative increase in number of earthquakes or energy in any area of a seismic active region within a certain time interval in comparison with long-term observations in the same region.

Introduction

Both seismic quiescence and seismic activation before large earthquakes have been repeatedly described in the scientific literature. Large earthquakes are viewed as natural hazards, which may cause destruction of buildings, loss of life, and economic damage. Comparative size of different earthquakes is measured by their magnitude M (see *Earthquake, Magnitude*). The most destructive earthquakes fall in the range of magnitudes 7–9; and seismic energy released from an earthquake's source (see *Earthquakes, Energy*) lies in the range of 10^{22}–10^{25} ergs and the intensity of seismic shaking at the Earth surface (see *Earthquakes, Intensity*) exceeds the 7 level of the 12° scale.

Just the first condition of detection of quiescence or activation follows from the stated above: a necessity of long-term history of seismic observations. However, for the purpose of rigorously proven detection of such occurrences one should be sure that any seismic event, magnitude of which exceeded some minimum threshold, was not missed within the observation period. Just in the second half of the (twentieth) century development of seismic networks of observation equipped with calibrated seismographs lead to creation of instrumental catalogs of earthquakes what enabled to judge objectively about appearance of seismic quiescence and activation.

The first sufficiently founded reports of appearance of seismic quiescence before large earthquakes were published at the beginning of 1960s in Japanese and were practically unknown outside Japan. Then, in 1969, Mogi, on the basis of a visual analysis of seismicity maps, formulated a hypothesis that a seismic quiescence may be a precursor of a large earthquake (Mogi, 1979). At the end of the 80th year, Wyss and Habermann studied instrumental seismic catalogs of a number of regions of the Earth and determined basic rules of a formal detection of a seismic quiescence (Wyss and Habermann, 1988). In brief, they may be summarized as follows: (1) Evaluation of homogeneity of a seismic catalog for the analyzed time interval, including magnitude calculations, taking into consideration changes in methods of determination of coordinates and depth of seismic events. (2) Determination of a minimum magnitude of earthquakes being recorded without any omissions. (3) Removal of groups and aftershocks in order to analyze the so-called background seismicity. (4) Quantitative evaluation of size and significance of an anomaly, whereas significance shall be considered a statistically proven difference of an anomaly from random variations of the background seismicity. (5) Quantitative determination of the beginning of an anomaly. (6) Evaluation of sizes of an anomalous region. Wyss and Habermann (1988) determined that numerous events of seismic quiescence described in the literature are explained by changes of a minimum representative magnitude that were not taken into account, owing to the expansion of a seismic network or development of information processing technique. However, there remain dozens of published events of seismic quiescence that may not be explained by any omissions of technical nature (artifacts).

Description of certain events of seismic quiescence exceeds the limits of this article. We will mention only as an example two widely discussed events. Ohtake et al. (1977) studied the seismicity along the Pacific Ocean coast of Mexico and found that after the middle of 1973, the seismicity rate in the region with the linear dimensions of 100 by 200 km around coordinates of (16.5°N, 96.5°W) declined sharply what was interpreted as a seismic quiescence. The Oaxaca earthquake of November 29, 1978 with M = 7.6 occurred in this region. Kisslinger (1988) found a seismic quiescence before an Andreanof Islands earthquake of May 7, 1986 with M = 8. Our task does not include any evaluation of authenticity and significance of these and other events, especially because it is not correctly to do it without the authors' participation. We will

note only that about 100 authors of scientific works published their investigations of seismic quiescence in regions with different geological and tectonic structure. It may be assumed that such phenomenon exists objectively in the nature and sometimes arises before a large earthquake. Since a seismic process to a great extent is self-similar in a wide range of magnitudes, a seismic quiescence most likely arises before a relative weak earthquake as well.

Main phases of seismic quiescence and activation

On the basis of thousands of observations of seismicity in various regions of the world, there exists a generally accepted opinion among the seismologists that the large earthquake arises after a long-term (dozens of years) increase in seismicity in a relative region connected with a gradual growth of tectonic stresses. The latter, in their turn, arise at the joints of the earth's crustal plates moving with a different rate. Let us name this period of increase in seismicity as a *phase of long-term seismic activation* (phase I in the Figure 1). Spatial dimensions of a region of long-term activation before the large earthquake (M = 7 to 8) may exceed 1,000 km.

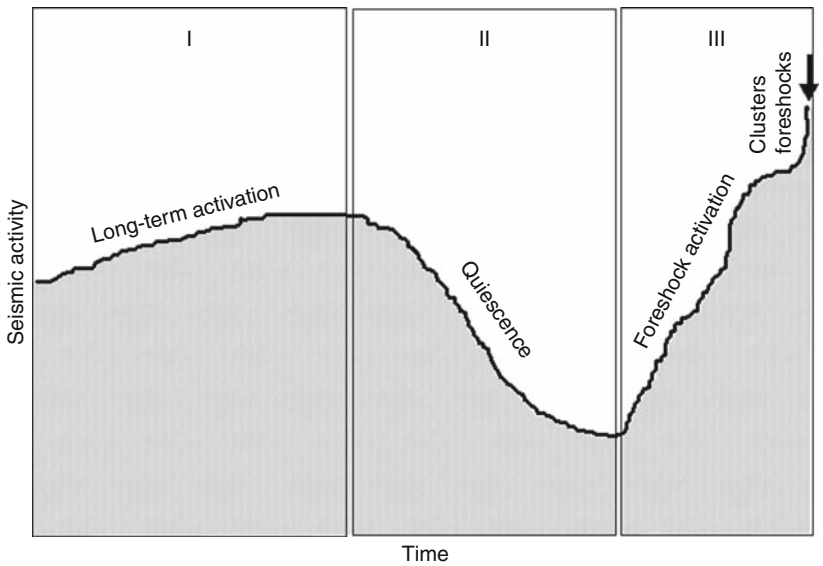
On the background of such process, in some area of the region of long-term activation a mean seismicity rate decreases what, as a rule, results in decrease in rate of release of the seismic energy. A researcher has a task to detect such decrease in the seismic activation and to prove its significance at a quantitative level, that is, to detect the seismic quiescence. All the suggested methods are based somehow or other on an analysis of statistics of earthquakes that at today's level of development of seismological networks of observations includes events contained in catalogs, the magnitude of which is from 3 to 4 units less in comparison with the magnitude of the large earthquake. This statistics includes at best the first thousands of events within a period of some years what makes the task of quiescence detection difficult and not always solvable. Below we will describe in brief two methods that were analyzed to reveal a seismic quiescence before several earthquakes both *a retrospective one* (after the large earthquake) and *a prospective one* (before the large earthquake).

In the Z-value method, there is calculated the standard deviation Z, to estimate the significance of the rate change,

$$Z = \frac{m_1 - m_2}{\left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)^{1/2}},$$

where m is the mean rate, S the variance, and n the number of events in the first and second period to be compared (Wyss and Habermann, 1988).

For the purpose of significance evaluation there is applied the normal distribution law, but when having a rather large statistics of earthquakes the method gives plausible evaluations by other distribution law as well. The more the Z value is the more the difference between a rate of a seismic flow in the area of a supposed seismic quiescence and in comparison with a long-term rate. The calculations are carried out by the enumeration of nodes of the geographical grid and a changing number of earthquakes located near each node. It results the averaged Z values are presented in the form of maps in the researched seismic active region. The area of a seismic quiescence is determined according to high values of Z isolines above the given rate of statistic significance. The experience showed that the more reliable results appeared in the task if there were removed swarms and aftershocks of previous earthquakes from the catalog.



Seismic Quiescence and Activation, Figure 1 Main phases of seismic quiescence and activation.

The RTL method uses three functions to measure the state of seismicity at a given location as a function of time (Sobolev, 2001).

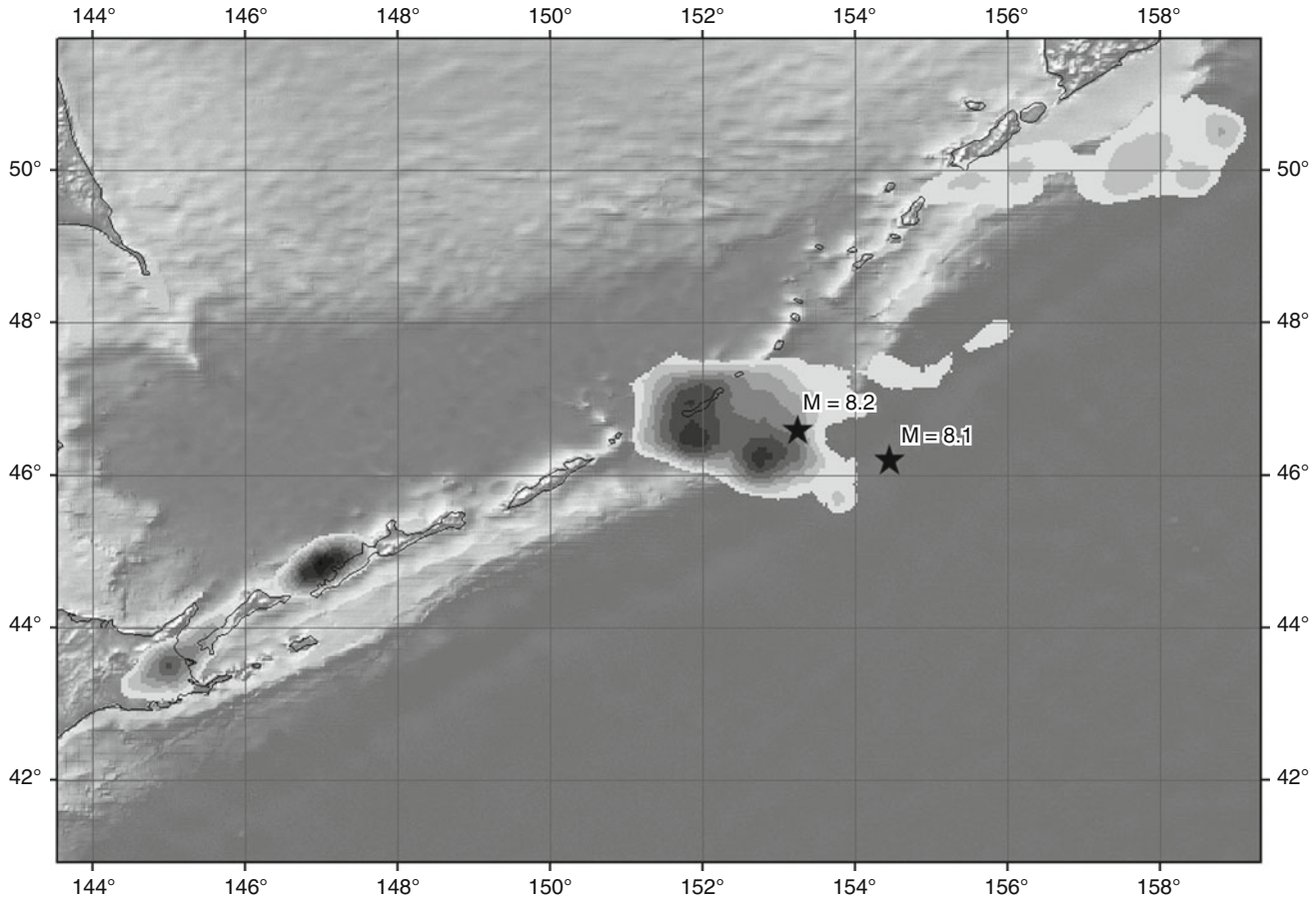
$$R(x,y,z,t) = [\Sigma \exp(-r_i/r_0)] - R_{\text{ltr}},$$
$$T(x,y,z,t) = [\Sigma \exp(-(t-t_i)/t_0)] - T_{\text{ltr}},$$
$$L(x,y,z,t) = [\Sigma \exp (l_i/r_i)^p] - L_{\text{ltr}}.$$

In these formulas, x, y, z, and t are the coordinates, the depth and the analysis time, respectively. r_i is the epicentral distance from the location selected for analyses, t_i is the occurrence time of the past seismic events, and l_i is the length of rupture estimated by empirical relation with the magnitude. The R_{ltr}, T_{ltr}, and L_{ltr} are the long-term averages of these functions. r_0 is a coefficient that characterizes the diminishing influence of more distant seismic events; t_0 is the coefficient characterizing the rate at which the preceding seismic events are "forgotten" as the time of analysis moves on; and p is the coefficient that characterizes the contribution of size of each preceding event. R, T, and L are dimensionless functions. They are normalized by their standard deviations. The product of the above three functions is calculated as the RTL-parameter, which describes the deviation from the background level of seismicity and is in units of the standard deviation, $\sigma_R\sigma_T\sigma_L$. The RTL-parameter decrease in relation to the background level means a seismic quiescence, and a structure of isolines of such parameter on the map indicates to the size of a quiescence area. In the RTL method, the outgoing catalog is cleaned of the aftershocks but the swarms are not removed.

Both methods (Z-value and RTL) require that the duration of period, which is included in calculations of the rate of background seismicity, must be by an order more than the supposed duration of the seismic quiescence. The experience of application of the described and other methods of revealing of the seismic quiescence in different tectonic conditions has shown that the duration of a quiescence period before the large earthquakes amounts to several years, that is, by an order less than the phase of the long-term activation (Figure 1). The area of significant values of the quiescence has linear dimensions about 300 km. In the most published works, the seismic quiescence was determined by the application of only one method. The exception, apparently, is a work by Wyss et al., in which a situation before the large earthquakes on the Sakhalin Island was analyzed by using of one and the same regional seismic catalog but by the application of both methods mentioned above: Z value and RTL (Wyss et al., 2004). The duration of quiescence anomalies before the Neftegorsk earthquake of May 27, 1995 with the M = 7.1, amounted according to results of both methods to 2.7 years. The size of a quiescence region was determined according to Z value in the form of a rectangle with dimensions of 200 × 600 km, and according to the RTL data – in the form of a circle with the radius up to 400 km. The epicenter of the earthquake lied at the periphery of anomalies in both determinations. Before the Uglegorsk earthquake of August 4, 2000, M = 6.8, the duration of quiescence was according to Z value 2.5 years and according to the RTL 3.0 years. The dimensions of the anomalous regions were determined in the form of a circle with the radius up to 165 and 200 km, respectively. The epicenter of the earthquake was also at the lateral part of quiescence area. The examination conducted with random catalog has shown that the probability that the anomaly occurred by chance before the Neftegorsk earthquake did not exceed 1% and before the Uglegorsk earthquake – 2% according to the results of both methods.

In spite of available facts of the seismic quiescence before some large earthquakes, the practical application of such anomaly for the forecast of earthquakes remains undecided. Authors of publications on the subject of seismic quiescence do not cite any data on a number of false alarms when no large earthquake occurred in the region of quiescence. According to our evaluations, the number of false alarms exceeds by several times the number of forecasts proven to be correct (even by a retrospective analysis). There is no statistic evaluation of a real *prospective* forecast with the use of such anomaly indeed. Let us illustrate the above stated by Figure 2, where the areas of seismic quiescence in the region of the Kurile arch are shown, which were detected before the Simushir earthquake of November 15, 2006, M = 8.2. In this case the RTL method was applied, but it is not a matter of principle. The dark spots mean the areas of seismic quiescence. The main one of them with linear dimensions of about 200 km is located in on the territory with coordinates of 46°–48°N – 151°–154°E; the epicenter of the main shock is located at its periphery, and the another major earthquake of January 13, 2007, M = 8.1 – outside the anomalous region. Several other anomalies of a lower size are shown in the figure: the territory with coordinates (45°N, 147°E), (43°N, 145°E), and (50°N, 158°E). Any large earthquakes in these regions did not occur up to now.

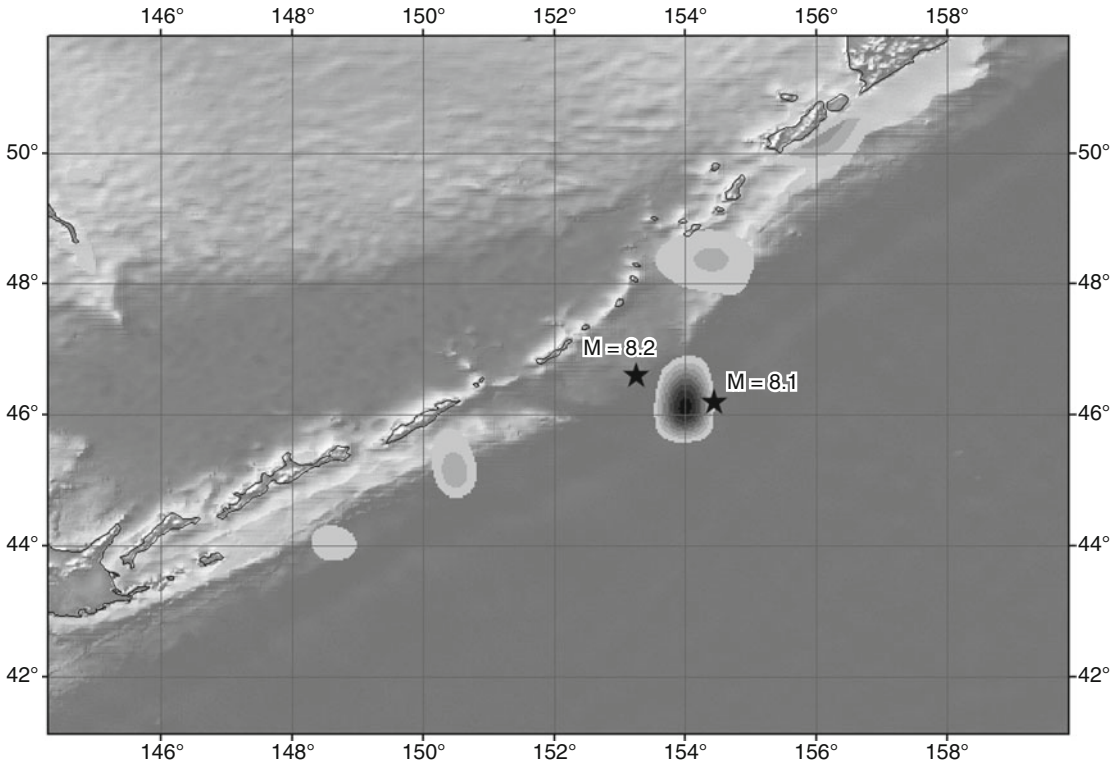
It follows from the experience of investigations of the seismic quiescence that a large earthquake seldom occurs during the most significant rate of quiescence. It arises more often after the rate of seismic activity has increased again and even exceeded that which has been observed before the beginning of the quiescence. Let us name the time interval from minimal values of activity at the stage of quiescence to the moment of the large earthquake as a phase of foreshock activation (phase III on Figure 1). Its duration has a wide range: from several days to the dozens of months. The areal of foreshock activation in general coincides with the area of aftershocks (see *Earthquake, Aftershocks*) and its size is estimated in hundreds of kilometers. These properties of the phase of foreshock activation bring to the understanding of the term "foreshock" in the broad sense. At the final step of development of such phase there may occur seismic events, the locations of which coincide practically with the source of the



Seismic Quiescence and Activation, Figure 2 Areas of seismic quiescence before the Simushir earthquake.

following large earthquake. They are traditionally called foreshocks and should be understood as foreshocks in the narrow sense (see *Earthquake, Foreshocks*). Usually they arise several hours before a large earthquake, reflecting apparently the nucleation of a rupture. This phenomenon is of a great interest for a short-term forecast, but the practical application thereof has some difficulties because no reliable criteria are found till today in order to differentiate foreshocks from usual earthquakes. In exceptional events when a number of events occurring during several hours and at one and the same place amounts to dozens, they may be interpreted presumably as foreshocks of a future large earthquake and serve as a ground for a short-term forecast. The striking example is a prediction of the Haicheng earthquake of February 4, 1975, $M = 7.3$ (Ma Zongjin et al., 1989). The foreshocks determined after the large earthquake (retrospectively) are described in a great number of publications and the description thereof is beyond the scope of this article.

Returning to the phase of foreshock activation (Figure 1), we cite the example of its detection before the Simushir earthquake mentioned above as of November 15, 2006, $M = 8.2$. The RTL method developed for the detection of a seismic quiescence allows also to identify the foreshock activation as a period of increase in the seismic activity after the quiescence. The situation showed in Figure 3 is a continuation in time of that which appeared in this area of the Kurile arch during the seismic quiescence (Figure 2). The development of the stage of seismic quiescence from the level of a long-term background to the lowest values of the seismic activity lasted in this event 1.5 years and the period of recovery of the stage of activity to the level of a long-term background (foreshock activation) lasted 1.5 years as well. The linear dimensions of foreshock activation amounted to about 100 km (Figure 3). It was located in the area of aftershocks between the main shock with the $M = 8.2$ and the strongest aftershock, $M = 8.1$. Also, like at a seismic quiescence, within the region as showed on the map (Figure 3) during the development of the main anomaly there was selected a number of anomalies with a lower intensity: on the territory of coordinates (44°N, 149°E); (45°N, 150.5°E); (48.5°N, 154.5°E.). In these places, no large seismic events occurred up to now. Thus, the foreshock activation, like the seismic quiescence, is not of high reliability in the prognostic aspect because there exist anomalies qualified as false.



Seismic Quiescence and Activation, Figure 3 Areas of foreshock activation before the Simushir earthquake.

Physical mechanisms leading to seismic quiescence and activation

Let us consider possible physical mechanisms, which bring to occurrences of the seismic quiescence and of the foreshock activation. There is no unified explanation at present and it is expedient to discuss various hypotheses. We would like to note at first that except for the three phases discussed above and the occurrence of foreshocks in the narrow sense, other anomalies may also appear sometimes. Scholz selected additional phases: doughnut pattern was observed in the period of development of the seismic quiescence around its external limits, and seismic silence was noticed just prior to the main event (Scholz, 1990).

Let us consider in a more detailed manner the hypotheses of nature of the stages of a seismic quiescence and of a foreshock activation. Ma Zongjin et al., investigated the seismic patterns before nine large earthquakes in China (Ma Zongjin et al., 1989). The authors did not select specially occurrences of a seismic quiescence or activation. They did describe mainly a migration of sources of seismic events for several years before large events in the regions adjacent to their epicenters with the dimensions of about first 100 km. Actual data cited by them are of interest as applied to a problem of physical nature of occurrences discussed in this article. A migration of sources was considered, mainly with the magnitudes 3–6, while the large earthquakes had the magnitudes of above 7. The regions round the epicenters of large events were separated in three parts: a hypocentral area A, which included a hypocenter of a large event; area of aftershocks B; and external area C.

The following main characteristics of migration were selected: During 3–10 years, till the moment of the large earthquake, two phases were varied in time. In the first one the events occurred mainly in the area C, while the areas A and B were characterized as relatively quiet. During the second phase which finished in the large earthquake, the activity was detected in the areas A and B, while the seismic activity in the area C was reduced. As the second phase developed, the activity in the areas A and B continued to last up to several months or days in relation to the moments of the large earthquake; at the end of such series the activity either accumulated inside the hypocentral area A or continued to last in the area B, while the area A was calm. We should note that the authors did not cite any quantitative evaluations of the seismicity rate, thus conclusions may be made based on a qualitative manner only. In summarization, there may be made a conclusion that the process of development of a large earthquake may occur according to various scripts, though in described events it (is) related to continental events only.

The seismic quiescence may be a consequence of increase in the strength of rocks within the seismically active area. In the dilatancy-diffusion model the increase in the strength is explained by laboratory experiments, being a consequence of appearance of open microcracks

and a relative drying of the rocks. However, it is not explained how such process can develop in the heterogeneous lithosphere in the spatial regions with the linear dimension of above 100 km.

The quiescence occurs also when the acting stresses decrease. This can be a consequence of development of an unstable deformation in the source of a future large earthquake what constitutes one of the corner stones of the model of avalanche-like fracturing. Expenditure of the accumulated potential energy brings to a decrease in stress both in the source and in the surrounding space. This explanation leads to expect an increase in seismic activity in the hypocentral area of a future large earthquake and a simultaneous occurrence of a seismic quiescence in the external area. This cannot always be observed. But the situation that the activation in the source occurs at the level of seismic events with small magnitudes, which are not registered by the available seismological network, cannot be excluded.

Decrease in the tectonic stress in some area of lithosphere can be a consequence of the motion of the neighboring blocks with a different rate. If one of the blocks stops because of a strong asperity at its tectonic boundaries with any neighboring blocks, then the stress will be accumulated at points of an asperity, while in the middle of the stopped block it will relax gradually. The activation will start at boundaries of the stopped block and then it will migrate to central parts when the asperity is destroyed.

As a modification of such assumption is the situation when the stress increases and brings to the seismicity in the stronger blocks, and the block between them is relatively discharged. The quiescence occurring in the discharged block in such situation causes an effect of "false forerunners" as no large earthquake follows after them.

Kanamori suggested an explanation of different phases of seismicity based on a model of existence of a strong inclusion at a fault being heterogeneous in its strength (Kanamori, 1981). When the stress increases gradually, the less strong inclusions disintegrate sequentially what brings to an accelerated growth of the stress at points being not disintegrated. A background seismicity is observed at the low stress. As the stress increases, there occurs the doughnut pattern – a mass destruction of rocks around the strong inclusion. The latter has a seismic quiescence. When the stress approaches to a critical level, there occurs a destruction of subunits of the strong inclusion which characterizes the phase of foreshock activation. In this hypothesis, the facts of existence of the seismic quiescence at distances that significantly exceed the rupture sizes when a large earthquake occurs are not explained.

The reason of seismic quiescence can be a change of orientation of the tensor of the current stress due to development of a creep at the fault where a future earthquake will occur. The existing fractures, which get into new conditions, require the time for unstable development (kinetics of destruction). It is not proven that this, undoubtedly, existing phenomenon can cause a quiescence at distances that significantly exceed the length of a rupture of a future earthquake.

Another reason can be a transformation of the medium surrounding the source to the quasi-plastic state, for example, when the temperature or the hydrostatic compression increases. Then the process of fracturing will occur at a lower scale level which is beyond the registration of earthquakes by the seismic network. There remains a question what are the physical reasons of increase of the temperature or the compression in the lithosphere or the earth's crust within a relatively short time period (years).

It is necessary to mention another effect, which influences on the fracturing in the geological medium. Laboratory and field observations prove that an increase or a decrease in the degree of water saturation of rocks results in a significant acceleration or retardation of occurrence of the seismic events (seismicity, triggered/induced). Further, if the rate of the seismic activity increases due to that, then it results in a relative increase in the number of events with the relatively small magnitudes and in a decrease in the number of big events. The seismic quiescence may occur: (a) if a degree of water saturation decreases, (b) if under the influence of the increased water saturation the seismicity transforms to the level of small events being not registered by the seismic network. The activation is a direct consequence of the increased water saturation and may appear also in the form of swarms.

It follows from the most available experimental facts that the foreshock activation develops in the epicentral region. Its size is less than the area of seismic quiescence, and the centers of anomalous areas of these two phenomena do not coincide as a rule. There is a ground to suppose that the physics of foreshock activation is connected with the development of unstable deformation, which is localized under the laws of mechanics in the zone of mainly two-dimensional extension. The complexity of fault systems, which display a fractal geometry, causes a successive occurrence of several zones of unstable deformation.

At the final period of foreshock activation, there often occur clusters of seismic events, that is, groups of events, the distances between the hypocenters and the times between the successive events of which are less than mean values of the background seismicity. The occurrence of clusters at the phase of foreshock activation can be explained by two reasons. First, they occur by chance because of the increase in rate of formation of events in the narrow zone of unstable deformation. Second, when the spatial density of accumulated active faults exceeds the critical level, there arise stress interactions among neighboring faults with the formation of the faults of a larger length. In the latter case, such effect must appear in the increase of middle magnitudes of seismic events, that is, in the decrease of the b value.

Summary

It follows from the above-mentioned hypotheses that there exist various physical mechanisms of occurrence both of the seismic quiescence and of the foreshock activation. For the purpose of more fundamental understanding of

such phenomena, the additional laboratory and fieldworks are required. In our opinion, special attention should be paid to the following directions.

To investigate on the same catalogs of earthquakes both the quiescence and the activation in a complex, paying attention to their distribution in space, time, and magnitudes.

To compare, where possible, the seismic patterns with the field of deformations being estimated based on the data of satellite geodesy.

To compare, where possible, the seismic patterns with the data of deep geoelectrical and hydrogeological investigations with the purpose of better understanding of the role of water.

Bibliography

Kanamori, H., 1981. The nature of seismicity patterns before large earthquakes. *Earthquake prediction. International Review.* Washington, DC: American Geophysical Union, pp. 1–19.

Kisslinger, C., 1988. An experiment in earthquake prediction and the 7 May 1986 Andreanof Islands earthquake. *Bulletin of the Seismological Society of America*, **78**, 218–229.

Ma, Z. J., Fu, Z. X., Zhang, Y. Z., Wang, C. G., Zhang, G. M., and Liu, D. F., 1989. *Earthquake Prediction, Nine Major Earthquake in China*. New York: Seismological Press, Springer, 332 pp.

Mogi, K., 1979. Two kinds of seismic gap. *Pure and Applied Geophysics*, **117**, 1172–1186.

Ohtake, M., Matumoto, T., and Latham, G. V., 1977. Seismicity gap near Oaxaca, Southern Mexico, as a probable precursor to a large earthquake. *Pure and Applied Geophysics*, **115**, 375–385.

Scholz, C. H., 1990. *The Mechanics of Earthquakes and Faulting*. Cambridge: Cambridge University Press, 439 pp.

Sobolev, G., 2001. The examples of earthquake preparation in Kamchatka and Japan. *Tectonophysics*, **338**, 269–279.

Wyss, M., and Habermann, R. E., 1988. Precursory Seismic quiescence. *Pure and Applied Geophysics*, **126**, 319–332.

Wyss, M., Sobolev, G., and Clippard, J. D., 2004. Seismic quiescence precursors to two M7 Earthquakes on Sakhalin island, measured by two methods. *Earth Planets Space*, **56**, 725–740.

Cross-references

Artificial Water Reservoir Triggered Earthquakes
Earthquake, Aftershocks
Earthquake, Foreshocks
Earthquake, Magnitude
Earthquakes, Energy
Earthquakes, Intensity

SEISMIC SEICHES

Art McGarr
U.S. Geological Survey, Menlo Park, CA, USA

Definition

Seismic seiche is a term first used by Kvale (1955) to discuss oscillations of lake levels in Norway and England caused by the Assam earthquake of August 15, 1950. This definition has since been generalized to apply to standing waves set up in closed, or partially closed, bodies of water including rivers, shipping channels, lakes, swimming pools and tanks due to the passage of seismic waves from an earthquake.

The first published mention of seismic seiches is thought to be reports of those observed throughout much of Europe due to the great earthquake at Lisbon, Portugal in 1755 (Wilson, 1953; Richter, 1958). In addition to the Lisbon and Assam earthquakes, seismic seiches at teleseismic distances have been observed for many other large earthquakes including the 1964 Alaska (McGarr and Vorhis, 1968) and the 2002 Denali, Alaska, an earthquake that caused damaging seiches in Lake Union, Seattle, Washington at an epicentral distance of 2,400 km (Barberopoulou et al., 2004).

Kvale (1955) showed that seismic surface waves from the Assam earthquake were the most probable cause of the seiches observed in Norway and England at that time. Moreover, he concluded that the natural period of a basin must be matched by the periods of the passing seismic surface waves. Motivated by observations reported by Donn (1964) of a seiche generated in a channel near Freeport, Texas, at an epicentral distance of about 5,040 km from the 1964 Alaska earthquake, McGarr (1965) developed a relation between the ground motion of seismic waves and the resulting seiche. The passing seismic wave exerts a horizontal acceleration on a closed or partially closed body of water, which can be idealized as a long channel of uniform depth. This causes a seiche, composed of standing water waves whose periods depend on the dimensions of the channel. The amplitude of the seiche is a function of channel depth, the amplitudes of the horizontal accelerations of the passing seismic waves, and the extent to which the periods of the seismic waves match those of standing water waves. The gravest mode of the standing waves has a period given by $T = 2L/\sqrt{gH}$, where T is the period in seconds, L is the channel width in meters, H is the channel depth in meters, and g is gravity. For instance, if L = 100 m and H = 10 m, then the period of the gravest seiche mode is 20 seconds, which tends to be the period where surface waves show maximum amplitudes at teleseismic distances. Any factor that enhances the amplitudes of surface waves, such as basins containing low-velocity sediments, tends to result in greater production of observed seiches from a given earthquake (McGarr and Vorhis, 1968; Barberopoulou et al., 2004).

Bibliography

Barberopoulou, A., Qamar, A., Pratt, T. L., Creager, K., and Steele, W. P., 2004. Local amplification of seismic waves from the Denali earthquake and damaging seiches in Lake Union, Seattle, Washington. *Geophysical Research Letters*, **31**, L03607, doi:10.1029/2003GL018569.

Donn, W. L., 1964. Alaska earthquake of 27 March, 1964: Remote seiche stimulation. *Science*, **146**, 261–262.

Kvale, A., 1955. Seismic seiches in Norway and England during the Assam earthquake of August 15, 1950. *Bulletin Seismological Society of America*, **45**, 93–113.

McGarr, A., 1965. Excitation of seiches in channels by seismic waves. *Journal of Geophysical Research*, **70**, 847–854.
McGarr, A., and Vorhis, R. C., 1968. Seismic seiches from the March 1964 Alaska earthquake. *U.S. Geological Survey Professional Paper*, **544E**, E1–E43.
Richter, C. F., 1958. *Elementary Seismology*. San Francisco: WH Freeman, 768 pp.
Wilson, B. W., 1953. Coastal seiches, pt. 1 of Oscillations of the sea and the phenomenon of range. The Dock and Harbour Authority [London], pp. 41–45.

SEISMIC SIGNALS IN WELL OBSERVATIONS: PRE, CO, POST

R. K. Chadha
National Geophysical Research Institute, Hyderabad, Andhra Pradesh, India

Synonyms

Abnormal water level changes; Hydrological earthquake precursors

Definition

Seismic signals in wells – rise or fall in water level fluctuations in bore wells due to earthquakes. Pre, co, and post means abnormal fluctuations prior, during, and after the occurrence of earthquake.

Introduction

Water levels in wells are found to be affected by local and distant earthquakes in addition to other phenomena like, rainfall, atmospheric pressure, and earth tides. These anomalous changes that could be pre, co, and post earthquake are believed to reflect pore pressure changes related to the redistribution of stress in the near and far fields of dislocation sources, but in general are poorly understood because of inadequate data. Numerous reports of the well-documented cases of this phenomenon have appeared from China, Japan, Russia, the USA, India, Taiwan, and other countries during the last 30 years (Wakita, 1975; Roeloffs, 1988, 1996, 1998, 2003; Liu et al., 1989, 2006; King et al., 1999, 2000; Rudnicki et al., 1993; Koizumi et al., 1996, 1999; Quilty and Roeloffs, 1997; Chadha et al., 1997, 2003, 2008; Gavrinlenko et al., 2000; Wen-Chi Lai et al., 2004; Akita and Matsumoto, 2004; Kitagawa et al., 2006; Sil, 2006a, b). Since the earthquake-related groundwater level changes were scattered and rarely observed by a sufficiently dense network of observations wells, such studies have not been adequately developed. Recently, some experiments have been taken up to drill bore wells in the seismically active region to continuously monitor water-level changes related to earthquakes.

Types of observed changes

Four types of changes in well water levels have been observed, which are related to earthquakes, namely, pre, co, post, and transient. While, the pre and post earthquake observations are mostly interpretative in nature and difficult to substantiate, coseismic and transient changes are well established. The most suitable wells for observing anomalous coseismic and transient changes are the ones that are connected to confined aquifers and show the effect of earth tides. The presence of tidal signals in well level data indicates that the well is sensitive to small strain changes in the connected rock formations and hence also should be sensitive to variations in local stress fields, and thus to earthquakes. Unconfined aquifers also show seismic signals in case of large magnitude earthquake at closer distances. Few typical examples of coseismic and transient changes are described below.

Coseismic seismic signals

Coseismic signals are generally observed in wells connected to confined aquifers and are located in the near field of earthquake source location. These are generally steplike changes coinciding with the occurrence of the earthquake. In case of shallow unconfined aquifers, coseismic oscillatory changes have also been reported.

The Chi-Chi earthquake of Mw 7.6 on September 21, 1999 is one of the well-documented events for earthquake-induced water level changes in wells. Based on the hourly digital record of the water levels, all changes were reported as coseismic or postseismic. Chia et al. (2001) described the details of these changes and compared the coseismic changes with the distances between the earthquake fault and observation wells. Wang et al. (2001) discussed the distribution of the coseismic changes and attributed them to liquefaction. Wen-Chi Lai et al. (2004) compared these changes with geological setting and seismic ground motion. They showed that in the nearby Choshui River alluvial fan area, the groundwater levels coseismically rose and those amplitudes increased as the ground acceleration and hydraulic conductivity became larger. In the slope area near the earthquake fault, the water level coseismically dropped and those amplitudes increased as the ground acceleration became larger.

Akita and Matsumoto (2004) reported coseismic changes in 29 wells associated with M 8.0 Tokachi-oki earthquake in 2003 in Japan. These changes were observed as increase in groundwater in wells located in the southeast part of Hokkaido. The maximum increase was 430 cm. In other area of Hokkaido, coseismic decreases were observed, the maximum being 59 cm. These anomalous changes were explained in terms of poroelastic response to the volumetric strain after the Tokachi-oki earthquake. The strain sensitivities determined by coseismic responses in the groundwater level were found to be consistent with those estimated by M_2 tidal strain in few wells.

Matsumoto et al. (2003) studied hydrological response at Haibara well to earthquakes in central Japan. They reported 28 coseismic changes during the period from 1981 to 1997 and obtained a relationship $M \geq 2.45 \log_{10}$

Dis + 0.45 which holds good for 26 events. Most of the water level changes in Haibara well could not be explained by poroelastic response of the water level to coseismic volumetric strain steps in confined aquifers, because the sizes and directions of the coseismic water level changes correlate poorly with the sizes and directions of the volumetric strain steps. Ground motion due to seismic waves was put forward as a more probable reason for coseismic drops in water level. Few preseismic or inter-earthquake changes were also observed by them in the residual water level data at Haibara well, which was interpreted to be related to local aseismic crustal deformation.

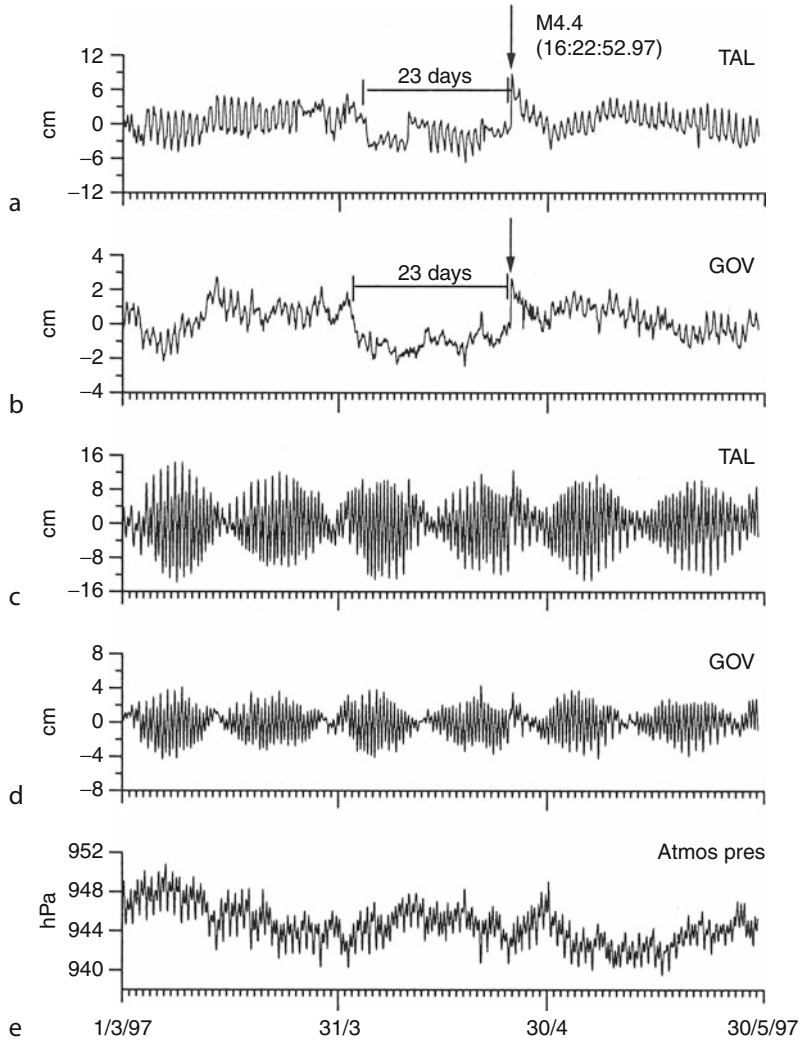
Along the San Andreas fault in the Parkfield area, water levels at 17 wells were monitored continuously as a part of Parkfield earthquake prediction experiment. Rudnicki et al. (1993) analyzed data from a well located 460 m from the San Andreas fault near Parkfield, California during January 1989–July 1990. They demonstrated that recovery rates of water level changes due to fault creep can be used to infer the position of the slipping zone relative to the well. They applied this technique to five slip-related water level changes. The changes were all characterized by a rapid drop (less than 8 h) and a slow recovery (15–30 days). The recoveries modeled by pore fluid diffusion due to a plane strain dislocation in porous, fluid-saturated elastic provided satisfactory agreement with the observations. Calculations were also done for limiting cases of both an impermeable and a permeable fault. They showed that permeability of the fault has little effect on inferences about the slip, but the inferred (horizontal) diffusivity differs for the two cases, 0.15 m^2/s for the impermeable and 0.06 m^2/s for the permeable. Later, Roeloffs (1998) analyzed persistent water level changes data at a well near Parkfield for anomalies caused by local and distant earthquakes.

Chadha et al. (2003) reported four cases of coseismic changes in well around Koyna-Warna region in western India. Koyna region is known to be site of the world's largest Reservoir Triggered Earthquake of M 6.3 in 1967. For the last 5 decades, earthquakes of $M \geq 5.0$ continue to occur in a small source volume of the dimension $30 \times 15 \times 10$ km^3 in this region. To comprehend the cause-and-effect relationship between local earthquakes and water level changes, 21 boreholes were drilled surrounding the seismic source volume. Out of the 21 boreholes, 10 were found to be connected to confined aquifers and showed strong tidal signals. Analysis of data from these wells from 1997 to 2003 revealed four cases of coseismic changes associated with earthquakes of $M \geq 4.2$. Figure 1 shows a coseismic steplike increase in two wells in the Koyna region, which is preceded by precursory drop in water levels prior to the M 4.4 earthquake. While coseismic increase of 2–7 cm in water levels was observed for earthquakes of M 4.3–4.7 in three cases, a troughlike decrease up to 8 cm was observed at seven wells for an M 5.2 earthquake. All these earthquakes occurred within the network of wells drilled for the study. From their studies, Chadha et al. (2003) concluded that the magnitude of the earthquake and epicentral distance to the wells are two vital parameters for recording the hydrological anomalies in well water levels. In Koyna, coseismic anomalies were recorded in wells for earthquakes of $M \geq 4.3$ located up to 24 km distance. Although there were several earthquakes of $M < 4.0$ during the period of study, no anomalous changes were observed in well recordings. Also, few precursory anomalies were also interpreted for local earthquakes that showed coseismic changes.

Using data of M 4.4 earthquake in Koyna, Chadha et al. (2005) tested the hypothesis that "well level fluctuations respond to changes in crustal volume strain that is induced by an earthquake in the form of step like coseismic change." Using Okada's (1992) formulation, they calculated the static volumetric strain at the surface of a homogeneous half space to see whether the coseismic steps observed for this earthquake agree with the observations of Wakita (1975) that the water level in a well shall rise or fall based on the location of the well with reference to the disposition of the fault plane. The results showed that expanding and contracting zones representing the static volumetric strain field agree with the observations of well level changes indicating that well-aquifer system indeed has the potential to reflect coseismic volumetric strain changes, similar as with the tidal fluctuation in volume strain. Further, using the strain sensitivities obtained from tidal analysis, they showed that there was misfit in amplitudes of the observed and calculated coseismic steps. Based on this study, they concluded that coseismic strains may be a function of site effects controlled by local heterogeneity in geological structures. Thus, simple elastic models cannot explain the amplitudes of hydrological anomalies wholly. Huang et al. (1995) have earlier suggested a nonlinear response of water levels to coseismic strains due to local heterogeneities. Earlier, Grecksch et al. (1999) have also observed coseismic well level steps to be higher than predicted from strain sensitivities.

Transient seismic signals

Seismic waves from distant earthquakes can cause changes in well water levels at great distances due to the passing of seismic waves (Rexin et al., 1962). There are several reports of such changes from different parts of the world, following the Sumatra earthquake of Mw 9.3 on December 26, 2004 (Kitagawa et al., 2006; Sil, 2006; Sil and Freymueller, 2006; Wang and Manga, 2006; Chadha et al., 2008). Earthquake-induced water level changes at distant locations were also reported after the Denali earthquake (Brodsky et al., 2003; Harp et al., 2003; Stricherz, 2003; Cassidy and Rogers, 2004; Kayen et al., 2004; Sil, 2006). It is thought that dynamic oscillation in crustal strain in an aquifer due to a seismic wave mainly caused dynamic oscillation in the groundwater level. Seismic oscillations, due primarily to surface waves from distant events, occur in some wells tapping highly transmissive aquifers (Liu et al., 1989; Liu et al., 2006). Earlier, several workers have shown that anomalous



Seismic Signals in Well Observations: Pre, Co, Post, Figure 1 Records of air pressure and well water level in GOV and TAL wells, during March 1–May 30, 1997. (**a**) and (**b**) residual water levels after removing earth tides and atmospheric pressure, (**c**) and (**d**) raw data after removing the trend. Spring and neap tides are clearly seen, (**e**) Atmospheric pressure. Arrow shows origin time of the earthquake on April 25, 1997. Precursory decrease 23 days prior to M 4.4 earthquake is shown by a horizontal line.

well-water level changes occur in response to deformation of the connected aquifers because of seismic waves (Cooper et al., 1965, 1967; Liu et al., 1989), fault creep (Wesson, 1981; Roeloffs, 1996, 1998), tidal strain (Bredehoeft, 1967; Van der Kamp and Gale, 1983), or atmospheric loading (Rojstaczer, 1988). Cooper et al. (1965) and Kunugi et al. (2000) showed that the amplitude of the oscillation in groundwater level is enhanced in a particular period due to the characteristics of the well-aquifer system.

M 9.1 Sumatra earthquake of 2004 and transient seismic signals

Kitagawa et al. (2006) reported water level changes in wells due to Sumatra earthquake of 2004 at a distance of more than 5,000 km in Japan. At 38 of the 45 observation stations, these changes were recorded. Ishii-type borehole strain instruments installed in 10 of the observation stations also recorded changes in crustal strains. Most of these changes in crustal strains and groundwater levels were dynamic oscillations due to a seismic wave. Sil and Freymueller (2006) reported anomalous changes in wells in Fairbanks, Alaska due to the passing of seismic waves of Sumatra earthquake. Chadha et al. (2008) reported large water level changes in six bore wells in the Koyna region in western India, following the Great Sumatra earthquake of Mw 9.3 in December 26, 2004. This earthquake occurred at 00:58:50 coordinated universal time (UTC) off the coast of Sumatra located about 3,000 km from the Koyna region. The arrival time of the *P* wave recorded at the Koyna seismic station was 01:04:24.45 (UTC). The anomalous water level changes were observed at 01:15

(UTC) at all of these wells where the sampling rate was 15 min. No abnormal changes were noticed in the preceding sampling time at 01:00 (UTC), clearly indicating that these changes occurred due to the passage of seismic waves generated by the Sumatra earthquake. However, due to the large sampling interval, it was difficult to correlate the water level changes either to shear or surface waves. No local earthquake was recorded by the network during this period. The anomalous water level changes were of the order of 2–65 cm and showed either a spikelike or a steplike change. These types of changes are attributed to the dynamic strain induced by the passage of seismic waves, most probably long period surface waves.

Summary and conclusion

Earthquakes can cause changes in well water levels in near and far fields. This has been exemplified by several reported cases in the literature during the last 3 decades. Therefore, water level monitoring becomes a valuable tool for earthquake prediction research. Although, the coseismic and transient water level changes are well established, the greatest challenge is to identify precursory anomalous changes that can be identified prior to the occurrence of earthquakes. Very often, these precursory changes that represent the pre-earthquake phase are very small and therefore not very obvious in the raw data. This is due to the response of wells to large atmospheric pressure, significant precipitation, or earth tides, which may obscure these smaller earthquake-related changes. Sophisticated filtering techniques have to be developed to decipher the earthquake related anomalies from non-tectonic effects in the data. Definite patterns of water level anomalies have to be established, both site specific and global, so that predictive capabilities can be developed. This could be done with continuous monitoring of dedicated networks of wells around known seismically active regions in different countries.

At present, three categories of anomalies have been reported, namely, co- and preseismic, aseismic, and transient changes. In the first category, a steplike coseismic rise or fall has been observed. This type of coseismic steps is understood as sudden pore pressure changes related to an alteration in in-situ volume strain caused by the redistribution of stress in the brittle crust (Bodvarsson, 1970; Kuempel, 1992; Muir-Wood and King, 1993). Sometimes, these coseismic steps were found to be preceded by persistent water level drops prior to the earthquakes. According to Sadovsky et al. (1979) and Monakhov et al. (1983), the most common precursor is a water level drop of a few centimeters amplitude several days before the earthquake. Typically, the drop is beginning to recover when the earthquake occurs. This type of anomaly has been referred to as "rebound anomaly" (Igarashi et al., 1992) and is believed to be related to an increase of porosity and permeability due to fracturing, with the subsequent recovery attributable either to influx of fluid or to compression (Roeloffs, 1988). Another type of precursory water level drop that occurs much prior to coseismic steps, but recovers just before the earthquake. This type of water level drops could be due to aseismic creep along a fault plane where the earthquake stress is building up. The third type of precursory change occurs and recovers much before the occurrence of the earthquake. This type of water level drops and recovery is believed to be due to small slips on the hanging wall of a normal fault prior to occurrence of the main earthquake. All these precursory changes are established in the hindsight, that is, after the occurrence of earthquakes. Continuous data sets for long periods will help in establishing these kinds of anomalies.

Bibliography

Akita, F., and Matsumoto, N., 2004. Hydrological responses induced by the Tokachi-oki earthquake in 2003 at hot spring wells in Hokkaido. Japan. *Geophysical Research Letters*, **31**, L16603, doi:10.1029/2004GL020433.

Bodvarsson, G., 1970. Confined fluids as strain meters. *Journal of Geophysical Research*, **75**, 2711–2718.

Bredehoeft, J. D., 1967. Response of well-aquifer systems to earth-tide. *Journal of Geophysical Research*, **72**, 3075–3087.

Brodsky, E. E., Roeloffs, E., Woodcock, D., Gall, I., and Manga, M., 2003. A mechanism for sustained groundwater pressure changes induced by distant earthquakes. *Journal of Geophysical Research*, **108**, 2390.

Cassidy, J. F., and Rogers, G. C., 2004. The Mw 7.9 Denali fault earthquake of 3 November 2002: felt reports and unusual effects across western Canada. *Bulletin of the Seismological Society of America*, **94**, S53–S57.

Chadha, R. K., Gupta, H. K., Kuempel, H.-J., Mandal, P., Nageswara Rao, A., Kumar, N., Radhakrishna, I., Rastogi, B. K., Raju, I. P., Sarma, C. S. P., Satyamurthy, C., and Satyanarayana, H. V., 1997. Delineation of active faults, nucleation process and pore pressure measurements at Koyna (India). *Pure and Applied Geophysics*, **150**, 551–562.

Chadha, R. K., Pandey, A. P., and Kuempel, H. J., 2003. Search for earthquake precursors in well water levels in a localized seismically active area of reservoir triggered earthquakes in India. *Geophysical Research Letters*, **30**(7), 1416, doi:10.1029/2002GLO016694.

Chadha, R. K., Srivastava, K., and Kumpel, H. J., 2005. Earthquake related changes in well water level and their relation to a static deformation model for the seismically active Koyna–Warna region, India. In Rummel, F. (ed.), *Rock Mechanics with Emphasis on Stress*. New Delhi: Oxford and IBH, pp. 135–150.

Chadha, R. K., Chandrani Singh, and Shekar, M., 2008. Transient changes in well water level in bore wells in western India due to 2004 Mw 9.3 Sumatra earthquake. *Bulletin of the Seismological Society of America*, **98**(5), 2553–2558, doi:10.1785/0120080952.

Chia, Y., Wang, Y. S., Chiu, J. J., and Liu, C. W., 2001. Changes of groundwater level due to the 1999 Chi-Chi earthquake in the Choshui River alluvial fan in Taiwan. *Bulletin of the Seismological Society of America*, **91**, 1062–1068.

Cooper, H. H., Bredehoeft, J. D., Papdopulos, I. S., and Bennett, R. R., 1965a. The response of well-aquifer systems to seismic waves. *Journal of Geophysical Research*, **70**, 3915–3926.

Cooper, H. H., Bredehoeft, J. D., Papdopulos, I. S., and Bennett, R. R., 1965b. The response of well-aquifer systems to seismic waves. *Journal of Geophysical Research*, **70**, 3915–3926.

Cooper, H., Bredehoeft, J. D., and Papadopulos, I. S., 1967. Response of a finite-diameter well to an instantaneous charge of water. *Water Resources Research*, **3**, 263–269.

Gavrinlenko, P., Melikadze, G., Chelidze, T., Gibert, D., and Kumsiashvili, G., 2000. Permanent water level drop associated with Spitak earthquake: observations at Lisi Borehole (Republic of Georgia) and modeling. *Geophysical Journal International*, **143**, 83–98.

Grecksch, G., Roth, F., and Kuempel, H. J., 1999. Coseismic well level changes due to the 1992 Roermond earthquake compared to static deformation of half space solutions. *Geophysical Journal International*, **138**, 470–478.

Gupta, H. K., Radhakrishna, I., Chadha, R. K., Kuempel, H.-J., and Grecksch, G., 2000. Pore pressure studies initiated in area of reservoir-induced earthquakes in India. *EOS. Transactions of the American Geophysical Union*, **81**(14), 145–151.

Harp, E. L., Jibson, R. W., Kayen, R. E., and Keefer, D. K., 2003. Landslides and liquefaction triggered by the Mw 7.9 Denali fault earthquake of 3 November 2002. *GSA Today*, **13**, 4–10.

Huang, W., Rojstaczer, S., and Breau, S., 1995. Coseismic response of water level to earthquakes in the San jacinto Fault, Southern California. *EOS, Transaction of the American Geophysical Union*, **76**(Suppl.), 359.

Huang, F.-Q., Jian, C.-L., Tang, Y., Xu, G.-M., Deng, Z.-H., and Chi, G.-C., 2004. Response changes of some wells in the mainland subsurface fluid monitoring network of China, due to the September 21, 1999, Ms 7.6 Chi-Chi earthquake. *Tectonophysics*, **390**(1–4), 217–234.

Igarashi, G., Wakita, H., and Sato, T., 1992. Precursory and coseismic anomalies in well water levels observed for the February 2, 1992 Tokyo Bay earthquake. *Geophysical Research Letters*, **19**, 1583–1586.

Janssen, J., 1998. Untersuchung des zusammenhangs zwishen lokaler seismizitat und dem porendruck in gespannten aquiferen in der Koyna region, Indien, *Diploma Thesis*, Germany: University of Bonn, 120 pp.

Kayen, R. E., Thompson, E., Minasian, D., Moss, R. E., Collins, B., Sitar, N., Dreger, D., and Carver, G., 2004. Geotechnical reconnaissance of 2002 Denali fault, Alaska earthquake. *Earthquake Spectra*, **20**, 639–667.

King, C.-Y., Azuma, S., Igarashi, G., Ohno, M., Saito, H., and Wakita, H., 1999. Earthquake-related water-level changes at 16 closely clustered wells in Tono, central Japan. *Journal of Geophysical Research*, **104**, 13073–13082.

King, C.-Y., Zuma, S., Hno, M., Asai, Y., Kitagawa, Y., Igarashi, G., and Wakita, H., 2000. In search of earthquake precursors in the water-level data of 16 closely clustered wells at Tono, Japan. *Geophysical Journal International*, **143**, 469–477.

Kissin, I. G., Belikov, V. M., and Ishankuliev, G. A., 1996. Short-term groundwater level variations in a seismic region as an indicator of the geodynamic regime. *Tectonophysics*, **265**, 313–326.

Kitagawa, Y., Koizumi, N., Takahashi, M., Matsumoto, N., and Sato, T., 2006. Changes in groundwater levels or pressures associated with the 2004 earthquake off the west coast of northern Sumatra (M 9.0). *Earth Planets Space*, **58**, 173–179.

Koizumi, N., Kano, Y., Kitagawa, Y., Sato, T., Takahashi, M., Nishimura, S., and Nishida, R., 1996. Groundwater anomalies associated with the 1995 Hyogoken Nanbu earthquake. *Journal of Physics of the Earth*, **44**, 373–380.

Koizumi, N., Tsukuda, E., Kamigaichi, O., Matsumoto, N., Takahashi, M., and Sato, T., 1999. Preseismic changes in groundwater level and volumetric strain associated with earthquake swarms off the east coast of the Izu peninsula, Japan. *Geophysical Research Letters*, **26**, 3509–3512.

Kuempel, H.-J., 1992. About the potential of wells to reflect stress variations within inhomogeneous crust. *Tectonophysics*, **211**, 317–336.

Kunugi, T., Fukao, Y., and Ohno, M., 2000. Underdamped responses of a well to nearby swarm earthquakes off the coast of Ito City, central Japan, 1995. *Journal of Geophysical Research*, **105**, 7805–7818.

Lai, W.-C., Koizumi, N., Matsumoto, N., Kitagawa, Y., Lin, C.-W., Shieh, C.-L., and Lee, Y.-P., 2004. Effects of seismic ground motion and geological setting on the coseismic groundwater level changes caused by the 1999 Chi-Chi earthquake Taiwan. *Earth Planets Space*, **56**, 873–880.

Liu, L.-B., Roeloffs, E., and Zheng, X.-Y., 1989. Seismically induced water level fluctuations in the WaliWell, Beijing, China. *Journal of Geophysical Research*, **94**, 9453–9462.

Liu, C., Huang, M. W., and Tsai, Y. B., 2006. Water level fluctuations induced by ground motions of local and teleseismic earthquakes at two wells in Hualien, eastern Taiwan. *TAO*, **17**, 371–389.

Matsumoto, N., Kitagawa, G., and Roeloffs, E. A., 2003. Hydrological response to earthquakes in the Haibara well, central Japan – I. Ground water level changes revealed using state space decomposition of atmospheric pressure, rainfall and tidal responses. *Geophysical Journal International*, **155**, 885–898.

Monakhov, F. I., Khantayev, A. M., Kechaykina, V. S., and Shishkin, A. A., 1983. Comparative characteristics of the response of groundwater level and outflow to earthquake preparation. *Izvestiya, Earth Physics*, **19**, 405–408.

Montgomery, D. R., and Manga, M., 2003. Streamflow and water well response to earthquakes. *Science*, **300**, 2047–2049.

Muir-Wood, R., and King, G. C. P., 1993. Hydrological signatures of earthquake strain. *Journal of Geophysical Research*, **98**, 22035–22068.

Okada, Y., 1992. Internal deformation due to shear and tensile faults in a half space. *Bulletin of the Seismological Society of America*, **82**(2), 1018–1040.

Quilty, E. G., and Roeloffs, E. A., 1997. Water level changes in response to the December 20, 1994 M 4.7 earthquake near Parkfield, California. *Bulletin of the Seismological Society of America*, **87**, 310–317.

Rexin, E. E., Oliver, J., and Prentiss, D., 1962. Seismically induced fluctuations of the water level in the Nunn-Bush well in Milwaukee. *Bulletin of the Seismological Society of America*, **52**, 17–25.

Roeloffs, E. A., 1988. Hydrologic precursors to earthquakes: a review. *Pure and Applied Geophysics*, **126**, 177–209.

Roeloffs, E. A., 1996. Poroelastic techniques in the study of earthquake related hydrological phenomena. *Advances in Geophysics*, **37**, 135–195.

Roeloffs, E. A., 1998. Persistent water level changes in a well near Parkfield, California due to local and distant earthquakes. *Journal of Geophysical Research*, **103**, 869–889.

Roeloffs, E. A., Sneed, M., Galloway, D. L., Sorev, M. L., Farrar, C. D., Howle, J. F., and Hughes, J., 2003. Water-level changes induced by local and distant earthquakes at Long Valley caldera. California. *Journal of Volcanology and Geothermal Research*, **127**, 269–303.

Rojstaczer, S., 1988a. Intermediate period response of water levels in wells to crustal strain: sensitivity and noise level. *Journal of Geophysical Research*, **93**, 13619–13634.

Rojstaczer, S., 1988b. Determination of fluid flow properties from response of water levels in wells to atmospheric loading. *Water Resources Research*, **24**, 1927–1938.

Rojstaczer, S., and Agnew, D. C., 1989. The influence of formation material properties on the response of water levels in wells to earth tides and atmospheric loading. *Journal of Geophysical Research*, **94**, 12403–12411.

Rudnicki, J. W., Yin, J., and Roeloffs, E. A., 1993. Analysis of water level changes induced by fault creep at Parkfield, California. *Journal of Geophysical Research*, **98**, 8143–8152.

Sadovsky, M. A., Monakhov, F. I., Kissin, I. G., and Shirokov, B. D., 1979. Short-term hydrogeodynamic precursors of

earthquakes in earthquake prediction. *Proceedings of the International (UNESCO) Symposium on Earthquake Prediction*. Paris, France.
Sil, S., 2006a. Response of Alaskan wells to near and distant large earthquakes. Master's Thesis, University of Alaska, 83 pp.
Sil, S., and Freymueller, J. T., 2006b. Well water level changes in Fairbanks, Alaska due to the great Sumatra–Andaman earthquake. *Earth Planets Space*, **58**, 181–184.
Singh, C., Ramana, D. V., Chadha, R. K., and Shekar, M., 2008. Coseismic responses and the mechanism behind Mw 5.1 earthquake of March 14, 2005 in the Koyna–Warna region, India. *Journal of Asian Earth Sciences*, **31**(4–6), 499–503.
Sterling, A., and Smets, E., 1971. Study of earth tides, earthquakes and terrestrial spectroscopy by analysis of the level fluctuations in a borehole at Heibaart (Belgium). *Geophysical Journal of the Royal Astronomical Society*, **23**, 225–242.
Stricherz, V., 2003. Seismology getting to know hydrology. *University Week*, **29**, 20.
Talwani, P., 1997. Seismotectonics of the Koyna–Warna area, India. *Pure and Applied Geophysics*, **150**, 511–550.
Van der Kamp, G., and Gale, J. E., 1983. Theory of the earth tide and barometric effects in porous formations with compressible grains. *Water Resources Research*, **19**, 538–544.
Wakita, H., 1975. Water wells as possible indicators of tectonic strains. *Science*, **189**, 553–555.
Wang, C., and Manga, M., 2006. Threshold distance for streamflow increase and liquefaction during earthquakes – implications on their mechanisms. *Western Pacific Geophysics Meeting Supplement*, Vol. 87, No. 36, Beijing, China, 24–27 July 2006.
Wang, C. Y., Cheng, L. H., Chin, C. V., and Yu, S. B., 2001. Coseismic hydrologic response of an alluvial fan to the 1999 Chi-Chi earthquake, Taiwan. *Geology*, **29**, 831–834.
Wesson, R. L., 1981. Interpretation of changes in water level accompanying fault creep and implications for earthquake prediction. *Journal of Geophysical Research*, **86**, 9259–9267.
Zhang, Z.-D., Chi, Z. L., Chen, H. M., and Luan, G.-Y., 2000. Vibration of water level in well and earthquake wave. *Journal of Seismological Research*, **23**(4), 418–425.

Cross-references

Earthquake Precursors and Prediction
Poroelasticity
Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity

SEISMIC STRUCTURE AT MID-OCEAN RIDGES

Donald W. Forsyth
Department of Geological Sciences, Brown University, Providence, RI, USA

Definition

Lithosphere. The cool, uppermost layer of the solid earth that moves as a unit and has some long-term elastic rigidity. It usually consists of both crust and the uppermost mantle. The mantle portion is typically characterized by high seismic velocities and low attenuation.
Asthenosphere. A more deformable, low-viscosity layer in the mantle underlying the lithosphere. Characterized by low seismic velocities and relatively high attenuation.
Anisotropy. Physical properties at any one point varying depending on direction; for seismic waves, velocity may depend on direction of propagation and direction of polarization of the wave.

Introduction

Mid-ocean ridges are spreading centers where two oceanic, lithospheric plates move apart. The separation of the plates induces upwelling in the underlying asthenospheric mantle. Because melting temperature decreases with decreasing pressure, as the mantle upwells, it undergoes pressure-release partial melting, producing magma that migrates upward to form new oceanic crust. As the plates move away from the spreading center, heat is lost to the surface conductively by diffusion and convectively by hydrothermal circulation. As the crust and mantle lose heat, the magma solidifies, and the plates thicken and cool with increasing age of the seafloor.

Although the general outline of the processes beneath mid-ocean ridges leading to the formation of new seafloor is well known, there are many questions remaining about the details of the mantle flow, melt generation, and melt migration that have not yet been fully answered. For example, along mid-ocean ridges, spreading centers are offset and segmented by transform faults and overlapping spreading centers. Are there distinct centers of mantle upwelling beneath each ridge segment or is the segmentation a shallow manifestation of stresses within the lithosphere with upwelling a more or less continuous phenomenon in the asthenosphere? How broad and deep is the melt production region?

The propagation of seismic waves through the crust and mantle provides one of the most direct ways of probing the structure beneath the ridges associated with plate separation and crust formation. The velocity, attenuation, and anisotropy of the waves are affected by temperature, composition, the presence of melt or cracks, and the crystal fabric. Tomographic images and maps of crustal and uppermost mantle structure are typically created in active source experiments where artificial sound sources, such as air guns, are recorded by ocean bottom seismometers (OBS) and/or long arrays of hydrophones towed behind ships. Deeper mantle structure is probed using signals generated by distant, teleseismic earthquakes recorded on arrays of ocean-bottom and land seismometers over periods of months to years. The logistical difficulty and expense of such experiments means that relatively few mid-ocean ridges have been studied in detail.

Mantle structure

A cross section of the shear velocity structure beneath the East Pacific Rise spreading center is shown in Figure 1 in comparison to the velocity variations that would be expected for simple conductive cooling of the plates if shear velocity were sensitive only to temperature and pressure. This tomographic image is based on the propagation of Rayleigh surface waves across two OBS arrays deployed for periods of

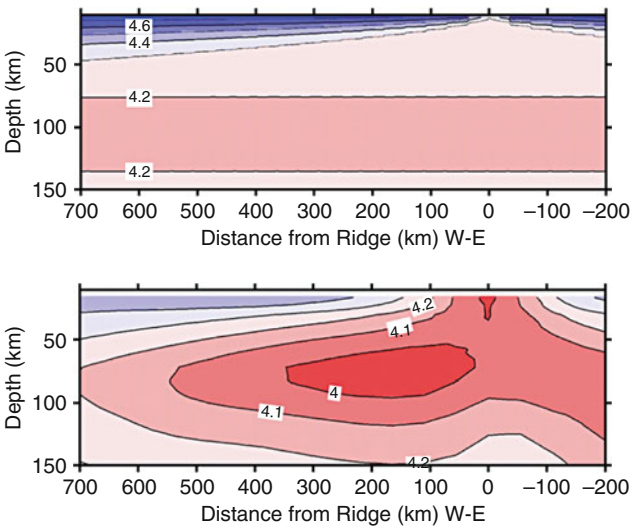
6 and 12 months. The expected thickening of the plate with increasing distance from the ridge axis as the plate cools is clearly observed in the form of increasing shear wave velocity near the surface. Velocity changes extend substantially deeper than is predicted and the shear velocity is lower than expected for the direct effects of temperature variations. Both of these departures from the predictions are indications that partial melt may be present, which could reduce the shear (S) wave velocity.

The lowest velocities are observed at depths of 60–80 km, where petrological models predict maximum melt production. There also are very low shear velocities at shallow depths immediately beneath the ridge axis, which may represent higher melt concentrations in the mantle that accumulated as melt migrates upward and inward to form the new oceanic crust at the ridge axis.

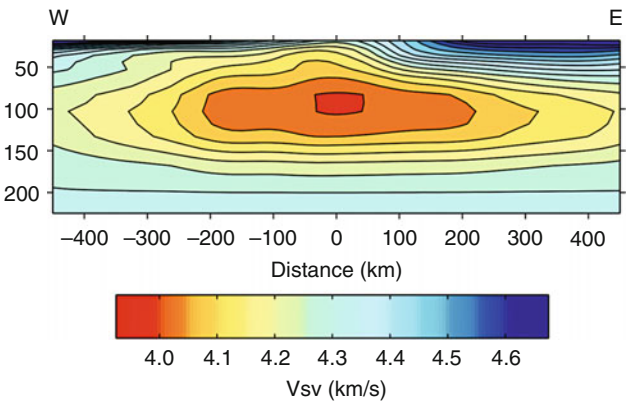
Another important feature of the shear velocity structure is the asymmetry across the ridge axis. To the east, beneath the Nazca plate, the high-velocity surface layer increases in thickness more rapidly and the very low-velocity region is absent. This asymmetry is also observed in the degree of shear wave splitting, an indicator of anisotropy, in delays of compressional (P) and S waves, in electrical conductivity, and in the rate of subsidence of the seafloor away from the ridge axis (MELT Seismic Team, 1998; Evans et al., 1999). S wave tomography indicates that the asymmetry may extend to depths of 200 km or more (Hammond and Toomey, 2003). The asymmetry is probably caused by large-scale mantle flow coming from the hotspot region to the west beneath the Pacific plate, coupled with migration of the spreading center to the west. The fast direction for seismic wave propagation, as indicated by shear wave splitting and Rayleigh wave anisotropy, is perpendicular to the East Pacific Rise, consistent with the alignment of olivine crystals in the mantle expected for plate formation and flow from the west.

The East Pacific Rise is one of the fastest spreading ridges, with full spreading rate of about 14 cm/year. Pronounced asymmetry is also observed across the Reykjanes Ridge south of Iceland (Figure 2), which has a full spreading rate of only about 2 cm/year (Delorey et al., 2007). In this case, the tomographic study took advantage of the existence of arrays of stations on Iceland, which straddles the Reykjanes/Mid-Atlantic Ridge. Like the East Pacific Rise, very low shear velocities (~4.0 km/s) indicative of the presence of melt are found in a broad region beneath the ridge. For both the East Pacific Rise and the Reykjanes Ridge, there is too little attenuation of surface waves to attribute the very low velocities to the effect of high temperature alone. Velocities at depths shallower than 80 km are lower on the west side, beneath the North American plate, perhaps due to the westward migration of the ridge and upwelling in the mantle in the wake of the thicker North American lithosphere farther west. The anisotropy pattern is different than for typical mid-ocean ridges, perhaps indicating that there is along-axis flow away from the Iceland hotspot in the asthenosphere.

To date, there have been no experiments that provide good control of along-axis variations in mantle structure at depths of tens of kilometers or more where melt production is expected to occur. In the Gulf of California, a surface wave study showed that there are along-axis variations in shear velocity with minima spaced at intervals of about 250 km, perhaps indicative of discrete upwelling centers (Wang et al., 2009), but that spreading system is flanked by continental crust on both sides and dominated by long transform faults, so it may not be typical. At shallower depths just



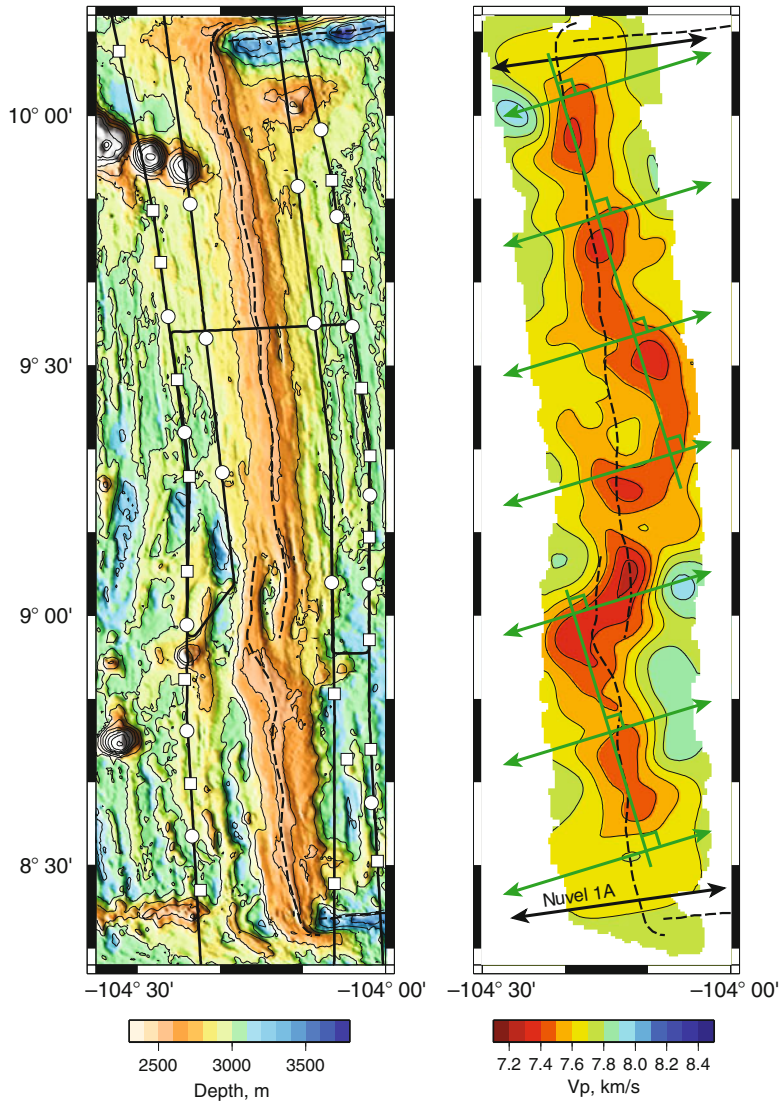
Seismic Structure at Mid-Ocean Ridges, Figure 1 Tomographic cross section of the East Pacific Rise, comparing predicted shear velocity structure (*top*) to observed structure (*bottom*). Contours are labeled in km/s. Velocities are significantly lower than predicted for temperature effects alone, suggesting the presence of partial melt. After Harmon et al. (2009).



Seismic Structure at Mid-Ocean Ridges, Figure 2 Tomographic cross section of shear wave velocity structure across Reykjanes Ridge south of Iceland. Note strong asymmetry in upper 100 km between North American plate (west) and European plate (east). After Delorey et al. (2007).

beneath the Moho (the transition from crust to mantle), there are clearly along-axis variations in P-wave velocity on the northern East Pacific Rise, the boundary between the Pacific and Cocos plates (Toomey et al., 2007). Using long offset arrivals refracting from the Moho (Pn) observed in an active experiment, velocity minima were found spaced about 25 km apart (Figure 3). P-wave velocities in the minima are 7.4 km/s or less, suggesting the presence of 1–3% melt distributed in films or thin sheets (typical P-wave velocities at the Moho are 7.8–8.4 km/s). However, it is not clear whether these apparent centers of melt lie above centers of upwelling mantle or they represent a scale length associated with melt migration.

Most of the centers of melt concentration lie beneath or very close to the spreading center determined from detailed bathymetric surveys, but one at 9° 30' N is displaced several kilometers from the axis. Because the fast direction for P-wave propagation is skewed from the spreading direction and is not perpendicular to the strike of the ridge axis, Toomey et al., inferred that upwelling and mantle flow at depth may also be skewed, controlling the location of the velocity minima just below the Moho.



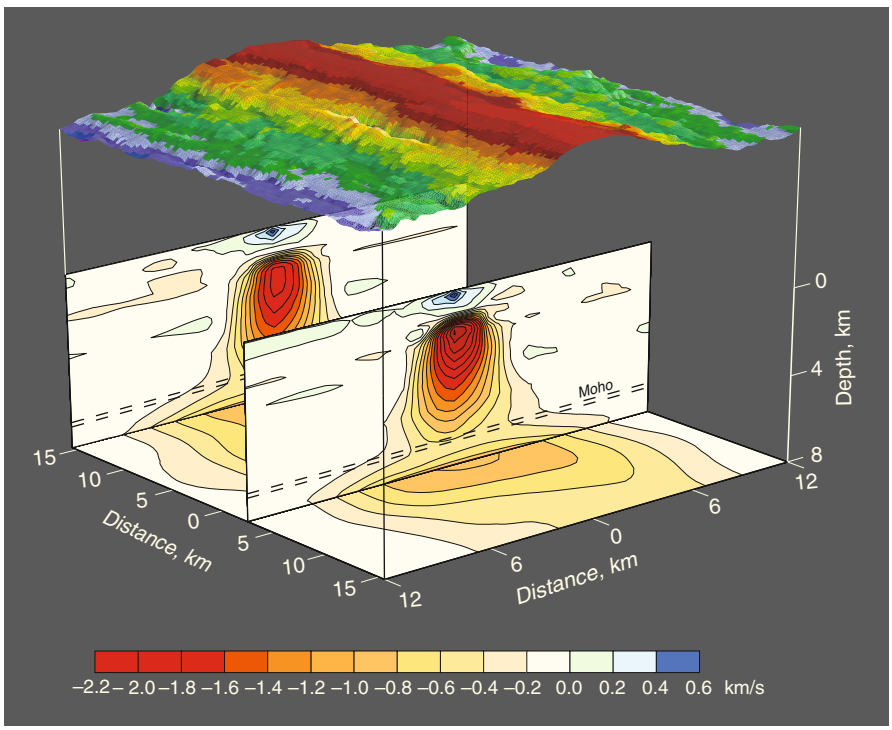
Seismic Structure at Mid-Ocean Ridges, Figure 3 Bathymetry of the East Pacific Rise (*left*) and tomographic image of the mantle 9 km beneath the seafloor (*right*). Dashed lines show plate boundary. Solid lines show locations of air gun shots fired at 500-m intervals. Squares and circles are locations of ocean bottom receivers. Contour interval on tomographic image of P-wave velocity is 0.1 km/s. Green lines with double arrowheads indicate fast direction for anisotropic wave propagation. Black lines with arrows indicate direction of relative motion between the Pacific and Cocos plates. Note that some of the slowest region is displaced from the plate boundary. After Toomey et al. (2007).

Crustal structure

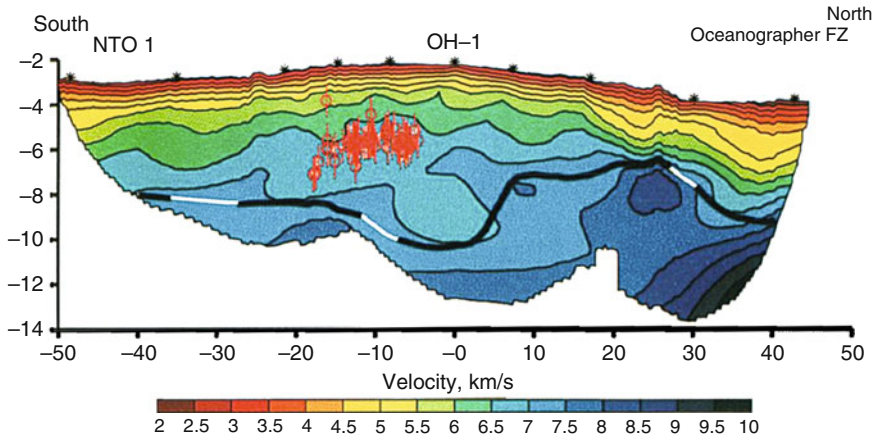
Basaltic melt migrates upward through the mantle and is focused at the ridge axis. The mechanisms for focusing are still not well known: there may be melt-rich channels at the base of the lithosphere that guide the melt upward and toward the ridge axis; there may be pressure gradients within the deforming mantle that help push the melt toward the axis; or there may be anisotropic cracks or dunite channels that form easy paths for melt migration. Once the melt reaches crustal levels, there may be redistribution along axis through dikes or a continuous magma chamber. It is clear from seismic studies, however, that there is very little magmatic addition to the crust outside the immediate vicinity of the ridge axis. The crust is essentially full thickness at the spreading center itself (Detrick et al., 1987).

The classic model of oceanic crust is a layer of extrusive basalts at the surface in the form of porous pillow basalts and sheet flows, underlain by a region of sheeted dikes that feed the extrusive layer from a magma chamber, and the lower crust consisting of gabbros that solidify from the magma chamber or underlying mush zone. The seismic structure has also commonly been described in terms of layers that have often been equated to the lithological layering: layer 2A is a low-velocity layer at the surface several hundred meters thick (layer 1 is sediments that are deposited on top of the basaltic crust); layer 2B is a transition region in which the P velocity increases rapidly downward; and layer 3 has low vertical velocity gradients and high P-wave velocities of 6.5–7.0 km/s. A number of investigations, however, have demonstrated that the seismologic layering does not correspond exactly to the lithological layering and that seismic structure is primarily controlled by porosity. The middle to lower crust, layer 3, contains both dikes and gabbro bodies.

At the axis of fast-spreading ridges, there usually is a low-velocity region only a few kilometers wide at most that represents a zone of partial melting that extends throughout the crust (Figure 4). At the top of this low-velocity region is a very low velocity layer, which is imaged as a prominent reflector in seismic reflection profiles at depths of 1.5–2.5 km on the East Pacific Rise (Detrick et al., 1987). The depth to this reflector tends to increase with decreasing spreading rate or proximity to a fracture zone and it is typically absent at slow-spreading ridges (Singh et al., 2006). The width varies from about 250 m to several kilometers. It is interpreted as the top of an axial magma chamber (AMC) or melt sill, with typical thickness less than 100 m. Modeling of P to S conversions in wide-angle reflections indicates that melt content in the AMC varies from nearly 100% to less than 30% along the ridge (Singh et al., 1998; Canales et al., 2006). Estimates of melt content in the deeper crustal mush zone (consisting



Seismic Structure at Mid-Ocean Ridges, Figure 4 Three-dimensional view of the anomalous P-wave velocity structure of the East Pacific Rise at 9°30′N (a one-dimensional reference model has been subtracted). From Dunn et al. (2000). Overlying bathymetry is shown at top, with red shallow to blue, deep. The red regions below the ridge axis correspond to areas of partial melting. The axial magma chamber is at the top of these anomalously slow regions.



Seismic Structure at Mid-Ocean Ridges, Figure 5 P-wave velocity structure along the Mid-Atlantic Ridge axis from a non-transform offset on the left to the Oceanographer fracture zone on the right. Crustal thickness from wide-angle reflections off the Moho is shown by thick solid line (open where there are gaps in coverage). The crust is thicker and the lower crust has lower velocity near the center of the segment. From Hooft et al. (2000). The red circles are locations of microearthquakes with uncertainties indicated by error bars. Location of ocean-bottom hydrophones indicated by asterisks.

of a mixture of melt and crystals) are of the order of 2–8% in the lower crust and 3–12% near the Moho (Dunn et al., 2000). The fact that the zone in which melt is present is so narrow even at fast-spreading ridges means that hydrothermal circulation must be very efficient in removing heat, as purely conductive cooling of the crust would result in a much wider zone of partial melting.

Although there are variations in structure along-axis at fast-spreading ridges, the along-axis variations are much more pronounced at segmented, slow-spreading ridges. Typically, the crust thins approaching transform offsets. Within the fracture zone itself, the basaltic crust may be as thin as 1 km or less, but seismically there may be an altered, fractured layer that is a few kilometers thick and characterized by unusually low velocities, so that it looks like crust. The low-velocity region in the fracture zone, however, probably is mostly mantle that is altered by interaction with water penetrating down cracks that are repeatedly opened by slip along the transform fault. Near the center of ridge segments between two transform offsets, the crust tends to be thicker and lower in velocity than elsewhere (Figure 5), suggesting that melt is preferentially delivered to the crust from the mantle at that point. The upper crust at slow spreading ridges is anisotropic, with P-waves traveling faster along axis than perpendicular to it, indicating that faults and fissures are preferentially aligned parallel to the spreading center (Barclay and Toomey, 2003).

Summary

The seismic velocity structure of mid-ocean ridges is controlled by crustal thickness, cracking or porosity, temperature, melt, and crystal orientation. Low-velocity regions in the mantle that are caused by high temperatures and partial melt indicate that melt is generated in a broad region beneath spreading centers. Asymmetries in the velocity structure show that upwelling and melting beneath ridges are strongly influenced by global mantle circulation and plate motions. Crustal low-velocity regions are concentrated very near the ridge axis, so melt must migrate both vertically and horizontally to the ridge axis from the broad melt production region. The narrowness of the low-velocity zone in the crust requires that hydrothermal circulation must rapidly cool the crust. Along-axis variations, particularly at slow-spreading ridges, suggest that melt is preferentially delivered to the center of ridge segments.

Bibliography

Barclay, A. H., and Toomey, D. R., 2003. Shear wave splitting and crustal anisotropy at the Mid-Atlantic Ridge, 35°N. *Journal of Geophysical Research*, **108**, doi:10.1029/2001JB000918.

Canales, J. P., Singh, S. C., Detrick, R. S., Carbotte, S. M., Harding, A., Kent, G. M., Diebold, J. B., Babcock, J., and Nedimovic, M. R., 2006. Seismic evidence for variations in axial magma chamber properties along the southern Juan de Fuca Ridge. *Earth and Planetary Science Letters*, **246**, 353–366.

Delorey, A. A., Dunn, R. A., and Gaherty, J. B., 2007. Surface wave tomography of the upper mantle beneath the Reykjanes ridge with implications for ridge-hot spot interaction. *Journal of Geophysical Research*, **112**, B08313, doi:10.1029/2006JB004785.

Detrick, R. S., Buhl, P., and Vera, E., 1987. Multi-channel seismic imaging of a crustal magma chamber along the East Pacific Rise. *Nature*, **326**, 35–41.

Dunn, R. A., Toomey, D. R., and Solomon, S. C., 2000. Three-dimensional seismic structure and physical properties of the crust and shallow mantle beneath the east pacific rise at 9 degrees 30'N. *Journal of Geophysical Research*, **105**, 23537–23555.

Evans, R. L., et al., 1999. Asymmetric electrical structure in the mantle beneath the east pacific rise at 17°S. *Science*, **286**, 756–759.

Hammond, W. C., and Toomey, D. R., 2003. Seismic velocity anisotropy and heterogeneity beneath the mantle electromagnetic and tomography experiment (MELT) region of the East Pacific Rise from analysis of P and S body waves. *Journal of Geophysical Research*, **108**, doi:10.1029/2002JB001789.

Harmon, N., Forsyth, D. W., and Weeraratne, D. S., 2009. Thickening of young Pacific lithosphere from high-resolution Rayleigh

wave tomography: a test of the conductive cooling model. *Earth and Planetary Science Letters*, **278**, 96–106.
Hooft, E. E. E., Detrick, R. S., Toomey, D. R., Collins, J. A., and Lin, J., 2000. Crustal thickness and structure along three contrasting spreading segments of the Mid-Atlantic Ridge, 33.5° – 35°N. *Journal of Geophysical Research*, **105**, 8205–8226.
MELT Seismic Team, 1998. Imaging the deep seismic structure beneath a Mid-Ocean Ridge: the MELT experiment. *Science*, **280**, 1215–1218.
Singh, S. C., Crawford, W. C., Carton, H., Seher, T., Combier, V., Cannat, M., Canales, J. P., Dusunur, D., Escartin, J., and Miranda, J. M., 2006. Discovery of a magma chamber and faults beneath a Mid-Atlantic Ridge hydrothermal field. *Nature*, **442**, 1029–1032.
Singh, S. C., Kent, G. M., Collier, J. S., Harding, A. J., and Orcutt, J. A., 1998. Melt to mush variations in crustal magma properties along the ridge crest at the southern East Pacific Rise. *Nature*, **394**, 874–878.
Toomey, D. R., Jousselin, D., Dunn, R. A., Wilcock, W. S. D., and Detrick, R. S., 2007. Skew of mantle upwelling beneath the East Pacific Rise governs segmentation. *Nature*, **446**, 409–414.
Wang, Y., Forsyth, D. W., and Savage, B., 2009. Convective upwelling in the mantle beneath the Gulf of California. *Nature*, **462**, 499–501, doi:10.1038/nature08552.

Cross-references

Crustal Reflectivity (Oceanic) and Magma Chamber
Earth's Structure, Upper Mantle
Lithosphere, Mechanical Properties
Lithosphere, Oceanic
Lithosphere, Oceanic: Thermal Structure
Ocean Bottom Seismics
Seafloor Spreading
Seismic Anisotropy
Seismic Data Acquisition and Processing
Seismic Tomography
Seismic Velocity-Density Relationships
Surface Waves
Traveltime Tomography Using Controlled-Source Seismic Data

SEISMIC TOMOGRAPHY

Guust Nolet
Geosciences Azur, Sophia Antipolis, France

Definition

The term tomography derives from the Greek τομος, or slice. "Seismic tomography" is used for a variety of methods that use transmitted seismic waves to estimate the spatial variations in properties (wave velocity, density, attenuation) inside the Earth, which are often represented as images of two-dimensional cross-sections or "slices." It is conceptually different from seismic migration, which uses reflected waves to image sharp discontinuities.

History

In 1971, P. Bois at the Institut Français de Pétrole was the first to suggest the tomographic method in order to locate the causes of delays in seismic waves between two boreholes. His paper predates many future developments but was written in French and remained largely unnoticed. In the mid-1970s, Keiti Aki from MIT applied a linear inversion to locate velocity heterogeneities beneath large nuclear monitoring arrays in Norway and Montana, and Harvard's Adam Dziewonski began interpreting the time residuals published by the International Seismological Center (ISC) in the UK in terms of global velocity anomalies.

In 1982, Guy Masters and his colleagues at the Scripps Institution of Oceanography discovered a strong degree-2 component in the geographical distribution of the slight shifts in the spectral peaks of the Earth's normal modes. Since then, the Earth's free oscillations have contributed to constrain the heterogeneity in the Earth at the longest wavelengths and as deep as the inner core. By 1984 John Woodhouse and Adam Dziewonski at Harvard published a first global model for shear velocity in the upper mantle based on long-period surface waves.

However, to image smaller scale anomalies, the shorter wavelengths of P and S-waves are indispensable. In particular, Steve Grand at the University of Texas, Rob van der Hilst and Wim Spakman at Utrecht University and Yoshio Fukao and colleagues at the University of Tokyo pioneered high-resolution body-wave tomography using iterative solvers for the huge systems of linearized equations and established in the early 1990s that some, but not all, subducting slabs are able to sink well into the depths of the lower mantle.

Thermal plumes in the lower mantle were for the first time reliably imaged in 2003 by Raffaella Montelli, using a new technique of finite-frequency tomography developed by Tony Dahlen and Guust Nolet and their collaborators at Princeton University.

For references and a detailed account of the history of seismic tomography see the reviews by Romanowicz (2003) and Rawlinson et al. (2010). Nolet (2008) provides a general introduction into the methods of seismic tomography, including the theoretical aspects that are here discussed only briefly.

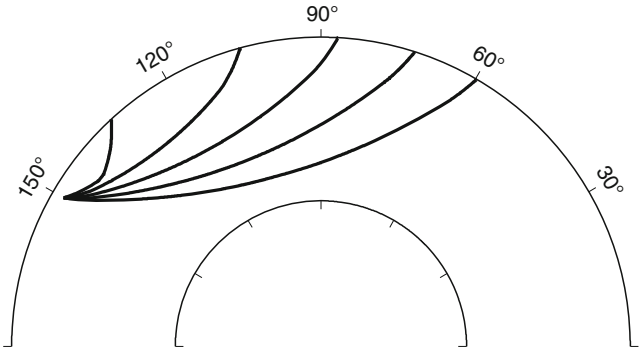
Onset times

Much of seismic tomography is based on estimating the arrival time of a seismic body wave by picking the "onset" of a phase on the seismogram, and interpreting the travel time T using the infinite-frequency approximation of ray theory:

$$T = \int_{\text{raypath}} \sigma(\boldsymbol{r}) \mathrm{d}s \tag{1}$$

where the raypath is determined by Snell's law and where σ is the inverse velocity v^{-1} of the wave at the location $\boldsymbol{r}$. The raypaths may be located between two boreholes, e.g., to monitor the exploitation of an oil or gas field, between several explosive sources and an array of seismographs at the surface (deep seismic sounding), between a seismogenic zone and a local array of seismographs, or between an earthquake and the global network of seismic stations.

The raypath is often approximated by the path calculated for a spherically symmetric Earth or a horizontally



Seismic Tomography, Figure 1 Examples of raypaths for a P wave in an equatorial cross-section from a hypothetical earthquake at a depth of 248 km located at 0°N, 150°E.

layered background model. Examples of global raypaths are given in Figure 1. This approximation is permitted because ray trajectories render the travel time stationary, such that small deviations in the true ray path location only cause second order errors in the travel time calculated with Equation 1, a property of rays known by the name of *Fermat's Principle*. In practice, one usually prefers to invert for the difference between the true Earth and a starting- or background model: $\delta\sigma(\boldsymbol{r}) = \sigma(\boldsymbol{r}) - \sigma_0(\boldsymbol{r})$, and use the difference between observed time T and its prediction T_0 from the background model as datum: $\delta T = T - T_0$:

$$\delta T = \int_{\text{raypath}} \delta\sigma(\boldsymbol{r})\mathrm{d}s \tag{2}$$

For strongly heterogeneous regions such as subduction zones, one may need to use three-dimensional ray tracing and an iterative approach using (Equation 2).

Model parameterization and inversion

The model $\delta\sigma(\boldsymbol{r})$ can be described by a finite number of parameters if it is developed in terms of a basis of M interpolating or "basis" functions $h_k(\boldsymbol{r})$:

$$\delta\sigma(r) = \sum_{k=1}^{M} m_k h_k(\boldsymbol{r}) \tag{3}$$

The basis functions may be represented by homogeneous cells, linear interpolators in a tetrahedral or other mesh, spherical harmonics, or 3D wavelets. Substitution of Equation 3 in Equation 2 gives a formal system of linearized equations for N estimated travel times $\delta T_i, i = 1, \ldots N$, arranged in a vector $\delta\boldsymbol{T}$:

$$\delta\boldsymbol{T} = \boldsymbol{A}\boldsymbol{m} \tag{4}$$

where $A_{ik} = \int_{\text{raypath}_i} h_k(\boldsymbol{r})\mathrm{d}\boldsymbol{s}$. Since many raypaths may overlap, the system (Equation 4) is usually overdetermined and needs to be solved by a least squares solver that minimizes χ^2, the length of the misfit vector weighted by the data standard error e_i:

$$\chi^2(\boldsymbol{m}) = \sum_{i=1}^{N} \left(\frac{\sum_{k=1}^{M} A_{ik} m_k - \delta T_i}{e_i} \right)^2 \tag{5}$$

Mathematically, this is accomplished by first dividing the system (Equation 4) by the standard error of the data (i.e., multiplying by the inverse square root of the covariance matrix $\boldsymbol{C}$ which is generally assumed to be diagonal), and backprojecting the system using the transpose $\boldsymbol{A}^T$ of $\boldsymbol{A}$:

$$\boldsymbol{A}^T\boldsymbol{C}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{m} = \boldsymbol{A}^T\boldsymbol{C}^{-\frac{1}{2}}\boldsymbol{\delta T} \tag{6}$$

One often scales the equations a priori such that the data have unit error $e_i = 1$ and $\boldsymbol{C} = \boldsymbol{I}$, which has the same effect. In the following we therefore ignore $\boldsymbol{C}$.

Usually, Equation 6 is at the same time overdetermined for some parameters and underdetermined for others (the determinant of $\boldsymbol{A}^T\boldsymbol{A}$ being zero): there are in that case infinitely many models that satisfy the data within the measurement uncertainty and one needs to regularize the solution. This can be done by choosing the solution that minimizes a weighted penalty of χ^2, model norm and model roughness, by strongly truncating a development in spherical harmonics, or by choosing the sparsest decomposition in wavelets that still satisfies the data. This generally involves a subjective choice between the degree of detail one allows in a model and the goodness of fit to the data. Generally, one aims for a χ^2 approximately equal to the number of data N – in other words, one attempts to fit the data at the level of about one standard deviation. Invariably, a trade-off exists between the detail allowed in the model and the statistical precision with which the model parameters can be determined: the sharper the detail, the more uncertain the m_k.

Regularization can also be done in a fundamentally different way by inverting for local averages in $\delta\sigma(\boldsymbol{r})$ that can be estimated linearly from the data with a specified variance, though this can only be done at the expense of a significant increase in computing time (Backus-Gilbert theory). In this case the trade-off is between the size of the averaging volume and the model variance: averages over larger volumes are determined with smaller statistical uncertainty.

Whatever method is used, the system (Equation 4) may be very large (e.g., 10^6 data for 10^5 unknown model parameters). Local parameterizations (cells, as opposed to spherical harmonics) render $\boldsymbol{A}$ sparse, and the system can be solved efficiently using iterative solvers that adapt to sparse matrices, such as LSQR.

Normal modes and surface waves

The eigenfrequencies ${}_n\omega_\ell^m$ of the Earth are characterized by three quantum numbers ℓ, m and n, related to the number of nodal surfaces in the displacement field of the Earth with latitude, longitude and depth, respectively. For a non-rotating, isotropic, spherically symmetric Earth the spectrum is degenerate in the sense that the frequency is

independent of the azimuthal order m. For the real Earth, a weak dependence on m splits each eigenfrequency into $2\ell + 1$ separate frequencies that are too closely spaced to be resolvable except for the very lowest angular order ℓ. Instead, a composite spectral line or "multiplet" is observed with a peak that depends on the location of the seismic station – a direct consequence of the fact that the $2\ell + 1$ single peaks have amplitudes that depend on geographical location by virtue of their spherical harmonic dependence on latitude and longitude.

Two major strategies exist to exploit the small fluctuations in the spectrum. Decomposing the free oscillation into surface waves traveling in opposite directions, ray theory may be used to establish a linear relationship between the heterogeneity along the great circle between source and receiver and the observed peak shift in the spectral line. At higher frequency we separate single passages of the surface wave and exploit the linear relationship between fluctuations in the observed phase velocity and the Earth's heterogeneity. In both cases, the relationship between the Earth's heterogeneity and the observed datum is a two-dimensional integral along the great circle of the form:

$$\delta\omega = \int_0^a \int_{\text{gc}} K(\boldsymbol{r})\delta\sigma(\boldsymbol{r})\mathrm{d}s\mathrm{d}r, \tag{7}$$

where the kernel $K(\boldsymbol{r})$ is computed using first order perturbation theory of the differential equations governing the Earth's free oscillations.

Alternatively, we may exploit the known distribution of amplitudes of single peaks over the surface of the Earth to invert for the location of these peaks ("mode splitting"). The small frequency shifts $\delta\omega^m$ are themselves the eigenvalues of a splitting matrix $\boldsymbol{H}$ of which the elements are linearly related to the variation of density and elastic parameters in the Earth. We can estimate $\boldsymbol{H}$ from the seismic data using autoregressive filtering techniques. This way we avoid any ray-theoretical approximations and obtain a three-dimensional integral constraint that can be used to solve for the large-scale variations in the Earth's density and elastic properties.

Finite-frequency tomography

Modern, broadband digital instrumentation allows for a robust estimation of the delay of an observed seismic wave $s_1(t)$ with respect to a theoretically predicted ("synthetic") waveform $s_2(t)$ or with respect to the same phase observed elsewhere, by locating the maximum in the cross-correlation $C(t)$ between the two signals:

$$\begin{aligned} \delta T &= \arg \max_{t\in[\tau_1,\tau_2]} C(t), \\ C(t) &= \int_{\tau_1}^{\tau_2} s_1(\tau)s_2(\tau - t)\,\mathrm{d}\tau\,, \end{aligned} \tag{8}$$

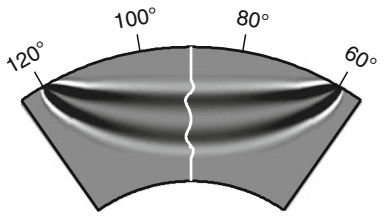
where the integration interval (τ_1, τ_2) extends over all times where the integrand is nonzero. The cross-correlation delay is thus fundamentally different from the delay measured by picking the onset of $s_1(t)$, because it represents an integral measure over a time window that is at least as long as the dominant period of $s_1(t)$. Heterogeneities inside the Earth may influence the integral by scattering waves in the direction of the recording station that arrive within the time window. The most serious consequence is that energy may diffract around a small heterogeneity, thus masking or dominating the slower or faster arrival that has crossed the heterogeneity. This phenomenon of "wavefront healing" biases observed delay times to zero, especially if the anomaly is located at some distance from the source and receiver.

Taking wave diffraction into account leads to a three-dimensional integral constraint of the form:

$$\delta T = \int_V K(\boldsymbol{r})\delta\sigma(\boldsymbol{r})\mathrm{d}^3\boldsymbol{r}\,, \tag{9}$$

where the kernel $K(\boldsymbol{r})$ can be efficiently calculated using ray theory for scattered waves. The effective volume of integration V is roughly equivalent to that of the Fresnel zone of the ray, where the sensitivity is largest. An example of a finite-frequency kernel is shown in Figure 2. Remarkably, the cross-correlation travel time is insensitive to perturbations in the Earth's properties at the location of the ray path. Because of this hole in the sensitivity and their general shape, finite-frequency kernels are often referred to as "banana-doughnut" kernels. Numerical tests have shown that the finite-frequency kernels accurately model the loss of signal caused by "wavefront healing," the gradual reduction of a delay caused by waves diffracting around a small heterogeneity. By estimating delays in different frequency bands, one gains information about the size of the heterogeneity in the Earth.

Though finite-frequency kernels were initially proposed to provide a better theoretical modeling of long-period signals affected by wavefront healing, they offer the advantage of being able to model the healing as a function of frequency, and thus obtain independent information on the size of the heterogeneity. Such "multiple-frequency" tomography may significantly increase



Seismic Tomography, Figure 2 The sensitivity of the cross-correlation time to the earth structure, for a long-period P wave from a surface source at 0°N, 120°E, recorded at a distance of 60°. The kernel is plotted in an equatorial cross-section of the mantle, the *grayscale* runs from large negative values (*black*) to positive (*white*). The sensitivity is zero at the location of the geometrical raypath, forming a "doughnut hole" in the banana-shaped kernel.

resolution (Sigloch et al., 2008). Finite-frequency sensitivity can also be formulated for the amplitude perturbations caused by attenuation and by focusing/defocusing (focusing cannot be handled with ray theory because ray theory does not model amplitudes correctly at low frequency and is very nonlinear at high frequency).

Instead of inverting for a cross-correlation delay or an amplitude perturbation, one can formulate a 3D sensitivity directly for the observed time series, though this has the disadvantage that the observed and predicted waveforms have to be close enough in phase that the phase difference can be adequately modeled by adding a small perturbation to the waveform itself (in fact modeling $e^{i\varphi} \approx 1 + i\varphi$). This limits the inversion to waveforms with small time mismatches, in contrast to the delays estimated through (Equation 8), which remain linear for large anomalies.

All types of data – delays, amplitudes and waveform mismatches – can be iteratively inverted using an adjoint approach, in which the best fitting model is sought along the gradient $\nabla_m \chi^2$ (Tromp et al., 2005), e.g., for delay times scaled to unit variance:

$$\nabla_m \chi^2 = \boldsymbol{A}^T(\boldsymbol{Am} - \delta \boldsymbol{T}) \tag{10}$$

For waveforms, the matrix A is in this case the matrix representation of a finite-difference or spectral-element algorithm that produces the predicted seismograms and A^T is the adjoint operator that projects the seismograms back in time. Since this allows one to backproject all residual seismograms for one earthquake with only one calculation, and since only the product of the adjoint matrix and the residual data vector is needed, this allows us to avoid computing the partial derivatives separately for each datum δT_i with respect to each model parameter m_k. The adjoint approach has some advantages over a direct matrix inversion for large-scale problems, in particular those with few sources and many stations, because it does not require the storage of a large matrix, which is especially important for waveform inversions.

Summary

The equations of seismic tomography are integral equations; the observations are weighted averages of the properties of the Earth (Equation 9). Though often treated as line integrals, assuming ray theory is valid, even a 1 Hz P wave senses a volume or Fresnel zone inside the Earth that is several hundred kilometers wide. The inversion of integral equations demands care to avoid that noise (errors in the data) is interpreted as small differences in such averages which may require large spatial variations in the Earth's properties. This imposes a fundamental limitation to the resolving power of seismic waves that even finite-frequency theory cannot completely avoid. The theoretically best available horizontal resolution, using the highest observable seismic frequencies, is of the order of hundred kilometers in the lower mantle. In practice, this lower limit is not yet reached because of the limited coverage of the Earth's surface with seismic stations. On land, the coverage can be improved markedly using temporary deployments of seismic arrays, but this is much more difficult and expensive to do in the oceanic domain. Major improvements on a global scale are therefore to be expected only if we solve the problem of oceanic seismometry.

Bibliography

Nolet, G., 2008. *A Breviary of Seismic Tomography*. Cambridge: Cambridge University Press.

Rawlinson, N., Pozgay, S., and Fishwick, S., 2010. Seismic tomography: a window into deep earth. *Physics of the Earth and Planetary Interiors*, **178**, 101–135.

Romanowicz, B., 2003. Global mantle tomography: progress status in the past 10 years. *Annual Review of Earth and Planetary Sciences*, **31**, 303–328.

Sigloch, K., McQuarrie, N., and Nolet, G., 2008. Two-stage subduction history under North America inferred from multiple-frequency tomography. *Nature Geoscience*, **1**, 458–462.

Tromp, J., Tape, C., and Liu, Q., 2005. Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels. *Geophysical Journal International*, **160**, 195–216.

Cross-references

Earth, Density Distribution
Earth's Structure, Core
Earth's Structure, Lower Mantle
Earth's Structure, Upper Mantle
Free Oscillations of the Earth
Gravity Field of the Earth
Inverse Theory, Global Optimization
Inverse Theory, Linear
Inverse Theory, Monte Carlo Method
Mantle Plumes
Plate Driving Forces
Propagation of Elastic Waves: Fundamentals
Seismic Imaging, Overview

SEISMIC VELOCITY-DENSITY RELATIONSHIPS

Kalachand Sain
National Geophysical Research Institute, Council of Scientific & Industrial Research (CSIR), Hyderabad, India

Definition

Seismic velocity is defined as the speed with which a seismic wave travels through a medium and is measured in km/s. The density is the mass per unit volume and is expressed in g/cc.

Types of seismic waves

The *body wave* and the *surface wave* are the main seismic waves. The body waves are the longitudinal (P) and the shear (S) waves in which the particles vibrate parallel and perpendicular to the direction of wave propagation respectively. The velocities of P-wave (V_P) and S-wave (V_S) are related to the density, ρ as

$$V_P = \sqrt{\frac{\kappa + \frac{4}{3}\mu}{\rho}} \quad (1)$$

$$V_S = \sqrt{\frac{\mu}{\rho}} \quad (2)$$

where κ and μ are the bulk and rigidity moduli respectively. It appears that the seismic velocity and density are inversely related, but the denser rocks are characterized by higher velocities compared to lighter rocks because the density also depends on κ and μ. The halite with low density (1.8 g/cc) and high V_P (4.5 km/s) shows an exception. Generally velocities and densities increase with depth. The inversion of this trend also is observed due to pore pressure, fluids, or sediments below volcanics.

General relation between velocity and density

(a) Birch's law – To fit measurements from crustal and mantle rocks, Birch (1961) established a linear relationship between V_P and ρ as

$$V_P = a + b\,\rho \quad (3)$$

where a and b are empirical parameters.

(b) Gardner relation – Gardner et al. (1974) conducted field and laboratory measurements on saturated sedimentary rocks and determined the velocity–density relation as

$$\rho = aV_P^b \quad (4)$$

where $a = 0.31$ and b is 0.25. Major sedimentary rocks fall within a narrow corridor around the Gardner curve. Coals, anhydrite, and salts exhibit large deviations from this trend. The deviation increases with porosity, and becomes higher for lower densities.

Estimation of seismic velocity and density from surface measurements provides important inputs to understand the structure and tectonics of the earth. Since gravity is a potential field, velocity structure is often used to constrain the gravity modeling (Behera et al., 2004) by converting the velocity into density using above formulas. Use of Ludwig et al. (1970) formula to relate V_P with ρ is also very common in crustal studies.

Bibliography

Behera, L., Sain, K., and Reddy, P. R., 2004. Evidence of underplating from seismic and gravity studies in the Mahanadi delta and its tectonic significance. *Journal of Geophysical Research*, **109**, 1–25.

Birch, F., 1961. The velocity of compressional waves in rocks to 10 kilobars. *Journal of Geophysical Research*, **66**, 2199–2224.

Gardner, G. H. F., Gardner, L. W., and Gregory, A. R., 1974. Formation velocity and density – the diagnostic basics for stratigraphic traps. *Geophysics*, **39**, 770–780.

Ludwig, W. J., Nafe, J. E., and Drake, C. L., 1970. Seismic refraction. In Maxwell, A. E. (ed.), *The Sea*. New York: Wiley-Interscience, Vol. 4, Part 1, pp. 53–84.

Cross-references

Body Waves
Surface Waves

SEISMIC VELOCITY-TEMPERATURE RELATIONSHIPS

Kalachand Sain
National Geophysical Research Institute, Council of Scientific & Industrial Research (CSIR), Hyderabad, India

Definition

The seismic velocities of subsurface rocks decrease with temperature but experience opposite effects due to increase in pressure (depth).

Velocity–temperature equation

On an average, the pressure increases at a rate of 30 MPa/km and the temperature raises at a rate of 25°C/km from surface to a few tens of kilometers. Further deep down, this increasing rate of temperature gradient decreases to small values. Since seismic velocity (V) decreases with temperature (T) and increases with pressure (P), we need to know the combined effects of P and T for interpreting seismic velocities at different depths. The generalized relation for the variation of velocity with depth, Z can be expressed as

$$\frac{dV}{dZ} = \left(\frac{\partial V}{\partial P}\right)_T \frac{dP}{dZ} + \left(\frac{\partial V}{\partial T}\right)_P \frac{dT}{dZ} \quad (1)$$

where $\left(\frac{\partial V}{\partial P}\right)_T$ denotes the change in velocity with pressure at constant temperature (isotherm) and $\left(\frac{\partial V}{\partial T}\right)_P$ is the change in velocity with temperature at constant pressure (isobar). $\frac{dP}{dZ}$ and $\frac{dT}{dZ}$ are the vertical pressure and temperature gradients respectively.

To make petrological inferences, the crustal and lithospheric seismic velocities are to be corrected to the experimental reference values (i.e., constant pressure of 100 MPa and room temperature of 20°C). Rybach and Buntebarth (1984) have defined the correction factor (f) to the P-wave velocity (V_P) at a given P and T as

$$V_P(20^\circ\text{C}, 100\text{ MPa}) = V_P(P,T)f = V_P\left[1 + \Delta T\frac{\partial V_P}{V_P \partial T} - \Delta P\frac{\partial V_P}{V_P \partial P}\right] \quad (2)$$

As per Equation 2, the corrections to the field velocities at depths from 1 to 50 km are of the order of ±0.1 km/s.

Velocity variation with temperature for some reservoir rocks

The change in seismic velocities with temperature also depends on the saturation of rock. The results of Wang and Nur (1990) in heavy oil sands and hydrocarbon-saturated

rocks show that both V_P and V_S decrease with increasing temperature. Since V_S is not affected by fluids, the decrease in V_S is due to changes in rock frame and in rock fluid interactions. The V_P in heavy oil sands (Tosaya et al., 1987) shows a dramatic decrease. As temperature increases from 25°C to 125°C, V_P can drop by 35% to almost 90%. Heavy oils are highly viscous and a strong interfacial force exists between oil and rock grains. The viscosity of oil and interfacial force decrease due to rise in temperature, which decreases the rigidity and bulk modulus leading to reduction in seismic velocities. For temperature up to about 150°C, changes in pore fluid properties play dominant role in changing the velocity.

Bibliography

Rybach, L., and Buntebarth, G., 1984. The variation of heat generation, density and seismic velocity with rock type in the continental lithosphere. *Tectonophysics*, **103**, 335–344.

Tosaya, C., Nur, A., Vo-Thanh, D., and Da Prat, G., 1987. Laboratory seismic method for remote monitoring of thermal EOR. *SPE Reservoir Engineering*, **2**, 238–242.

Wang, Z., and Nur, A., 1990. Wave velocities in hydrocarbon saturated rocks: experimental results. *Geophysics*, **55**, 723–733.

SEISMIC WAVE PROPAGATION IN REAL MEDIA: NUMERICAL MODELING APPROACHES

Ursula Iturrarán-Viveros[1], Francisco J. Sánchez-Sesma[2]
[1]Departamento de Matemáticas, Facultad de Ciencias, Universidad Nacional Autónoma de México, Ciudad Universitaria, México DF, Mexico
[2]Instituto de Ingeniería, Universidad Nacional Autónoma de México, Ciudad Universitaria, México DF, Mexico

Synonyms

Elastic waves in homogeneous and inhomogeneous media

Definition

The Earth is a complex medium containing heterogeneities and scattering structures over many scales. The influence of heterogeneities on seismic wave propagation is therefore studied intensively by numerical modeling to explain observations. Coming along with the enormous increase of computational power, the development of numerical techniques for the accurate calculation of synthetic seismograms in three-dimensional (3-D) models of the Earth has been the subject of a continuous effort in the last 30 years. This chapter presents a glimpse of several of these numerical methods.

Introduction

In 1969, Keiiti Aki first focused attention on the appearance of continuous wave trains (called coda), in the tail portion of individual seismograms of local earthquakes as a direct evidence of the random heterogeneity of the lithosphere. Models for seismic wave propagation through inhomogeneous elastic media have been developed using deterministic approaches such as model theory for layered media. In 1954, Norman Haskell made a great contribution to theoretical geophysics with his famous paper in which he showed how seismic surface waves could be computed for an Earth modeled by an arbitrary number of plane parallel layers, each one with arbitrarily prescribed physical properties. His work was based on Thomson's work and it is now known as the Thomson–Haskell method. In his pioneering work Claerbout (1968) recognized the solution to the inverse problem of determining the medium from the seismogram. Numerical modeling of seismic wave propagation in an inhomogeneous media is becoming more important in seismic exploration. A wide variety of recent numerical strategies are concerned with treating Earth models of successively increasing complexity. The realistic computational models can include anisotropic media, nonplanar interfaces between layers and blocks, velocity/density/quality-factor gradients inside layers, and often with free-surface topography. In particular, the rheology of the medium should allow for realistic attenuation. We can divide the different approaches to numerically solve the wave equation in complex media into three main classes: *direct methods*, *integral equation methods*, and *ray tracing methods*, see Table 1 and Carcione et al. (2002). In order to solve the wave equation using *direct methods*, the geological model needs to be approximated by a numerical mesh. There are no restrictions about material variability and different rheologies can be implemented. These methods give solution to the wave equation in the time domain. They can be very accurate provided a sufficiently fine grid, but they are computationally more expensive than analytical or ray tracing methods. *Integral equation methods* are based on Huygens' principle. These methods formulate the solution of a problem in terms of values at the domain's boundary. Integral equation methods have an interesting conceptual advantage over direct methods, which is the reduction of one space dimension. Moreover, integral equation methods (in the frequency domain) do not need absorbing boundaries. They match easily the boundary conditions and do not suffer from grid

Seismic Wave Propagation in Real Media: Numerical Modeling Approaches, Table 1 Classification of numerical methods to model seismic wave propagation in real media

Direct methods	Finite differences (FD), pseudospectral (PS), finite element (FE), spectral element method (SEM), finite volume (FV), discontinuous Galerkin (DG), discrete element method (DEM)
Integral equation methods	Boundary element method (BEM), indirect boundary element method (IBEM), fast multipole method (FMM), domain integral methods (DIM)
Asymptotic or ray tracing methods	Gaussian beams

dispersion. However, these methods require the knowledge of Green's function which is limited to a few cases (see Kausel, 2006). *Ray tracing* or *Asymptotic methods* are approximative because they do not take into account the complete wavefield. The wavefield is considered as an ensemble of certain events, each arriving at certain traveltime and having certain amplitude.

Finite-difference method

Among direct methods we have those that solve the strong form of the wave equation (FD, PS) and those that solve the weak form, which are also known as variational methods (such as FE). The strong formulation works directly with the equation of motion and the boundaries are prescribed in their differential form. The weak formulation instead uses an integral form of the equation of motion weighting the error by trial functions. The natural boundary conditions are automatically satisfied in the weak formulation. The basic idea of finite-difference methods is to compute the partial derivatives by an approximation based on Taylor series expansions of functions near the point of interest. Forward and backward time derivatives lead to implicit or explicit schemes. In general, an implicit method is computationally more expensive than an explicit method, due to the requirement of solving large matrix equations. When applying finite differences in the implicit scheme only for 1-D problems, the matrix to be inverted is tridiagonal, which is very convenient in terms of computational cost. However, this is not the case for 2-D problems. Nevertheless, using the method of dimensional splitting, one can replace the multi-dimensional problem by a succession of much simpler one-dimensional problems and as a result it is possible to build tridiagonal matrices and combine their results to compute the desired time derivative. This ingenious approach is described in Bayliss et al. (1986), Mufti (1985) and Emerman et al. (1982). More recently Liu & Sen (2009) have obtained implicit finite-difference formulae derived from fractional expansion of derivatives which form a system of equations that form a tridiagonal matrix. Obviously, the explicit schemes are computationally simpler. Therefore a vast majority of earthquake ground motion modeling and exploration seismology studies use explicit FD schemes. Virieux in a couple of papers (1984) and (1986) for SH and P-SV cases, respectively, was the first to introduce explicit time domain staggered FD in seismology. This method is easy to implement, to introduce different kinds of sources, and to consider viscoelasticity. In addition, the local nature of finite-difference operators makes the method suitable for parallelization. In seismic applications, the following velocity-stress formulation is widely used:

$$\begin{aligned} \frac{\partial \sigma_{ij}}{\partial x_j} + f_i &= \rho \frac{\partial v_i}{\partial t} \\ \frac{\partial \sigma_{ij}}{\partial t} &= c_{ijkl} \frac{\partial e_{kl}}{\partial t} \end{aligned} \tag{1}$$

where the first time derivative of the displacement u is the velocity v, σ_{ij} is the stress, $\partial_t e_{kl} = (v_{k,l} + v_{l,k})/2$, e_{kl} being the strain, c_{ijkl} is the fourth-order stiffness tensor with elastic coefficients for linear elastic solids, and f_i is the ith component of the body force vector $\mathbf{f}$ that can be written in terms of the moment tensor as follows:

$$\mathrm{f} = -\mathbf{M} \cdot \nabla \delta(x - x_s) S(t) \tag{2}$$

where x_s denotes the source position δ is the Dirac delta and $S(t)$ denotes the source time function. Evaluating the second time derivative of Equation 1, i.e., $\ddot{u} = \partial_t v$ (double dot denotes twice time derivative), at times $(n + 1)\Delta t$ and $(n–1)\ \Delta t$ by a Taylor expansion yields

$$\frac{\partial^2 u^n}{\partial t^2} = \frac{1}{(\Delta t)^2} \times \left[u^{n+1} - 2u^n + u^{n-1} - 2 \sum_{l=2}^{L} \frac{(\Delta t)^{2l}}{(2l)!} \frac{\partial^{2l} u^n}{\partial t^{2l}} \right] \tag{3}$$

when applied Equation 3 to numerically approximate spatial and time derivatives in the 1-D scalar wave equation

$$\frac{1}{\beta^2} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} \tag{4}$$

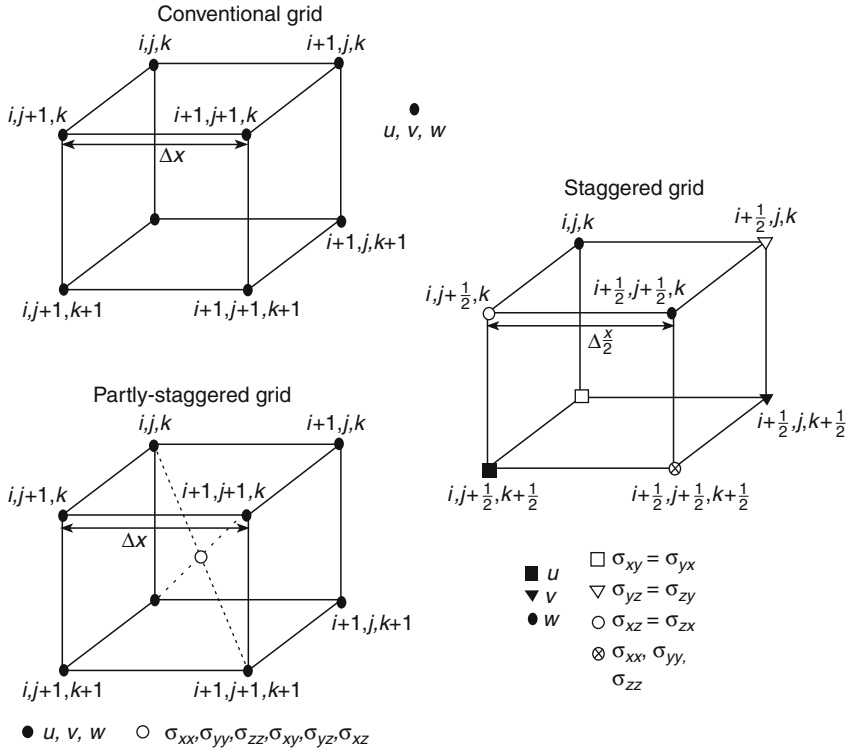
we obtain the Lax-Wendroff scheme

$$\frac{1}{\beta^2} \frac{u_i^{n+1} - 2u_i^n + u_i^{n-1}}{(\Delta t)^2} = \frac{u_{i+1}^n - 2u_i^n + 2u_{i-1}^n}{(\Delta x)^2} \tag{5}$$

This equation represents the discretization of Equation 4 taking into account that we have a Cartesian system of coordinates (x,y,z) with a fourth-order computational domain taking into account the time, i.e., (x,y,z,t). Considering the spatial-time grid given by a set of discrete points (x_i,y_j,z_k,t_n) given by $x_i = x_0 + i\Delta x$, $y_j = y_0 + j\Delta y$, $z_k = z_0 + k\Delta z$, $t_n = t_0 + n\Delta t$; $i,j,k,n = 1,2,...$ The spatial increments Δx, Δy, and Δz are usually referred to grid spacings, (x_0, y_0, z_0, t_0) are initial values for these variables, while Δt is the time step. The value of a function u at a grid position (x_i, y_j, z_k, t_n) is given by $u_{i,j,k}^n = u(x_i, y_j, z_k, t_n)$. It is possible to have nonuniform grids and the other important consideration is whether all functions are approximated at the same grid positions. The conventional, partly staggered, and staggered grids are illustrated in Figure 1. An important observation is pointed out for finite differences solutions of hyperbolic equations. Accuracy increases with increasing Δt, for central differences, up to a limit defined by the Courant-Friederickson-Lewy stability condition (assuming $\Delta x = \Delta z$, in 2-D):

$$\frac{V_{\max} \Delta t}{\Delta x} \leq \frac{1}{\sqrt{2}} \tag{6}$$

where $V_{\max}$ is the maximum velocity considered in the model.



Seismic Wave Propagation in Real Media: Numerical Modeling Approaches, Figure 1 Spatial grid cells in the conventional, partly staggered, and staggered grids (see Moczo et al., 2007).

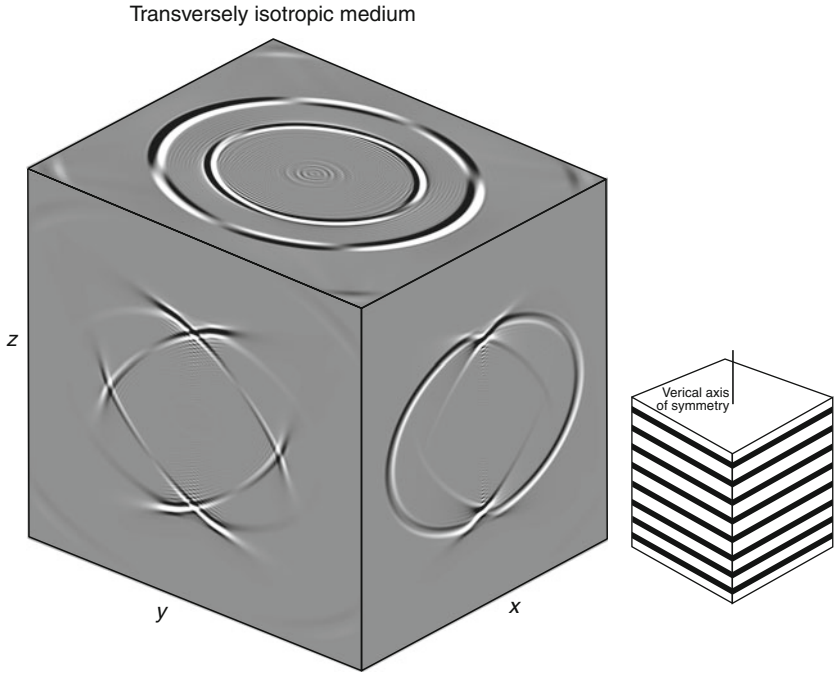
The choice among the different schemes is made to improve accuracy, reduce memory requirements, and to deal with anisotropy media. Recently, the use of the partly staggered grid was upgraded by Saenger et al. (2000) and Saenger & Bohlen (2004). The authors called the grid rotated staggered grid since they obtained the spatial FD operator by the rotation of the standard staggered-grid operator. They used this scheme to successfully simulate media with heterogeneities (cracks, pores, free surface, and anisotropy). Rotated staggered grids have been also used to model planar and nonplanar dynamic rupture models, see Cruz-Atienza et al. (2007). The main limitations of FD are the extensive consumption of computational resources in terms of both core memory and CPU time. However, Furumura and Chen (2005) performed large-scale parallel FD simulations of seismic wave propagation to model strong ground motions during damaging earthquakes. Parallel FD simulation allowed Saito and Furumura (2009) to successfully simulate tsunami generation. A complete review of FD methods can be found in Moczo et al. (2007).

Anisotropy

Anisotropy is caused by structural alignment of elements that are smaller than the wavelength employed for seismic data acquisition. Examples of structures that cause anisotropic effects are fine layering, tectonic grain, and orientation and intensity of cracking. One way to include these effects in the model is to consider that elastic properties at a given point might vary with direction. Therefore, by including anisotropy in the model, alignments of small scale hetegoeneities are captured. The alignment effects might lead to important petrophysical information. To consider anisotropy in a FD simulation one needs to define c_{ijkl} the fourth-order stiffness tensor in Equation 1. For example, to model the effect of layering (see Figure 2) one can consider the following stiffness tensor:

$$C = \begin{bmatrix} c_{11} = 66.6 & c_{12} = 19.7 & c_{13} = 39.4 & 0 & 0 & 0 \\ c_{12} & c_{11} & c_{13} & 0 & 0 & 0 \\ c_{13} & c_{13} & c_{33} = 39.6 & 0 & 0 & 0 \\ 0 & 0 & 0 & c_{44} = 10.9 & 0 & 0 \\ 0 & 0 & 0 & 0 & c_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & c_{66} \end{bmatrix} \tag{7}$$

where $c_{66} = (c_{11} - c_{12})/2$, the density given by $\rho = 2590\ \mathrm{Kg/m^3}$, and the units for the elastic constant are in GPa. The example represents the Mesaverde clay shale (see Thomsen, 1986 and Carcione et al., 1992). This kind of stiffness tensor with these five elastic constants characterizes a medium of class IV according to the classification given by Payton (1983) and it is called transversely



Seismic Wave Propagation in Real Media: Numerical Modeling Approaches, Figure 2 (*Left*). Illustration of a transversely isotropic medium. The example represents the Mesaverde clay shale. The six elastic constants (for the stifness tensor in Equation 1) that generate this medium are given by c_{11} = 66.6 GPa, c_{12} = 19.7 GPa, c_{13} = 39.4 GPa, c_{33} =39.9 GPa, c_{44} = 10.9 GPa., with $c_{66} = (c_{11} - c_{12})/2$ and the density ρ = 2590 Kg/m^3. (*Right*) A layered medium that represents the anisotropic effect.

isotropic medium (since it has the vertical z-axis as a symmetrical axis). Notice that due to the anisotropy, in Figure 2, and the wave fronts do not propagate at the same speed in all directions (see the different wave fronts in plane cuts: xy-plane, xz-plane and yz-plane).

Pseudospectral method

The pseudospectral method (PS) is base mainly on Fourier and Chebyshev differential operators. The spatial derivative is computed in the frequency-wavenumber domain as follows:

$$\frac{\partial u(x,t)}{\partial x} = \frac{1}{2\pi} \int_{-\infty}^{\infty} u(x,t) ike^{ikx} dk \tag{8}$$

see Kreiss and Oliger (1972), Kosloff and Baysal (1982), Furumura et al. (1998). The PS with Fourier has the disadvantage of periodic properties, called wraparound. Both Fourier and Chebyshev are accurate up to the maximum wavenumber of the mesh that corresponds to a spatial wavelength of two grid points (at maximum grid spacing for the Chebyshev operator). Like FD methods, they are also unable to model surface waves with the same accuracy as body waves because of the one-way treatment that needs to be performed in order to implement the free-surface condition (e.g., Carcione, 1994).

Finite element and spectral element methods

The finite element (FE) method is more efficient than FD or PS when dealing with complex geometries (irregular interfaces), heterogeneous media, and handling boundary conditions. This method is also applicable with inelastic constitutive models. Recently, Bielak et al. (2005) have developed an efficient FE-based computational platform for large-scale ground motions. Low approximation orders may lead to a large numerical dispersion, see Marfurt (1984). As a result, mesh refinement is required to reduce numerical dispersion but may lead to a large numerical cost even if parallelization is possible. High-order FE methods became more efficient when the spectral element method (SEM) appeared using Chebyshev polynomials as interpolating functions, see Seriani et al. (1992), Priolo et al. (1994). Komatitsch and Villotte (1998), and Komatitsch and Tromp (1999). This method introduced the use of the Lagrange polynomials as interpolating functions which produces a diagonal mass matrix when using the Gauss-Lobatto-Legendre quadrature.

Discontinuous Galerkin method

The Discontinuous Galerkin method (DG) is a technique that uses discontinuous basis functions to formulate a Galerkin approximation. Given a mesh of the analysis domain, the DG method approximates the solution within each element by a function from a low-dimensional vector

space of functions, e.g., as a linear combination of basis functions like polynomials. For a pair of adjacent mesh elements, the approximate solution computed in the interior of the elements does not have to agree on their common boundary. Some of the advantages of this method are:

- It can sharply capture solution discontinuities relative to a computational mesh.
- It simplifies adaptation since inter-element continuity is neither required for mesh refinement and coarsening.
- It conserves the appropriate physical quantities (e.g., mass, momentum, and energy) on an element-by-element basis.
- It can handle problems in complex geometries to high order.
- Regardless of order, it has a simple communication pattern to elements sharing a common face that simplifies parallel computation. With a discontinuous basis, the DG method produces more unknowns for a given order of accuracy than traditional finite element or finite volume methods, which may lead to some inefficiency. The DG method is harder when applied to unstructured meshes; in particular, it is harder to formulate limiting strategies to reduce spurious oscillations when high-order methods are used. Nevertheless, Käser & Dumbster (2006) have successfully implemented the method for elastic wave propagation on unstructured meshes.

For an introduction to the key ideas, basic analysis, and efficient implementation of discontinuous Galerkin methods and a review of the state of the art, we refer the reader to the book by Hesthaven & Warburton (2008).

Integral equation methods

Boundary element and indirect boundary element methods (BEM & IBEM)

Among Integral equation methods it is the boundary element method. The formulation of direct boundary integral equation method (BIE) in elastodynamics is back to the pioneering work of Somigliana (1886). Considering a volume of elastic material V bounded by a surface S, the displacement field u_i at a point $\mathbf{x}$ and time t can be expressed as a function of the values of the displacement and traction t_j along the boundary through Somigliana's representation theorem:

$$cu_i(\xi,t) = \int_0^t \mathrm{d}\tau \int_s \left[t_j(\xi,\tau)G_{ij}(\xi,t,x,\tau) - u_j(\xi,\tau)\Sigma_{jik}(\xi,t,x,\tau)n_k(\xi)\right]\mathrm{d}S(\xi) \tag{9}$$

where $G_{ji}(\xi,t,\mathbf{x},\tau)$ and $\Sigma_{jik}(\xi,t,\mathbf{x},\tau)$ are the responses in terms of displacement and stress, time t and point ξ of an infinite homogeneous medium to a unit force impulse, applied at time τ at point $\mathbf{x}$ in the direction i, $n_k(\xi)$ is the normal boundary pointing outside V. The volumetric body sources are assumed to be null, but if present, their contribution can be easily added as follows:

$$cu_i(\xi,t) = \int_0^t \mathrm{d}\tau \int_s \left[t_j(\xi,\tau)G_{ij}(\xi,t,x,\tau) - u_j(\xi,\tau)\Sigma_{jik}(\xi,t,x,\tau)n_k(\xi)\right]\mathrm{d}S(\xi) + \int_v^t \mathrm{d}\tau \int_V G_{ij}(\xi,t,x,\tau)f_i(x,\tau)\mathrm{d}V \tag{10}$$

where

$$c = \begin{cases} 1, & \xi \in \mathrm{V} \\ 1/2, & \xi \in \mathrm{S} \\ 0, & \xi \notin \mathrm{V} \end{cases} \tag{11}$$

and assuming that S has a smooth boundaries. The values of c come from the volume integration of the product $\delta(\mathbf{x}-\xi, t\text{-}\tau)\,\mathbf{u}_j(\xi,\tau)$ and the value accounts for the integration of the Dirac's delta and its position. The Green's function for the homogeneous, isotropic, elastic media can be found in Aki & Richards (1980).

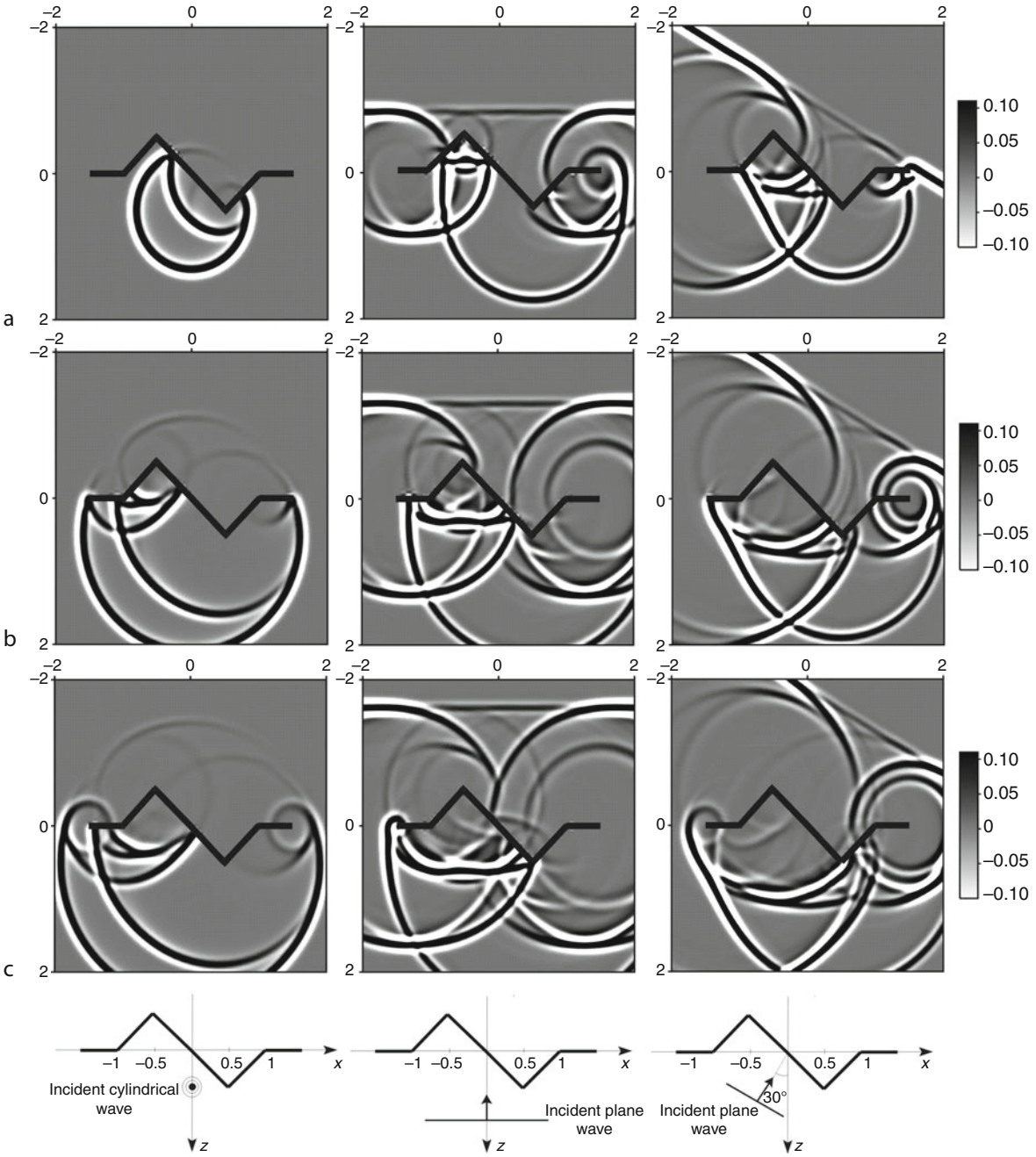
The indirect formulation of the elastodynamic problem expresses the wavefield as an integral over the boundary of elementary source radiations as follows:

$$u_i(x,t) = \int_0^t \mathrm{d}\tau \int_s \phi_j(\xi,\tau)G_{ij}(x,t,\xi,\tau)\mathrm{d}S(\xi) \tag{12}$$

where $\phi_j(\xi,\tau)$ denotes τ the force density distribution applied at time t at point ξ of the surface S. In this equation we have assumed that the volumic forces are null. If present, their contribution needs to be added. Equation 10 can be seen as the mathematical transcription of Huygens' principle, which states that each point of diffracting boundary acts as a secondary source of wave radiation. This principle applies to any wavefront as well. Equation 12 is also the mathematical expression of the exploding reflector concept widely used in seismic exploration (see Claerbout, 1985). The fictitious distribution ϕ_j is an intermediate variable which needs to be solved in the boundary integral equations before the wavefield can be computed. This is the reason why this formulation is called indirect.

The discretization of the boundaries leads to the discretization of the BIE, and the application of boundary conditions transforms BIE into a system of linear equations which in general is not symmetric. The resolution of this system is usually done implicitly and in the frequency domain. Comprehensive intorduction of integral equation methods can be found in the books by Bonnet (1999), Dominguez (1993), and Manolis & Beskos (1988). In seismology, the BEM is used to study the effect of irregular topography on earthquake ground motion, in

2-D some other works deal with 3-D problems. This method has been also used to simulate elastic wave propagation in media with cracks, for example, on Figure 3 we have an irregular crack strike by a cylindrical SH incident wave (first column) and the same crack strike by a plane SH incident wave with 0^o and 30^o, measured with respect to the z-axis (second and third columns, respectively), see Iturrarán-Viveros et al., 2005; Pointer et al., 1998. A complete review on the use of BEM for seismic problems can be found in Bouchon and Sánchez-Sesma (2007).



Seismic Wave Propagation in Real Media: Numerical Modeling Approaches, Figure 3 Snapshots of displacements for: (*left*) A cylindrical SH wave striking an irregular crack. (*middle*) A plane SH wave with incident angle 0^o striking an irregular crack. (*right*) A plane SH wave with incident angle 30^o striking an irregular crack. The set of snapshots are taken at (**a**) $t = 1.25$ s (**b**) $t = 1.71$ s and (**c**) $t = 2.03$ s.

Fast multipole method in elastodynamics

Two types of fast multipole method (FMM) are available for elastodynamics in the frequency domain. The first one is the low frequency FMM. As for static cases, the complexity of this algorithm is $O(N)$. This complexity is due to the fact that the wavelength is much larger than the domain size. On the other hand, if the wavelength is shorter than the geometrical feature, the complexity of low frequency FMM increases to $O(N^2)$ and the method is not efficient any longer. For this reason, computational efficiency of fast BEMs in the mid-frequency regime is enhanced by using the so-called diagonal form for the Helmholtz Green's function, proposed by Rokhlin (1990), with a complexity of O (N log N). The upper limit stems from the fact that the size N becomes intractable at high frequencies, but the diagonal form also breaks down at very low frequencies and must be replaced with other types of expansions. The method for low frequencies was developed by Fujiwara (1998). The first 3-D implementation was proposed by Fujiwara (2000) using a multi-level and diagonal form. The author presents some low frequency seismic oriented examples. More recently Chaillat et al. (2009) also addressed problems in 3-D. Yoshida (2001) proposed a low frequency FMM for crack problems in 3-D.

Analytic solutions for nonhomogeneous media

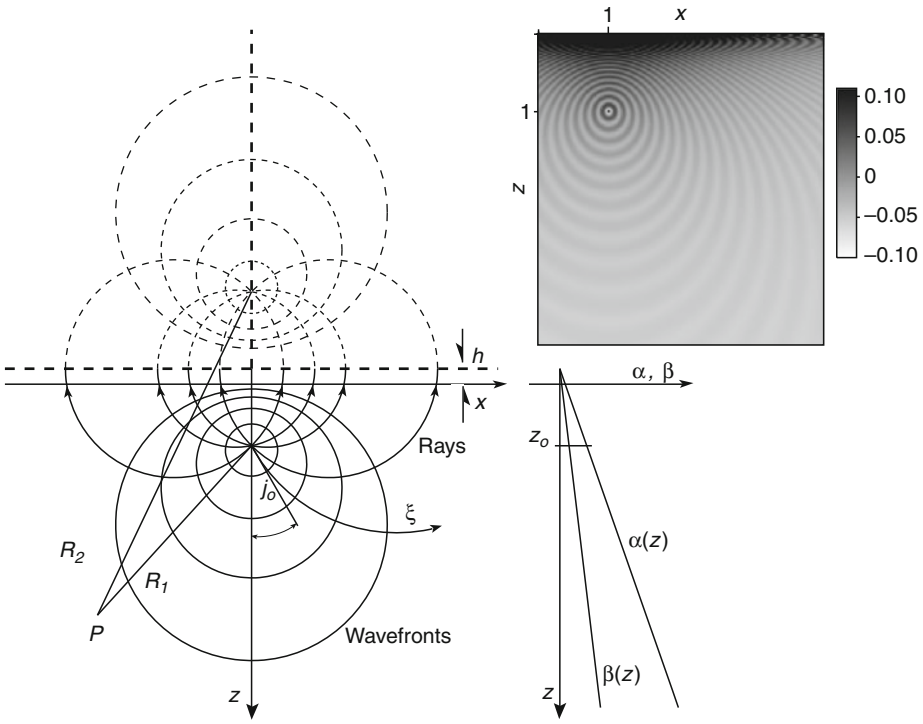
Since analytic solutions help to assess numerical methods, some attempts to develop analytic solutions for particular cases of mildly anisotropy media are briefly described. In Sánchez-Sesma et al. (2001) authors developed an approximate analytical formula for elastic 2-D Green's functions for a constant-gradient propagation velocity medium. These solutions correspond to unit line forces per unit length: the antiplane SH line source and the in-plane P-SV line sources, respectively. They are based on the asymptotic ray theory and account for both near-source effects and low frequencies. The orthogonal bipolar coordinates are depicted in Figure 4.

The approximate solution for the 2-D scalar case is given by

$$G_{22}(\omega,\tau) \approx \Lambda \frac{1}{4\mu_0} H_0^{(1)}(\omega\tau) \tag{13}$$

where Λ is defined as follows:

$$\Lambda = \left(\frac{1+\gamma z_0}{1+\gamma z}\right)^{\frac{1+n}{2}} \left[2\ln\left(\frac{R_2+R_1}{R_2-R_1}\right)\frac{(z_0+h)(z+h)}{R_1 R_2}\right]^{1/2} \tag{14}$$



Seismic Wave Propagation in Real Media: Numerical Modeling Approaches, Figure 4 (*Left*). Schematic illustration of the bipolar coordinate system. Both rays and wave fronts are circular. Rays depend upon take-off angle j_0. $t = \xi t_0$, where $t_0 = h/\beta(0)$ and ξ is one of the bipolar coordinates (the other is j_0). The distances R_1 and R_2 from the poles to point P are displayed. The plane for which $z = -h$ corresponds to null propagation velocities. (*Top right*) This solution is used to compute the Green's function for the scalar case in the time domain with a source located at (1,1), $\beta(0)$ = 500 m/s, and $\beta(z_{max})$ = 3,000 m/s. Wave fronts depend on traveltime.

where z_0 = depth of the source, β_0 = shear wave velocity at the source level, $\beta(0)$ = shear wave velocity at $z = 0$, ρ_0 = mass density at source level, $\gamma = 1/h$, h = distance from depth $z = 0$ to the level where the propagation velocity is null ($z = -h$), $j_0 = \sin^{-1}[2x(z_0 + h)/R_1R_2]$ = take-off angle, $R_1 = (x^2 + (z-z_0))^{1/2}$, and $R_2 = (x^2 + (z + z_0 + 2h)^2)^{1/2}$. The traveltime t is given by

$$\tau = t_0 \ln\left(\frac{R_2 + R_1}{R_2 - R_1}\right) = 2t_0 \tanh^{-1}\left(\frac{R_1}{R_2}\right) = t_0\xi \qquad (15)$$

In Watanabe & Payton (2002, 2006) Green's functions are obtained exactly for radially symmetric waves and for transient and time-harmonic SH-waves in inhomogeneous anisotropic elastic solids, respectively. Since time-harmonic conditions have been assumed, viscoelastic material behavior can be captured through the introduction of complex-valued wave speeds.

Equivalent media theories

When one wishes to model highly heterogeneous media, one possibility is to make the continuum hypothesis (see Christensen, 1979; Hudson, 1991). A real material with its atomic and molecular structure is replaced by an equivalent continuum model, which remains a continuum no matter how much it is subdivided. The macroscopic mechanical properties of the material remain unchanged. This hypothesis works as long as the scale of observation is much larger than the scale-length of the molecular structure (Rayleigh scattering regime). Micro fractures, pores, and other heterogeneities with uniform statistical distributions can be replaced by an equivalent or effective medium provided that the scale of the observation is much larger than the scale of the heterogeneities. This is an attractive approach in exploration seismology since the scale of the observation, or seismic wavelengths, is much larger than the scale length of the heterogeneities present in the subsurface. The concept of equivalent media clearly implies that, if the stress, strain or displacements are measured on a large enough scale, the values obtained and the relationship between them will be that of homogeneous continuum. This implies some kind of spatial averaging process. Nevertheless, scales much smaller than the minimum wavelength (used in the numerical model) are present in the earth model. Therefore effective media theories allow to average small scales of the original medium without losing the accuracy of the wavefield computation that are needed. In Capdeville and Marigo (2007), authors applied two-scale homogenization technique to the wave equation in layered media. The order 0 homogenization provides the same result as the one given by Backus (1962). The order 0 is not enough to obtain an accurate solution especially for surface waves. Higher order homogenization terms (up to 2) allowed them to obtain accurate surface waves. Lately Capdeville et al. (2010) proposed original nonperiodic homogeneization technique in 1-D. The extensions to 2-D and 3-D are foreseen.

Asymptotic ray tracing methods

An accurate estimation of traveltime is needed to map heterogeneities in the Earth's subsurface. Traveltime computation schemes fall within one of these two categories: First-arrival traveltime, and ray-theory based traveltime. The first-arrival traveltime corresponds to the first arrival of the complete wavefield at a specified receiver position and it is not a function of the type of wave (e.g., head wave, direct wave). Most of the schemes, which estimate first-arrival traveltimes, are based on the solution of eikonal equation. Vidale (1988, 1990) proposed different versions of finite-difference solution of the eikonal equation along expanding square (in 2-D) and along an expanding cube (in 3-D) to compute first-arrival traveltimes in isotropic media. Schneider et al. (1992) proposed a method based on Fermat's principle for traveltime computation in isotropic media that uses a local ray-trace solution of the eikonal equation. Ray tracing can be classified into two categories: (1) Initial-value ray tracing, and (2) Boundary value ray tracing. In initial-value ray tracing, a fan of rays is shot from the source in the model and some kind of extrapolation scheme is used to estimate the traveltime at a particular point in the model. Paraxial extrapolation (Cêrvený, 1985) is one of the most popular methods of traveltime extrapolation. An excellent review on ray theory can be found in (Cêrvený, 2001).

Poroelastic media

In addition to complex geometries, different rheologies and anisotropy, current numerical techniques are concentrating efforts to include porosity and fluid saturation in their models. Reservoir-rocks can be modeled as a porous solid matrix where the pore space is fully saturated with a pore fluid. The analysis of wave propagation in fluid saturated porous media is different in nature than that of elastic wave propagation. Biot's model is generally used to describe this effect (Biot 1956; Biot 1962). Two approaches exist to derive the poroelastic equations of motion and the constitutive laws. One method is to use homogenization methods. The laws to describe the medium are first derived on the microscopic scale of porosity. Then they are passed to macroscopic scale by homogenization. Pride et al. (1992) used an averaging procedure to derive the poroelastic equations similar to Biot (1956) and Biot (1962). We refer the reader to the book by Carcione (2007).

Boundaries

When modeling the propagation of elastic waves in any medium, the waves that reach the neighborhood of the boundaries will create reflected waves. This a nonphysical phenomena that we want to avoid. On top of the model we would like to impose a free surface. Therefore, for the top boundary, we need a reflected boundary and for the other three boundaries (in 2-D) we need a nonreflecting absorbing boundaries. In most of the numerical schemes described here, there are boundary

conditions to be imposed or it is necessary to implement absorbing boundaries conditions to eliminate spurious reflections due to the finite computational domain. The free-surface boundary condition is implemented by adding fictitious line of grid nodes next to the top boundary and use the one-side differences to approximate normal derivatives and centered difference to approximate tangencial derivatives, see Kelly et al. (1976) and Graves (1996). The simplest and more effective absorbing boundary conditions consist of a sponge along the artificially truncated edges of the domain. Clayton and Enquist (1977) proposed a kind of absorbing boundary conditions based on replacing the wave equation in the boundary region by one-way wave equation that do not allow energy to propagate from the boundaries to the numerical domain. An alternative scheme suggested by Cerjan et al. (1985) is based on gradual reduction of the amplitudes in a strip of nodes along the boundaries for the mesh. Bérenger (1994, 1996) first introduced Perfectly Matched Layers (PMLs) for electromagnetism. The main idea is to define a selective attenuation of the fields propagating in one prescribed direction. PMLs for elastic wave propagation were developed by Basu and Chopra (2003, 2004), Komatitsch and Martin (2007). The velocity-stress formulation of PMLs for elastic wave equations has been introduced by Collino and Tsogka (2001). It was applied to viscoelastic (Martin & Komatitsch, 2009) or poroelastic (Zeng et al., 2001; Martin et al., 2008) media. Recently Meza-Fajardo and Papageorgiou (2008) demonstrated that for any anisotropic media, the classic PML exhibits instabilities. In addition, they presented a generalization of the classical Perfectly Matched Layer (PML) (called multiaxial Perfectly Matched Layer M-PML) to a medium in which damping profiles are specified in more than one direction.

Summary

Modeling the Earth is a challenging task. With the increase of computational power, direct modeling techniques start to open the door to inversion problems. All numerical methods share at the same time their own disadvantages and weaknesses (high computational costs and storage requirements, poor accuracy, nonstructured grids, hard implementations, system matrix full) with the agreeable flavor of equivalent advantages (reduction of one dimension of the problem, system matrix diagonal or sparse, no need for absorbing boundaries, easy implementation, easy to model topographies or complex geometries). The nature of the problem gives us the appropriate guide to choose one method over others. In this article we briefly cover some numerical techniques to model elastic waves in heterogeneous media. However, the literature on all available methods is very extensive and we include only some references.

Acknowledgments

This work was partially supported by DGAPA-UNAM Project IN121709, Mexico.

Bibliography

Aki, K., and Richards, P. G., 1980. *Quantitative Seismology*. New York: WH Freeman.

Backus, G., 1962. Long-wave elastic anisotropy produced by horizontal layering. *Journal of Geophysical Research*, **67**(11), 4427–4440.

Basu, U., and Chopra, A. K., 2003. Perfectly matched layers for time-harmonic elastodynamics of unbounded domains: theory and finite-element implementation. *Computer Methods in Applied Mechanics and Engineering*, **192**, 1337–1375.

Basu, U., and Chopra, A. K., 2004. Perfectly matched layers for transient elastodynamics of unbounded domains. *International Journal for Numerical Methods in Engineering*, **59**, 1039–1074.

Bayliss, A., Jordan, K. E., LeMesurier, B. J., and Turkel, E., 1986. A fourth-order accurate finite-difference scheme for the computation of elastic waves. *Bulletin of the Seismological Society Of America*, **76**, 1115–1132.

Bérenger, J. P., 1994. A perfectly matched layer for the absorption of electromagnetic waves. *Journal of Computational Physics*, **114**, 185–200.

Bérenger, J. P., 1996. Three-dimensional perfectly matched layer for the absorption of electromagnetic waves. *Journal of Computational Physics*, **127**, 363–379.

Bielak, J., Ghattas, O., and Kim, E. J., 2005. Parallel octree-based finite element method for large-scale earthquake ground motion simulation. *Computer modeling in engineering and sciences*, **10**, 99–112.

Biot, M., and Biot, M., 1956. Theory of propagation of elastic waves in a fluid saturated porous solid. i-low frequency range. *The Journal of the Acoustical Society of America*, **28**, 168–178.

Biot, M., 1962. Mechanics of deformation and acoustic propagation in porous solids. *Journal of Applied Physics*, **33**, 1482–1498.

Bonnet, M., 1999. *Boundary Integral Equation Method for Solids and Fluids*. Chichester: Wiley.

Bouchon, M., and Sánchez-Sesma, F. J., 2007. Boundary integral equations and boundary element methods in elastodynamics. Advances in wave propagation in heterogeneous earth. In Wu, R.-S., Maupin, V., and Dmowska, R. (eds.), *Advances in Geophysics*. New York: Elsevier-Academic, Vol. 48, pp. 157–190.

Capdeville, Y., Guillot, L., and Marigo, J.-J., 2010. 1-D non periodic homogenization for the wave equation. *Geophysical Journal International*, **181**, 897–910.

Capdeville, Y., and Marigo, J.-J., 2007. Second order homogenization of the elastic wave equation for non-periodic layered media. *Geophysical Journal International*, **170**, 823–838.

Carcione, J. M. D., Kosloff, A. B., and Seriani, G., 1992. A spectral scheme for wave propagation simulation in 3-D elastic-anisotropic media. *Geophysics*, **57**(12), 1593–1607.

Carcione, J. M., 1994. The wave equation in generalized coordinates. *Geophysics*, **59**, 1911–1919.

Carcione, J. M., Herman, G. C., and ten Kroode, A. P. E., 2002. Seismic modeling. *Geophysics*, **67**(4), 1304–1325.

Carcione, J. M., 2007. *Wave Field in Real Media. Wave Propagation in Anisotropic, Anelastic, Propus and Electromagnetic Media*. Netherlands: Elsevier. Handbook of Geophysical Exploration.

Cerjan, C., Kosloff, R., and Reshef, M., 1985. A nonreflecting boundary condition for discrete acoustic and elastic wave equations. *Geophysics*, **50**, 705–708.

Cêrvený, V., 2001. *Seismic Ray Theory*. Cambridge: Cambridge University Press.

Cêrvený, V.,1985. The application of ray tracing to the numerical modeling of seismic wavefields in complex structures. In Helbig, K., and Treitel, S., *Handbook of Geophysical Exploration*. London: Geophysical Press, Vol. 15A, pp 1–124.

Chaillat, S., Bonnet, M., and Semblat, J. F., 2009. A new fast multi-domain BEM to model seismic wave propagation and

amplification in 3D geological structures. *Geophysical Journal International*, **177**, 509–531.
Claerbout, J. F., 1968. Synthesis of a layered medium from its acoustic transmission response. *Geophysics*, **33**(2), 264–269.
Claerbout, 1985. *Imagining the Earth's Interior*. Blackwell Scientific Publication.
Clayton, R., and Enquist, B., 1977. Absorbing boundary conditions for acoustic and elastic wave equations. *Bulletin of the Seismological Society of America*, **67**(6), 1529–1540.
Christensen, R. M., 1979. *Mechanics of Composite Materials*. New York: Wiley.
Cockburn B., Karniadakis, G., and Shu C., 2000. Discontinuous Galerkin methods: theory, computation and applications, Vol. 11 of Lecture Notes in Computational Science and Engineering. Springer.
Collino, F., and Tsogka, C., 2001. Application of the perfectly matched absorbing layer model to the linear elastodynamic problem in anisotropic heterogeneous media. *Geophysics*, **66**, 294–307.
Cruz-Atienza, V. M., Virieux, J., and Aochi, H., 2007. 3D Finite-difference dynamic-rupture modelling along non-planar faults. *Geophysics*, **72**, SM123–SM137.
Emerman, H. S., Schmidt, W., and Stephen, F. L. A., 1982. An implicit finite-difference formulation of the elastic wave equation. *Geophysics*, **47**(11), 1521–1526.
Frankel, A., and Leith, W., 1992. Evaluation of topographic effects on it P and it S waves of explosions at the northern Novaya Zemlya test site using 3-D numerical simulations. *Geophysical Research Letters*, **19**, 1887–1890.
Dominguez, J., 1993. *Boundary Elements in Dynamics*. London/ New York: Elsevier Applied Science.
Frankel, A., and Vidale, J., 1992. A three-dimensional simulation of seismic waves in the Santa Clara Valley, California, from a Loma Prieta Aftershock. *Bulletin of the Seismological Society Of America*, **82**, 2045–2074.
Fujiwara, H., 1998. The fast multipole method for the integral equations of seismic scattering problems. *Geophysical Journal International*, **133**, 773–782.
Fujiwara, H., 2000. The fastmultipolemethod for solving integral equations of three-dimensional topography and basin problems. *Geophysical Journal International*, **140**, 198–210.
Furumura, T., and Kennett, B. L. N., 1997. On the nature of regional phases – II On the effect of crustal barriers. *Geophysical Journal International*, **129**, 221–234.
Furumura, T., and Chen, L., 2005. Parallel simulation of strong ground motions during recent and historical damaging earthquakes in Tokyo, Japan. *Parallel Computing*, **31**(2), 149–165.
Furumura, T., Kennett, B. L. N., and Takenaka, H., 1998. Parallel 3-D pseudospectral simulation of seismic wave propagation. *Geophysics*, **63**, 279–288.
Graves, R. W., 1996. Simulating seismic wave propagation in 3D elastic media using staggered-grid finite differences. *Bulletin of the Seismological Society Of America*, **86**, 1091–1106.
Hesthaven, J. S., and Warburton, T., 2008. *Nodal Discontinous Galerkin Methods Texts in Applied Mathematics*. New York: Springer Verlag.
Higdon, R. L., 1991. Absorbing boundary-conditions for elastic-waves. *Geophysics*, **56**, 231–241.
Hudson, J. A., 1991. Overall properties of heterogeneous material. *Geophysical Journal International*, **107**, 505–511.
Iturrarán-Viveros, U., Vai, R., and Sánchez-Sesma, F. J., 2005. Scattering of elastic waves by a 2-D crack using the Indirect Boundary Element Method (IBEM). *Geophysical Journal International*, **162**, 927–934.
Kausel, E., 2006. *Fundamental Solutions in Elastodynamics. A Compendium*. New York: Cambridge University Press.
Käser, M., and Dumbster, M., 2006. An arbitrary high-order discontinuous Galerkin method for elastic waves on unstructured meshes – I. The two-dimensional isotropic case with external source terms. *Geophysical Journal International*, **166**, 855–877.
Kelly, K. R., Ward, R., Treitel, W., and Alford, R. M., 1976. Synthetic seismograms: A finite difference approach. *Geophysics*, **41**, 2–27.
Komatitsch, D., and Vilotte, J.-P., 1998. The Spectral Element method: an efficient tool to simulate the seismic response of 2D and 3D geological structures. *Bulletin of the Seismological Society Of America*, **88**, 368–392.
Komatitsch, D., and Tromp, J., 1999. Introduction to the spectral-element method for 3-D seismic wave propagation. *Geophysical Journal International*, **139**, 806–822.
Komatitsch, D., and Martin, R., 2007. An unsplit convolutional perfectly matched layer improved at grazing incidence for the seismic wave equation. *Geophysics*, **72**, SM155–SM167.
Kosloff, D., and Baysal, E., 1982. Forward modeling by the Fourier method. *Geophysics*, **47**, 1402–1412.
Kreiss, H.-O., and Oliger, J., 1972. Comparison of accurate methods for the integration of hyperbolic equations. *Tellus*, **24**, 199–215.
Liu, Y., and Sen, M. K., 2009. A practical implicit finite-difference method: examples from seismic modelling. *Journal of Geophysics and Engineering*, **6**, 31–249.
Marfurt, K. J., 1984. Accuracy of finite-difference and finite element modeling of the scalar and elastic wave equations. *Geophysics*, **49**, 533–549.
Martin, R., Komatitsch, D., and Ezziani, A., 2008. An unsplit convolutional perfectly matched layer improved at grazing incidence for seismic wave equation in poroelastic media. *Geophysics*, **73**(4), T51–T61.
Martin, R., and Komatitsch, D., 2009. An unsplit convolutional perfectly matched layer technique improved at grazing incidence for the viscoelastic wave equation. *Geophysical Journal International*, **179**(1), 333–344.
Manolis, G. D., and Beskos, D. E., 1988. *Boundary Element Methods in Elastodynamics*. London: Unwin Hyman.
Manolis, G. D., and Shaw, R. P., 1996. Green's function for the vector wave equation in midly heterogeneous continum. *Wave Motion*, **24**, 59–83.
Meza-Fajardo, K., and Papgeorgiou, A. S., 2008. A nonconvolutional, split-field, perfectly matched layer for wave propagation in isotropic and anisotropic elastic media: Stability analysis. *Bulletin of the Seismological Society Of America*, **98**(4), 1811–1836.
Moczo, P., Robertsson, J. O. A., and Eisner, L., 2007. The finite-difference time-domain method for modeling of seismic wave propagation. *Advances in Geophysics*, **48**, 421–516.
Mufti, I. R., 1985. Seismic modeling in the implicit mode. *Geophysical Prospecting*, **33**, 619–656.
Olsen, K. B., Archuleta, R. J., and Matarese, J. R., 1995. Magnitude7.75 earthquake on the San Andreas fault: three-dimensional ground motion in Los Angeles. *Science*, **270**, 1628–1632.
Payton, R. G., 1983. *Elastic Wave Propagation in Transversely Isotropic Media*. Hague: Mertinus Nijhoff.
Pointer, D., Liu, E., and Hudson, J., 1998. Numerical modeling of seismic waves generated by hydrofractures: application of the indirect boundary element method. *Geophysical Journal International*, **135**, 289–303.
Pride, S., Gangi, A., and Morgan, F., 1992. Deriving the equations of motion for porous isotropic media. *The Journal of the Acoustical Society of America*, **92**(6), 3278–3290.
Priolo, E., Carcione, J. M., and Seriani, G., 1994. Numerical simulation of interface waves by high-order spectral modeling techniques. *The Journal of the Acoustical Society of America*, **95**, 681–693.

Rokhlin, V., 1990. Rapid solution of integral equations of scattering theory in two dimensions. *J. Comp. Phys.*, **86**, 414–439.
Richtmyer, R. D., and Morton, K. W., 1967. *Difference Methods for Initial-Value Problems*, 2nd edn. Nuw York: Wiley.
Thomsen, L., 1986. Weak elastic anisotropy. *Geophysics*, **51**, 1954–1966.
Saenger, E. H., Gold, N., and Shapiro, S. A., 2000. Modeling the propagation of elastic waves using a modified finite-difference grid. *Wave Motion*, **31**, 77–92.
Saenger, E. H., and Bohlen, T., 2004. Finite-difference modeling of viscoelastic and anisotropic wave propagation using the rotated staggered grid. *Geophysics*, **69**, 583–591.
Saito, T., and Furumura, T., 2009. Three-dimensional tsunami generation simulation due to sea-bottom deformation and its interpretation based on the linear theory. *Geophysical Journal International*, **178**, 877–888.
Sánchez-Sesma, F. J., Madariaga, R., and Irikura, K., 2001. An approximate elastic two-dimensional Green's function for a constant-gradient medium. *Geophysical Journal International*, **146**, 237–248.
Schneider, W. A., Ranzinger, K. A., Balch, A. H., and Kruse, C., 1992. A dynamic programming approach to first arrival traveltime computation in media with arbitrary distributed velocities. *Geophysics*, **57**, 39–50.
Somigliana, C., 1886. Sopra l'equilibrio di un corpo elastico isotrope. *Nuovo Cimento*, **20**, 181–185.
Wu, R-S., Maupin V., and Dmowska, R., Seriani G., Priolo, E., Carcione, J. M., and Padovani, E., 1992. High-order spectral element method for elastic wave modelling. Advances in Wave Propagation in Heterogeneous Earth. SEG 62 Annual meeting, Expanded Abstracts, pp. 1285–1288.
Vidale, J. E., 1988. Finite-difference calculation of travel times. *Bulletin of the Seismological Society Of America*, **78**, 2062–2076.
Vidale, J. E., 1990. Finite-difference calculation of traveltimes in three dimensions. *Geophysics*, **55**, 521–526.
Virieux, J., 1986. P-SV wave propagation in heterogeneous media: Velocity-stress finite-difference method. *Geophysics*, **51**, 889–901.
Virieux, J., 1984. SH-wave propagation in heterogeneous media: Velocity-stress finite-difference method. *Geophysics*, **49**, 1933–1957.
Watanabe, K., and Payton, R., 2002. Green's function for SH-waves in a cylindrical monoclinic material. *Journal of the Mechanics and Physics of Solids*, **50**, 2425–2439.
Watanabe, K., and Payton, R., 2006. Green's function for radially symmetric waves in inhomogeneous anisotropic solids. *Zeitschrift fur Angewandte Mathematik und Mechanik*, **86**, 642–648.
Yoshida, K. I., 2001. *Application of Fast Multipole Method to Boundary Integral Equation Method*. Ph. D thesis, Kyoto University.
Zeng, Y. Q., He, J. Q., and Liu, Q. H., 2001. The application of the perfectly matched layer in numerical modeling of wave propagation in poroelastic media. *Geophysics*, **66**(4), 1258–1266.

Cross-references

Body Waves
Earthquake Rupture: Inverse Problem
Earthquakes, Source Theory
Poroelasticity
Propagation of Elastic Waves: Fundamentals
Seismogram Interpretation
Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting
Statistical Seismology
Surface Waves
Tsunami

SEISMIC WAVES, SCATTERING

Ludovic Margerin
Institut de Recherche en Astrophysique et Planétologie, Observatoire Midi-Pyrénées/C.N.R.S., Toulouse, France

Definition

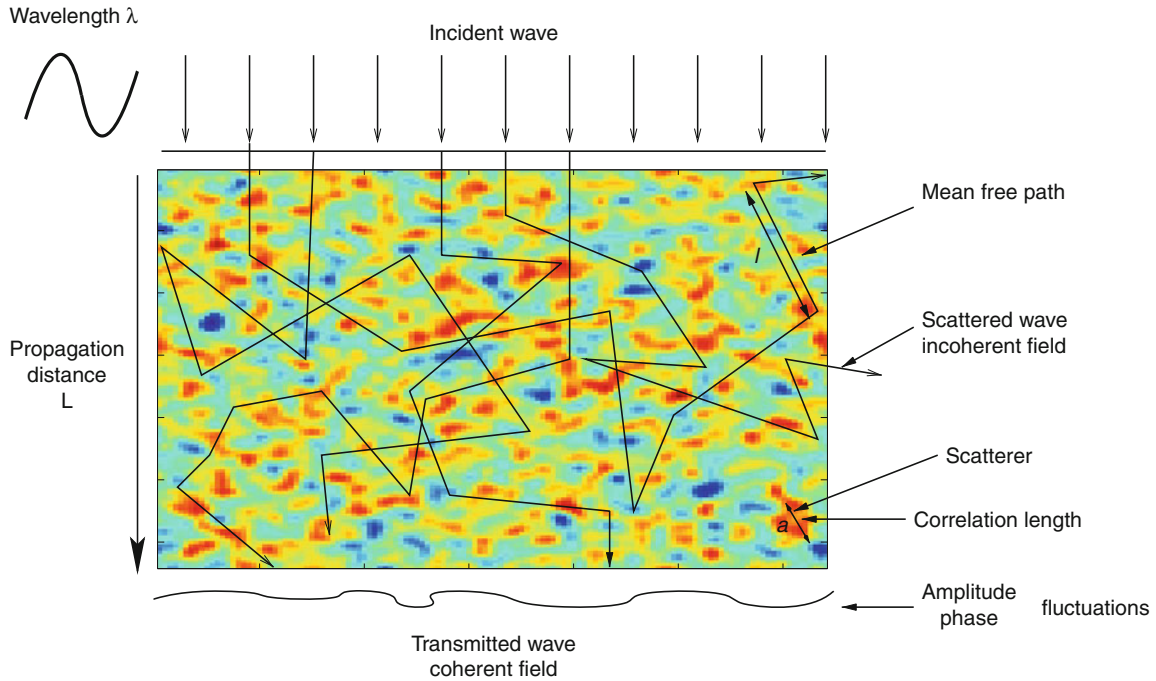
Scattering of seismic waves is best defined by reference to a laterally homogeneous or slowly varying medium where the wave fronts can be perfectly tracked and the propagation of waves can be successfully described by geometrical methods such as ray theory. In the presence of obstacles, or lateral variations of elastic parameters, wave fronts are distorted and seismic energy can be deflected in all possible directions: a phenomenon known as wave scattering. Specific to seismic waves are the possible mode conversions between various polarizations: compressional, shear, and Rayleigh waves.

Introduction

The scattering of seismic waves is intimately related to the heterogeneous nature of Earth materials on a variety of spatial scales. Well-logs data or geological maps offer direct but limited access to the nature of heterogeneities in the Earth. As elastic waves can propagate all the way through the Earth, they constitute the primary tool to sound deep heterogeneities. While the pioneering studies of seismic scattering were largely focused on the lithosphere, the presence of scatterers in the mantle and solid core has been confirmed in a number of recent studies (see Shearer and Earle, 2008, for a review). Hence, the field of seismic scattering has become of interest in global seismology too. For a thorough review of pre-1990s works in seismic scattering, we refer the reader to Wu (1989). A comprehensive reference on the topic is the book by Sato and Fehler (1998). Reviews of recent methodological and observational developments are given in Sato et al. (2008).

Statistical description of the Earth

To introduce the various scale lengths that enter in the modeling of seismic scattering, it is convenient to consider a thought-experiment such as the one shown in Figure 1. We imagine an incident plane wave propagating through a heterogeneous medium. As evidenced by the analysis of well-logs (see for instance Wu et al., 1994), the density and the elastic constants fluctuate in the Earth in a very complicated way. In such a situation, it is convenient to think of the Earth as a particular realization of an ensemble of random media sharing similar statistical properties. The goal of the stochastic imaging methods is to infer some information on the nature of the randomness from the statistical fluctuations of the wavefield. This is a difficult task because the statistical properties of the Earth are nonhomogeneous: they are expected to depend on the geological setting and on the depth. In addition, in the case



Seismic Waves, Scattering, Figure 1 Basic concepts of seismic scattering. We consider a thought-experiment where an incoming plane wave propagates through a chunk of random medium with typical dimension *L*. As the plane wave progresses through the heterogeneous medium, scattering operates and deflects energy in all space directions. The transmitted wavefield is distorted and attenuated.

of the Earth, the random fluctuations are superimposed on a deterministic background, usually represented by a 1-D reference model such as PREM (Dziewonski and Anderson, 1981). It is customary to assume that the spatial variations of the background velocities are slow compared to the typical scale length of the random medium. In this limit of separation of spatial scales, it is possible to consider the scattering medium as locally homogeneous. We may therefore hope to characterize large-scale variations of small-scale heterogeneities.

If we think of the medium as random and continuous, the fluctuations of a parameter such as the density ρ are characterized to lowest order by a correlation function (Rytov et al., 1989b):

$$\psi(\mathbf{x}) = \langle \rho(\mathbf{r} - \mathbf{x}/2)\rho(\mathbf{r} + \mathbf{x}/2)\rangle, \tag{1}$$

where the brackets denote an average over the ensemble of random media. In Equation 1, we have used an assumption of statistical homogeneity. In the case of a locally homogeneous medium, the correlation function may be allowed to vary slowly with the variable $\mathbf{r}$. $\psi(\boldsymbol{x})$ is maximum for $x = 0$, where it is equal to the total variance of the fluctuations denoted by ϵ^2 and goes to zero over a typical spatial scale a which is known as the correlation length. Common choices for ψ are Gaussian and exponential functions but a wide range of other possibilities exist (Klimeš, 2002). An equivalent description of the medium fluctuations is provided by the power spectrum of heterogeneities $\Phi(\mathbf{k})$ defined as:

$$\psi(\mathbf{x}) = \frac{1}{(2\pi)^3} \iiint_{-\infty}^{+\infty} \Phi(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{x}} d^3k \tag{2}$$

$\Phi(\mathbf{k})$ quantifies the distribution of the total variance of the fluctuations over all possible length scales. A very useful tool to represent the heterogeneity of fluid or solid geophysical media is the Von-Karman correlation function, whose power spectrum can be written as (Sato and Fehler, 1998):

$$\Phi(\mathbf{k}) = \frac{8\pi^{3/2}\varepsilon^2 a^3 \Gamma\left(\kappa + \frac{3}{2}\right)}{\Gamma(\kappa)(1 + k^2a^2)^{\kappa+3/2}}. \tag{3}$$

In Equation 3, ϵ denotes the RMS fluctuations, a is the correlation length and $\kappa > 0$ is the Hurst exponent which acts as a cutoff of the small-scale features in the medium. The power spectrum (Equation 3) shows a plateau at small wave numbers up to a corner value $k_c \sim 1/a$. For k much larger than k_c, the power spectrum decreases algebraically as $k^{2\kappa+3}$. Note, however, that this description does not apply to Gaussian random media. This is not a severe restriction, since Gaussian media are often regarded as too smooth to represent faithfully the heterogeneities in the Earth. This point of view is confirmed by the analysis

of well-logs (Shiomi et al., 1997), which further shows that only the parameter κ can usually be recovered from observations.

The coherent and incoherent fields

Following the statistical approach, it is natural to characterize the propagation in the random medium by calculating the moments of the wavefield u. This yields a simple classification of the various components of the random field. The first moment $\langle u \rangle$ defines the coherent wave (Rytov et al., 1989a; Sheng, 1995) and corresponds to the part of the field which survives after an average over the statistical ensemble has been performed. In a typical seismological experiment, the measured field is the sum of two terms: the ensemble average $\langle u \rangle$ and a fluctuating part u_f. The coherent field mostly contributes in a time window around the seismic primaries, which may also be termed the ballistic waves. Although the coherent field is an important component of the ballistic field, it has to be carefully distinguished from it (Derode et al., 2001). Statistical wave theory shows that the mean field decays like $e^{-L/2l}$, where L is the propagation distance and l is known as the mean free path (Sheng, 1995). The mean free path gives the typical spatial scale over which a significant amount of energy has been transferred from the coherent to the incoherent or fluctuating part of the wavefield. The incoherent part of the wavefield contributes to all portions of the seismogram and forms the coda of the earthquake records. It is important to realize that the term "incoherent" just indicates the generation of waves that propagate in a direction which differs from that of the incident wave. It does not imply that the phase information has been lost. In section Interferometry with scattered waves, we will demonstrate that the so-called incoherent field is in fact highly "coherent" in the sense that it gives rise to very useful interference effects.

The propagation regimes

The distinction between the fluctuating and ensemble average part of the wavefield is useful to characterize the propagation regime. For propagation distances typically less than the mean free path, the measured field is dominated by the coherent wave: this is the weak fluctuation regime (Rytov et al., 1989a). As shown in Figure 1, after propagating through several inhomogeneities, the most important observation is the distortion of the incident plane wave front. This gives rise to fluctuations of both phase and amplitude, from which information on the statistical properties of the medium can be extracted. For propagation distances much larger than the mean free path, the coherent wave completely vanishes and one is left with the fluctuating part only: this is the strong fluctuation regime (Rytov et al., 1989a). Since by definition the ensemble average of the fluctuating part u_f is zero, it is only through the consideration of the second moment of the field that one can model the propagation of scattering when fluctuations are strong. Specific approaches will be developed in sections Envelope modeling: Markov approximation and Envelope modeling: radiative transfer.

At this stage, it is natural to raise the following question: How can we measure the mean free path in the Earth? First, it is important to note that one has to consider at least two different mean free paths since both longitudinal and transverse waves propagate through an elastic medium. In the laboratory, there are well-calibrated techniques to measure the mean free path based on ensemble averaging. Unfortunately, these methods cannot be transposed to field experiments. Instead, the common seismological practice consists in correcting the seismograms for the travel-time fluctuations caused by long wavelength variations of the velocity before stacking or averaging the records. The field obtained after such a processing is called the traveltime-corrected mean field. Sato (1982, 1984) and Wu (1982) have demonstrated that this field decays exponentially on a scale which is much larger than the mean free path, because the traveltime correction in fact removes the long wavelength components of the medium fluctuations. A theoretical discussion of various averaging procedures and their application to seismic data is given by Shapiro and Kneib (1993).

Analysis of transmission fluctuations

As announced in introduction, we now review the various approaches that have been developed to infer the statistical properties of the Earth from an analysis of seismic wave scattering. We will first introduce a method which is well adapted to the analysis of fluctuations of amplitude and phase of teleseismic P waves recorded on a dense array. Such an approach has been pioneered in the seventies by Aki (1973) and has been subsequently developed by Wu and coworkers (Wu and Flatté, 1990; Zheng et al., 2007; Zheng and Wu, 2008). In this approach, one considers a small time window around the direct P-wave at teleseismic distance. It is usually assumed that at sufficiently large distance from the source, the wave front incident from the mantle can be approximated locally by a plane. By using sufficiently deep earthquakes, it is possible to avoid the effects of strong heterogeneities in the vicinity of the source. As the plane wave progresses through the lithosphere, the wave front is distorted and the amplitude and phase of the ballistic waves measured on the array fluctuate. Zheng and Wu (2005) have very carefully discussed the correct measurement practice. The main pulse recorded on the vertical component of the seismometer is first isolated from the rest of the signal. After Fourier transformation, the amplitude A and the phase ϕ of the pulse is obtained at a given frequency. This operation is repeated at all stations of the array, and the phase field is subsequently unwrapped spatially, i.e., possible phase discontinuities are removed. It is not correct to measure the phase of the field by picking the first arrival since, as a consequence of scattering, the medium is dispersive and the first arrivals propagate with the group velocity.

Basic observables are the fluctuations of the phase and of the logarithm of the amplitude. The reason why one considers the logarithm of the amplitude is that this quantity enters in the Rytov representation of the wavefield Rytov et al. (1989a):

$$u = u_0 e^{\psi}, \tag{4}$$

where u_0 is the reference plane wavefield in the background medium and ψ is a complex phase:

$$\mathrm{Re}\psi = \ln\frac{A}{A_0} \quad \mathrm{Im}\psi = \phi - \phi_0 \tag{5}$$

Scattering theory based on the Rytov representation is more accurate than the standard Born approximation in the transmission geometry. In Equation 5, A (resp. A_0) and ϕ (resp. ϕ_0) denote the amplitude and unwrapped phase of the field (resp. reference field). From the array measurement, one can estimate the transverse coherence function of the phase field which is defined as:

$$C_\phi(\boldsymbol{r}) = \langle \phi(\boldsymbol{x})\phi(\boldsymbol{r}+\boldsymbol{x})\rangle, \tag{6}$$

where the points $\boldsymbol{x}$ and $\boldsymbol{x} + \boldsymbol{r}$ lie on the surface of the Earth. More generally, one can consider the transverse angular coherence function of the two phase fields corresponding to two incoming plane waves with different incident wavevectors $\boldsymbol{k}$ and $\boldsymbol{k}'$ (Wu and Flatté, 1990; Chen and Aki, 1991).

The interpretation of the phase coherence function is based on two major theoretical approaches, with different domains of validity. These approaches rely on the basic physical assumption that the medium fluctuations have a spatial scale, a, which is much larger than the probing wavelength, λ. In this regime, the scattering occurs preferentially in a small angular cone around the forward direction. As a consequence, one can usually neglect the backscattered waves and consider one-way propagation only. The crucial nondimensional parameter that governs the properties of the phase is known as the wave parameter, D, which is defined as (Rytov et al., 1989a):

$$D = \frac{\lambda L}{a^2} \tag{7}$$

In the numerator of Equation 7, the length $\sqrt{\lambda L}$ provides the typical size of the first Fresnel zone which can be interpreted as the typical volume sampled by the waves around a ray of length L. The region $D \ll 1$ can be conveniently treated with the methods of geometrical optics. In this regime, there are some well-established results for the phase coherence function of a single plane wave with vertical incidence: (1) The variance of the phase is proportional to the total variance of the medium fluctuations and increases linearly with the propagation distance L. (2) The typical correlation length of the phase is proportional to the correlation length of the random medium. This result makes sense physically, since parts of the incident wave front which are located a distance a apart – a, the correlation length– will visit different inhomogeneities. This provides a very useful method to map the typical scale of heterogeneities. The geometrical optics approach is rather flexible and allows the treatment of more complex situations. Kravtsov et al. (2003) have adapted the method to the reflexion geometry usually encountered in exploration geophysics.

The method of geometrical optics usually breaks down for $D > 1$ as caustics may develop for sufficiently large propagation distances. In order to account properly for diffraction effects, it is important to use a wave-based approach. As in the problem of transmission fluctuations in which the propagation is essentially one-way, it is convenient to consider the depth variable z as an evolution variable and to neglect the backscattered waves. In this case, the usual Helmholtz equation can be replaced by the following parabolic approximation for the wavefield:

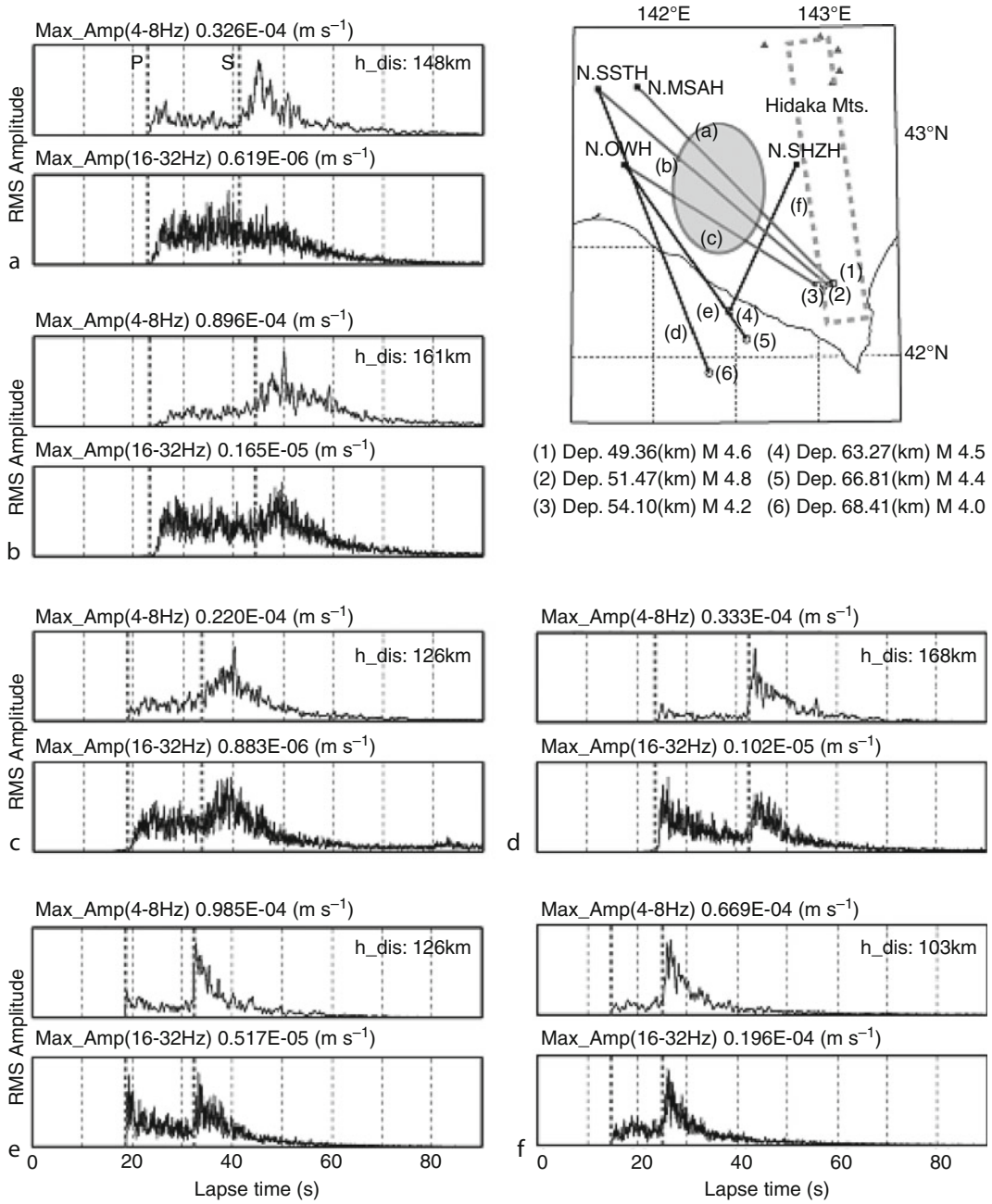
$$2ik_0\frac{\partial u}{\partial z} + \Delta_\perp u - 2k_0^2\frac{\delta c}{c_0}u = 0, \tag{8}$$

where $\Delta_\perp = \partial_x^2 + \partial_y^2$ represents the Laplace operator in the plane perpendicular to the main propagation direction. The parabolic approximation is accurate for low-angle scattering only, which is consistent with the scaling $\lambda/a \ll 1$. In this regime, a very powerful approach couples scattering theory and the Rytov representation of the wavefield to obtain expressions for the coherence function that extend the results of geometrical optics to the region $D > 1$. We refer to Wu and Flatté (1990) for further details on the derivation.

The first application to data was made by Aki who estimated the correlation length and total variance of the velocity fluctuations from the cross-coherence of amplitude and phase under the LASA array (Aki, 1973). In this pioneering work, he further approximated the correlation function of the medium with a Gaussian function. It was later realized that this choice was not always the most appropriate since many studies revealed that the more general Von-Karman correlation function could be used to describe a broad class of random media. Flatté and Wu (1988) developed a two-layer model of heterogeneity of the upper mantle based on observed transverse and angular correlation functions at NORSAR. By superimposing a layer with Hurst exponent 1/2 extending from 15 to 250 km depth upon a layer with flat power spectrum extending from 0 to 200 km depth, they were able to explain satisfactorily the joint measurements of phase and amplitude fluctuations. The most recent developments by Zheng et al. (2007) and Zheng and Wu (2008) allow the treatment of velocity gradient in the reference medium. Their general formalism offers the possibility to perform a nonparametric inversion, i.e., the full power spectrum of the fluctuations at depth can in principle be mapped from the surface measurements of angular correlation functions. For a complete description of the method as well as numerous references to the literature, we refer to the review by Zheng and Wu (2008).

Envelope modeling: Markov approximation

We now consider a regime of propagation where the fluctuating part of the field is larger than the coherent part. In this regime, an interesting phenomenon termed "envelope broadening" can be used to retrieve information on the statistical fluctuations of velocities in the medium. Enveloped broadening is illustrated in Figure 2, where small earthquakes have been recorded at epicentral distances of the order of 150 km in eastern Japan. On these data, we can observe a delay between the onset of the direct shear wave and the maximum of the amplitude. Such a delay is characteristic of the strong forward-scattering regime where the waves follow "snake-like" paths around the ballistic trajectory between source and station. As clearly shown in



Seismic Waves, Scattering, Figure 2 Example of earthquake records showing the phenomenon of envelope broadening. Paths a, b, c propagate through a very heterogeneous volcanic region. A large delay between the onset of *S* waves and the maximum amplitude is clearly observed. This phenomenon further depends on the frequency band considered. Paths d, e, f do not cross the heterogeneous region and show more impulsive envelopes with small onset to peak delay time (Reproduced from Takahashi et al. (2007)).

Figure 2, envelope broadening depends on propagation distance and frequency band. It is also noticeable that very heterogeneous regions such as volcano areas give rise to much stronger envelope broadening than the usual crust. The clear path dependence suggests the possibility to use regional variations of the onset to peak delay time to map heterogeneities in the crust, in order to develop a stochastic tomography in the strong fluctuation regime.

Such an approach has been pioneered by Sato in the late 1980s. A recent review of the topic in seismology is given by Sato and Korn (2009). Based on the parabolic approximation, Sato (1989) derived an equation for the two-point, two-frequency coherence functions of the wavefield for incident plane waves. This technique which was originally introduced in astrophysics has been termed "Markov approximation." In the case of regional propagation, it is preferable however to consider the radiation from a point source. Following Saito et al. (2002), we write the coherence function of the field as follows:

$$\Gamma(\mathbf{r}_{\perp 1},\ \mathbf{r}_{\perp 2},\ r,\ \omega_d, \omega_c) = \langle U(\mathbf{r}_{\perp 1},\ r,\ \omega_1)\langle U(\mathbf{r}_{\perp 2},\ r,\ \omega_2)^*\rangle, \tag{9}$$

where ω_c and ω_d denote the central $(\omega_1 + \omega_2)/2$ and modulation $\omega_1 - \omega_2$ frequency, respectively. The introduction of two frequencies is demanded by the nonstationarity of the envelope records. The source station distance is denoted by r and the points $\mathbf{r}_{\perp 1}$ and $\mathbf{r}_{\perp 2}$ lie in a plane perpendicular to the wave front. The time dependence of the observed coda envelopes around the frequency ω_c is obtained by setting $\mathbf{r}_{\perp 1} = \mathbf{r}_{\perp 2}$ and taking a Fourier transform over ω_d.

The basic idea of the so-called Markov approximation is to consider one-way propagation and to neglect backscattering. The different slices of the medium perpendicular to the main propagation direction independently perturb the local ray direction in a random fashion. As a consequence, the distribution of ray directions at a distance r_0 from the source depends solely on the heterogeneities that have been met for $r < r_0$ (Sato, 1989; Sato and Korn, 2009). Such an approximation is valid for strong forward-scattering, i.e., when the relation $\lambda > a$ applies. This does not mean that the Markov approximation cannot be used if the medium contains heterogeneities with scale lengths shorter than the wavelength. Indeed, the large-scale fluctuations contribute predominantly to the field measured in the transmission geometry. Scattering at large angles caused by small-scale fluctuations mainly plays the role of an apparent energy loss which adds to the intrinsic losses (Sato, 1989). In order to quantify the typical validity region of the Markov approximation, it is useful to introduce a quantity known as the transport mean free path (Sheng, 1995) and defined as:

$$l^* = \frac{l}{1 - \langle \cos\theta \rangle}, \tag{10}$$

where $\langle \cos\theta \rangle$ is the mean cosine of the scattering angle. The transport mean free path can be interpreted as the typical length beyond which the waves have lost memory of their initial propagation direction. In the case of large-scale fluctuations, the transport mean free path l^* is typically one or two orders of magnitude larger than the scattering mean free path l. For propagation distances of the order of the transport mean free path, backscattering can no longer be ignored and one must appeal to other methods to be described in section Envelope modeling: radiative transfer.

As carefully explained in the paper of Sato (1989), there are two basic phenomena that contribute to the observed envelope shape. The first one is termed "the wandering effect" and is purely statistical in nature. It corresponds to the small arrival time fluctuations in the line-of-sight propagation direction. It leads to a Gaussian spread of an initially delta-like pulse. This reshaping can only be observed upon ensemble averaging and is therefore not of primary importance in seismology. The second phenomenon is the increase of the typical duration of the seismogram with propagation distance which has been termed "pulse" – or may be more appropriately – "envelope broadening." It reflects the angular spreading of the wave propagation directions as they interact with the medium heterogeneities. For a general Von-Karman correlation function, a time scale t_m characterizes the typical duration of the seismogram (Saito et al., 2002):

$$t_m = C\left(\varepsilon^{1/(p-1)} a^{-1}\right)^{2-2/p} \omega^{-2+4/p} r_0^{1+2/p}, \tag{11}$$

where ϵ denotes the root mean square fluctuations, ω is the central frequency of the signal, r_0 is the hypocentral distance and C is a numerical pre-factor which depends on the wavespeed, and on the Hurst exponent κ. The parameter p is an increasing function of κ and varies from 1.2 to 2 for $0 < \kappa \leq 1$. Formula (11) reveals interesting characteristics of the envelope broadening phenomenon. For a smooth medium, i.e., a medium poor in small-scale features, the parameter p is close to 2 and, as a consequence, the typical envelop duration is frequency-independent. Such a result also holds for the Gaussian correlation function and is in sharp contradiction with observations. As a consequence, the frequency dependence of envelope duration offers access to the degree of roughness of Earth's heterogeneities. Accurate measurements of the exponent κ requires that absorption and – to some extent – scattering losses caused by small-scale heterogeneities be incorporated in envelope modeling (Saito et al., 2002).

In practical applications to data, a characteristic time, t_q, is measured as a function of frequency and hypocentral distance: $t_q = t_p + t_{1/2}$, where t_p is the duration from onset to peak and $t_{1/2}$ is the decay time of the envelope from the peak to RMS half-peak value. Note that the observations are usually focused on the S wave train for which pulse broadening is usually much clearer than for P waves. The observations are then compared to those numerically obtained by the Markov approximation for a wide range of correlation lengths, RMS perturbations and intrinsic quality factor. In general it is not possible to estimate

independently ϵ and a since various combinations of these two parameters may yield the same characteristic time t_m. However, by analyzing separately the distance and frequency dependence of t_q, the intrinsic quality factor Q_s, the parameter p, and in turn, the Hurst exponent κ can be estimated. For applications to seismic observations, we refer to, e.g., Scherbaum and Sato (1991); Obara and Sato (1995); Saito et al. (2005).

As illustrated in Figure 2, for comparable hypocentral distances and frequency bands, the envelope shapes may strongly depend on the direct ray path. Such an observation calls for the introduction of large-scale variations of small-scale heterogeneities. Such an approach has been taken by Takahashi et al. (2007) who mapped strongly heterogeneous regions beneath quaternary volcanoes in northeastern Japan from the observed path dependence of the characteristic broadening time t_q. A complete tomographic approach to estimate the power spectrum of small-scale heterogeneities from envelope broadening data was further developed and applied to northeastern Japan by Takahashi et al. (2009).

Gusev and Abubakirov (1999a, b) developed a tomographic method to map the transport mean free path in the Earth based on body wave envelope broadening. Their approach is simpler since they consider the inversion of a single parameter. From the knowledge of the frequency dependence of the the transport mean free path, it is possible to put constraints on the parameters ϵ, a, and κ. The Markov method has recently been extended to the case of vector waves by Sato (2006, 2007, 2008). The theoretical approach relies on the introduction of scalar and vector potentials for P and S waves, respectively. The employed Helmholtz decomposition neglects the conversions between P and S waves, which is justified in the high-frequency regime $\lambda \gg a$, up to a characteristic coupling time τ^{PS} between P and S waves. Extensions of the Markov method to the vectorial case allow the modeling of the observed energy envelopes of the transverse components of the wavefield after the onset of the direct P-wave. The observation of such transverse energy is another manifestation of the randomization of the ray directions in a scattering medium. Alternative explanations for this phenomenon are mode conversions or anisotropy. In the case of a statistically homogeneous Gaussian medium with thickness h, Sato (2006) showed that the ratio between the peak total intensity and the peak intensity on the transverse component of the P pulse scales like $\epsilon^2 h/a$. Interestingly, this relation is free from absorption, since it cancels out in the intensity ratio. Based on this approach Kubanza et al. (2007) made a global study of energy envelopes of transverse components in the P wave train observed at teleseismic distances and found good correlation between the strength of heterogeneity and the tectonic setting.

Envelope modeling: radiative transfer

When the propagation time and distances become very large compared to the ballistic time, or when the acquisition geometry is in the reflection mode, it becomes crucial to take into account scattering at large angle. A powerful approach to model seismogram envelopes in scattering media in such cases is provided by radiative transfer theory. In seismology, radiative transfer was introduced in the mid-1980s by Wu (1985) with a first application to data by Wu and Aki (1988). Physically, the equation of transfer is a local statement of energy conservation in a random medium and incorporates an arbitrarily high number of scattering events. The radiative transfer equation also takes into account the anisotropy of the scattering process, in particular large-angle scattering. Therefore, it enables the complete modeling of seismogram envelopes including coda waves, i.e., the late arrivals of the seismogram. A review of the radiative transfer approach in seismology is given by Margerin (2005).

The study of coda waves has been pioneered by Aki (1969) and Aki and Chouet (1975). In this last reference, the authors developed two famous models of scattering: the single-scattering approximation and the diffusion approximation. The single-scattering model applies for propagation distances of the order of, or less than the mean free path. This model has been very popular until studies by Gusev and Abubakirov (1987) and Hoshiba (1991) pointed out the importance of multiple scattering. Recently, the diffusion model has known a revived interest in connection with seismic experiments on volcanoes (Wegler and Lühr, 2001; Wegler, 2004). The single-scattering and diffusion models are respectively short-time and large-time asymptotics of the radiative transfer equation. Mathematically, radiative transfer takes the form of an integro-differential equation for an angularly resolved energy flux known as the specific intensity $I(\mathbf{r}, \hat{\mathbf{k}}, t, \omega)$. The specific intensity quantifies the amount of energy flowing around the direction $\hat{\mathbf{k}}$ in a small frequency band $[\omega, \omega + \delta\omega]$ during a time interval $[t, t + \mathrm{d}t]$ through a small surface element located at $\mathbf{r}$. Although the concept of specific intensity was first introduced on a phenomenological basis, it is now well understood that it is in fact connected to the Wigner-Ville distribution of the wavefield (Ryzhik et al., 1996):

$$I(\mathbf{r},\hat{\mathbf{k}},t,\omega) = C \iint \langle u(t-\tau/2,\mathbf{r}+\mathbf{x}/2)u(t+\tau/2,\mathbf{r}-\mathbf{x}/2)^*\rangle e^{-i\mathbf{k}\cdot\mathbf{x}+i\omega\tau} d^3x d\tau, \tag{12}$$

where the brackets denote an ensemble average. The key to understand this definition is the notion of separation of scales. In Equation 12, one must think of the intensity as a slowly varying function of the time t, as compared to the fast oscillations of the wavefield described by the frequency ω. Similarly, the average intensity varies slowly in space as compared to the wavelength. Such a separation of scales is well satisfied by high-frequency seismograms and guarantees that the Wigner distribution can be interpreted as a local power spectrum of the wavefield.

It is important to note that interference effects are neglected in the transfer approach. This can be justified on the basis of the small phase shifts that occur upon scattering and which will be different for waves visiting different heterogeneities.

An important step in the development of radiative transfer was the introduction of the Monte Carlo method, which is a flexible numerical technique to solve the transfer equation. It has been introduced in seismology by Gusev and Abubakirov (1987). The Monte Carlo method simulates the random walk of particles, which change direction of propagation at the end of each step. The step length has an exponential distribution with parameter the scattering mean free path, and the change of direction upon scattering is provided by a probabilistic interpretation of the differential scattering cross-section. The latter quantity can be obtained from the Born approximation at frequencies below the geometrical optics limit. The validity of this approximation has been discussed by Wegler et al. (2006a) and Przybilla and Korn (2008). The Monte Carlo method can be shown to yield an exact solution of the transfer equation (Papanicolaou et al., 2000). The variance of the result typically decreases like the square root of the number of simulated random walks.

An important application of the radiative transfer approach was developed by Fehler et al. (1992) and Hoshiba (1993) who proposed a method to infer the mean free path and intrinsic quality factor in the crust from the space time dependence of the energy radiated by small earthquakes. The method is based on the estimation of the energy in three successive time windows starting with the direct *S* wave arrivals. Such multiple lapse-time window analysis has been applied in various regions of the world to give estimates of the level of heterogeneity in the crust. Typically, the mean free path of the crust around 1 Hz ranges from a few tens to a few hundred kilometers. The method has some limitations since it assumes isotropic scattering and a uniform distribution of scatterers in a half-space.

It was pointed out by Gusev and Abubakirov (1987), Abubakirov and Gusev (1990) and Hoshiba (1995) that the observed shape of envelopes of body waves is incompatible with isotropic scattering, since no broadening occurs in this case. Based on the comparison between modeled and observed *S* wave seismogram envelopes, Gusev and Abubakirov (1996) were able to infer a Hurst exponent ranging between 1/4 and 1/2 in the crust. In the case of crustal propagation at regional distances, the field is dominated by guided waves. To take into account the role of stratification and scattering in the Earth, Hoshiba (1997), Margerin et al. (1998), and Yoshimoto (2000) have shown how to incorporate depth-dependent velocity and mean free path in the Monte Carlo method. Margerin et al. (1999) numerically solved the radiative transfer equation in the crustal geometry and confirmed the importance of the leakage of energy at the Moho to explain the decay of the coda, as first proposed by Korn (1990).

Other applications of radiative transfer pertain to the study of the seismic source. Nakahara et al. (1998) proposed a method based on radiative transfer theory to infer the distribution of high-frequency radiation on the fault plane of large earthquakes. They considered the different part of the fault to be independent and modeled the propagation using a multiple isotropic scattering model. Sens-Schönfelder and Wegler (2006a) devised a method to infer the seismic moment of small crustal earthquakes recorded at regional distance. Another important application concerns the modeling of wave propagation in very heterogeneous structures such as volcanoes, where the mean free path can become as small as a few hundred meters. A review of recent development in the field is provided by Del Pezzo (2008). It is to be noted that in cases where the typical propagation distance is much larger than the mean free path, the radiative transfer equation can be simplified to a simple scalar diffusion equation for the total energy density of *P* and *S* waves. The diffusion approach to the modeling of seismic energy propagation in volcanoes has been advocated by Wegler (2004).

Radiative transfer theory is still in an active stage of development. A radiative transfer equation for elastic waves has been introduced in seismology by Zeng (1993); Sato (1994); and Ryzhik et al. (1996). This equation takes into account in a rigorous way the coupling between *P* and *S* waves. A Monte Carlo method to solve this elastic transfer equation was developed by Margerin et al. (2000). A comparison of finite-difference and Monte Carlo simulations by Przybilla and Korn (2008) demonstrated the accuracy of the radiative transfer approach to simulate the coupling and multiple scattering of *P* and *S* waves. An extension of the radiative transfer equation to incorporate the coupling between surface and body waves is a major challenge. An important step in this direction was made by Maeda et al. (2007) who incorporated the coupling between Rayleigh and body waves in the single-scattering approximation. The impact of anisotropic scale lengths has been studied by Margerin (2006) who derived a radiative transfer equation with an angular-dependent scattering mean free path.

Global-scale scattering

The subject of seismic scattering at the global scale was first developed in the early 1970s in connection with short-period precursors to phases such as *PP* (King et al., 1975) and *PKP* (Cleary and Haddon, 1972). Accompanying the improvement of the global seismic network and the availability of high-quality data, evidences for the importance of scattering in the deep Earth have accumulated in recent years. The study of precursors to deep phases such as *PKP* (Hedlin et al., 1997) and *PKKP* (Earle and Shearer, 1997) has been used to put constraints on the heterogeneities in the deep mantle and core. Observation and modeling of anomalous propagation of seismic waves in the subducting lithosphere have put forward the role of scattering by elongated heterogeneities in slabs

(Furumura and Kennett, 2005). The presence of strong scatterers in Earth's inner core has been revealed by a number of studies (Vidale and Earle, 2000; Koper et al., 2004; Poupinet and Kennett, 2004). Vidale and Earle (2000) used data from nuclear explosions recorded on the LASA array to demonstrate that the long-lasting signal following the ballistic *PKiKP* phase was not caused by the response of the crust but were genuine scattered waves radiated from the inner core. They showed examples of slowly-emergent and long-lasting *PKiKP* signals recorded at an epicentral distance of about 60°. This last point is particularly important, since it corresponds to incidence angles of *P* waves at the inner core boundary such that the reflection coefficient vanishes. The coda of *PKiKP* thus provides a direct proof of the presence of scatterers in the inner core.

Since the deep Earth interior is usually thought to be much less heterogeneous than the lithosphere, much of the modeling work has been focused on the single-scattering approximation (Hedlin et al. (1997); Vidale and Earle (2000); Leyton and Koper (2007)). The radiative transfer approach was introduced in global seismology by Margerin and Nolet (2003a, b) to improve the modeling of the precursors to the *PKP* wave. In particular, they demonstrated the ability of radiative transfer to incorporate complicated ray geometry such as the one imposed by the severe drop of velocity at the core–mantle boundary. The Monte Carlo method was further developed by Shearer and Earle (2004) who proposed a scheme to simulate the multiple scattering of elastic waves at the global scale, including the coupling between *P* and *S* waves. Their method was applied to the study of the coda envelopes of teleseismic *P* waves. To explain the observations, they proposed a 1-D model of Earth small-scale heterogeneity composed of a heterogeneous 200 km top-layer (4% RMS perturbations) overlying a less heterogeneous 400-km layer (3% RMS perturbations). At the base of the mantle, a 600 km thick layer with 0.5% RMS perturbations is required to achieve a good fit to the data. Shearer and Earle (2004) concluded that attenuation is dominated by scattering in the upper-mantle and by absorption in the lower mantle.

In addition to global studies of scattering of short-period waves, an interest in the scattering of long-period surface waves has recently arisen. Maeda et al. (2006) performed an (*f,k*) analysis of the vertical components of three large earthquakes recorded on the Japanese Hi-Net in the 90–180 s period band. They find that the signals recorded between the first multi-orbiting Rayleigh wave arrivals are composed of waves coming from all directions with phase velocities of the order of the Rayleigh wave fundamental mode. This suggests that in the period band of interest, scattered Rayleigh waves are present in the data at short lapse time. After typically 30,000 s, the multi-orbiting Rayleigh waves have strongly decayed. The (*f,k*) analysis demonstrates that the signal is dominated by waves with very large phase velocities, thereby implying the dominance of higher-mode Rayleigh waves. Such a phenomenon is to be expected on the ground that higher modes sample the deeper parts of the Earth with high intrinsic quality factor. To model the coda of multi-orbiting Rayleigh waves, Sato and Nishino (2002) developed a radiative transfer equation on the sphere assuming the multiple isotropic scattering of the fundamental mode Rayleigh wave. In their approach, the coupling with higher modes is neglected. They observe a systematic discrepancy between observed and modeled seismogram envelopes which also suggests that higher modes dominate the data at lapse times larger than 30,000 s.

Interferometry with scattered waves

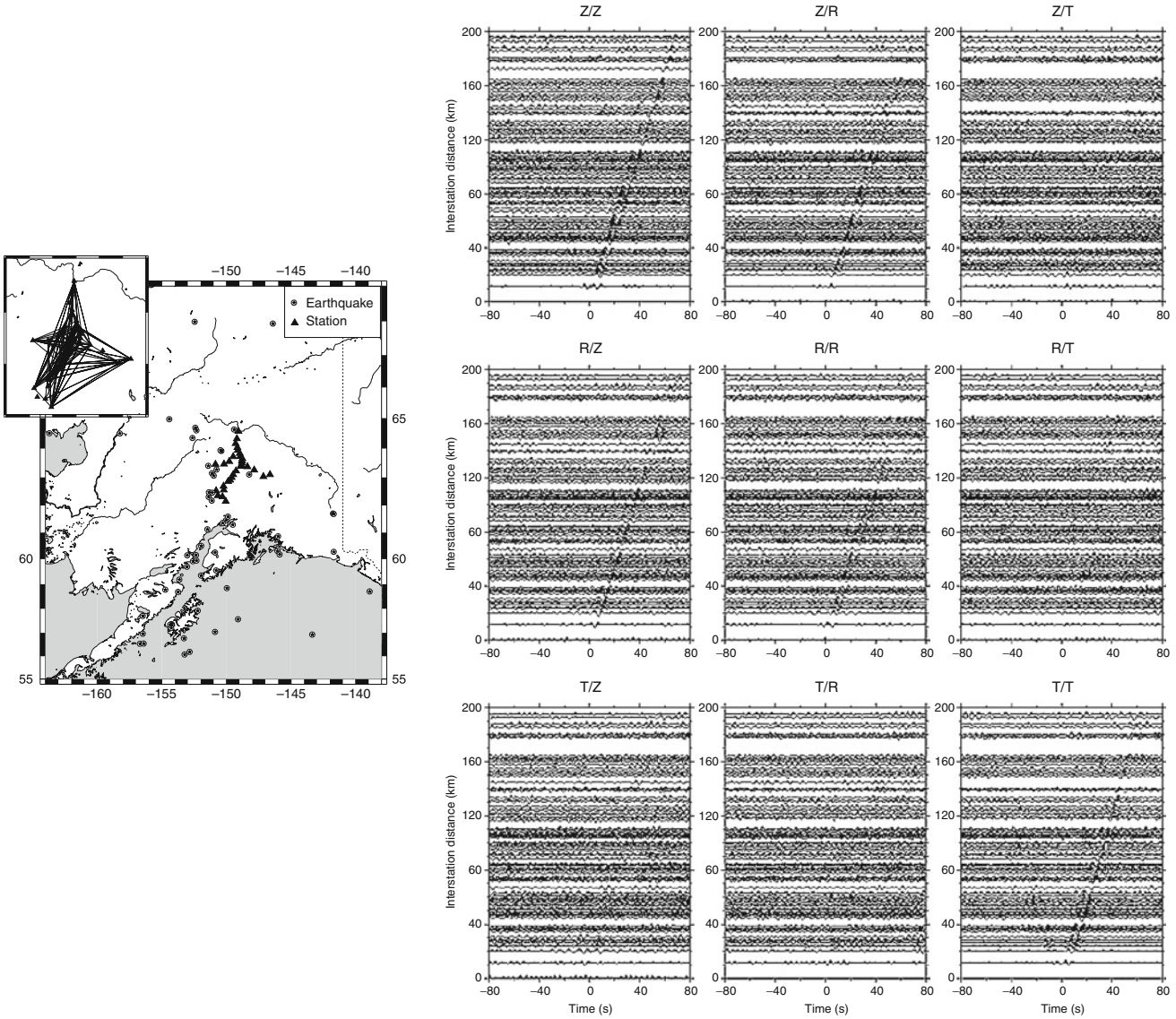
Green function retrieval

As explained in section Introduction, the incoherent field refers to the waves that average out when a mean over an ensemble of random media is performed. These waves are incoherent in the sense that, at a given point of the medium, they propagate in space directions that differ from that of the mean field. But "incoherence" does not imply that the phase information has been lost (Campillo, 2006). Campillo and Paul (2003) were the first to demonstrate the existence of correlations between the coda waves recorded at two distant points in the heterogeneous crust. Such correlations emerge after sufficient temporal and source averaging. Following Lobkis and Weaver (2001), Campillo and Paul (2003) proposed that the correlation tensor of the coda wavefields recorded at two points *A* and *B* is in fact proportional to the Green's tensor G_{ij} between these two points. In the frequency domain, such a relation can be mathematically formulated as follows:

$$\left\langle u_i(B,\omega)u_j(A,\omega)^*\right\rangle \propto \mathrm{Im}G_{ij}(B,A,\omega). \tag{13}$$

The proportionality factor depends on the details of the source spectrum. The occurrence of the imaginary part of the Green's function on the right-hand side of Equation 13 physically means that the correlation of two wavefields is proportional to a combination of the retarded and advanced Green's functions. The appearance of the advanced Green's function is most easily understood by noting the equivalence between correlations and time-reversal experiments. We will not discuss this analogy, but the interested reader can refer to Derode et al. (2003).

In seismology, the brackets in Equation 13 usually denote an average over time windows in the coda and over earthquake sources. An example of emergence of the retarded and advanced Green's function in the correlation of coda waves recorded on a temporary network deployed in Alaska is illustrated in Figure 3. The maximum of the correlation is always shown for positive times, but it is nevertheless possible to distinguish some coherent arrivals at negative times. The reconstruction of the Green's function from coda waves is intimately related to the concept of equipartition, which is a fundamental consequence of multiple scattering. Loosely stated, equipartition stipulates that plane *P, S*, Rayleigh, and Love



Seismic Waves, Scattering, Figure 3 Example of Green's function reconstruction from coda waves recorded in Alaska. A location map of the experiment is shown on the *left*. The inset shows all the station pairs analyzed in this study. On the *right*, all possible terms of the correlation tensor are represented as a function of the correlation time on the horizontal axis and the epicentral distance between the two stations on the vertical axis. For each station pair, the three components of the seismometers have been rotated onto the radial (R), transverse (T), and vertical (Z) reference frame (Reproduced from Paul et al. (2005)).

waves coming from all possible directions compose the wavefield. That the Green's function of the medium is recovered if the equipartition state is reached has been shown for an infinite elastic medium by Sánchez-Sesma and Campillo (2006). Based on the single-scattering approximation for scalar waves, Sato (2009) demonstrated that the correlation of coda waves recorded at two stations enables the reconstruction of the ballistic Green's function between the two stations. Snieder (2004b) derived the reconstruction of the Green's function from coda waves based on the stationary phase approximation. An experimental verification of the equipartition principle has been performed by Hennino et al. (2001). In Figure 3, we can observe a clear time asymmetry in the reconstructed Green's function. This seems to contradict the equipartition principle. Such a time asymmetry has been studied by Paul et al. (2005). These authors demonstrate that a flux of energy directed from the earthquake source to the station can persist even at very large lapse time in the coda. Such an energy flux breaks the temporal symmetry of the correlations unless the distribution of earthquakes around the station covers all azimuths.

Monitoring temporal variations

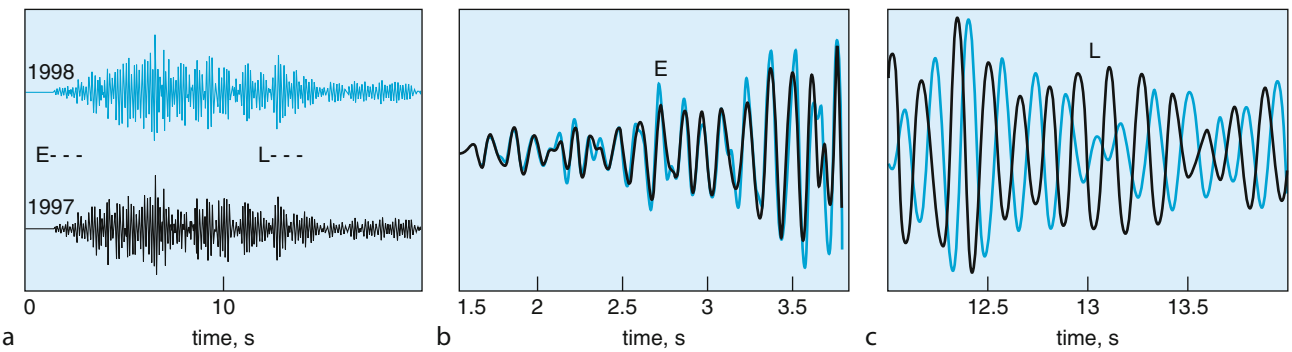
An interesting application of wave multiple scattering pertains to the detection of weak changes in dynamic media, i.e., media evolving with time. The methods which make use of scattered waves to probe changes in complex media have been termed "diffusing acoustic wave spectroscopy" in acoustics and "coda wave interferometry" (CWI) in seismology. In this article, we focus on applications to seismic waves. A theoretical approach of the method is provided by Snieder (2006). For a very accessible review of this broad topic in physics, we refer to Snieder and Page (2007). A comprehensive review of monitoring applications in seismology is given by Poupinet et al. (2008).

The key idea of CWI is to take advantage of the long propagation paths of scattered waves, which accumulate small phase differences on their way from source to station. An example of monitoring the medium changes at the Merapi volcano is shown in Figure 4. In this experiment, a repeatable source is used to generate seismic waves that propagate through the volcano. The fact that the source is highly reproducible is important, as any small difference between seismograms recorded at different time instants can be interpreted as a change of the propagating medium. Natural repeating sources such as earthquake doublets can also be exploited to probe medium changes. The doublets correspond to small earthquakes breaking the same fault patch at different times of occurrence (Poupinet et al., 1984). As shown in Figure 4, the changes are virtually undetectable on the first arrivals. In the coda, the medium changes are manifested by a small time delay between two very similar waveforms. In the case of the Merapi data shown in Figure 4, an analysis of the delay times as a function of the time lag in the coda reveals that the two seismograms are nearly stretched versions of one another. This means that the two waveforms can be related by a similarity transformation $t \rightarrow t(1+\epsilon)$ of the time variable. Such an observation can be interpreted as a very small change of the wave velocity δv in the propagating medium. The relation between the stretching parameter ϵ and the velocity variation reads (Sens-Schönfelder and Wegler 2006b):

$$\epsilon = -\frac{\delta v}{v} \tag{14}$$

In practice, the velocity change can be estimated by measuring the delay time which maximizes the correlation between the two signals. Such a procedure can be applied to moving time windows in the coda, of typical duration, a few central periods. The relation (14) is valid for acoustic waves only. A formula valid for elastic coda waves in the equipartition regime has been derived by Snieder (2002). Ratdomopurbo and Poupinet (1995), and Wegler, et al. (2006b) found that the shear velocity increases prior to eruptions at the Merapi volcano, an observation interpreted as a pressurization effect due to the magma ascent. Typically, relative velocity changes of the order of 10^{-5}–10^{-4} can be detected (Poupinet et al., 2008). This makes coda wave interferometry the ideal tool to study the rheology of materials (Snieder et al., 2002).

As a final example of the use of scattered waves, we consider a method which has been termed "passive image interferometry" (Sens-Schönfelder and Wegler, 2006b) which combines coda wave interferometry with the principle of Green's function reconstruction from coda waves. Passive interferometry uses ambient noise – instead of coda waves – to reconstruct the Green's function. The coda part of the reconstructed Green's function is subsequently used to monitor temporal variations in the medium. A great advantage of this method as compared to the usual coda wave interferometry lies in the fact that noise records allow a continuous reconstruction of the Green's function over very long time periods. Brenguier et al. (2008) demonstrated the possibility to detect very



Seismic Waves, Scattering, Figure 4 Illustration of the principle of coda wave interferometry. A controlled and repeatable source is used to probe the Merapi volcano with seismic waves. The signals recorded at a time interval of 1 year are shown by *blue and black lines*. The insets show the comparisons of waveforms in the early part of the signal (E) and in the late part of the signal (L). In the later portion, a clear time shift between the two traces can be observed (Reproduced from Snieder (2004a)).

small medium changes at Le Piton de la Fournaise volcano in la Réunion Island. Applications to the study of the recovery phenomena in sediments affected by strong ground motion are developed in Sawazaki et al. (2009).

Summary

This article presents a nontechnical introduction to the various phenomena and methods pertaining to the scattering of seismic waves. The basic concepts of seismic scattering are introduced with the aid of a thought-experiment. The topics covered include the interpretation of transmission fluctuations, the broadening of envelope at regional distances, the interpretation of coda waves at local and global scales, and the use of scattered waves in seismic interferometry. The methods developed in seismic scattering offer powerful tools for the stochastic imaging of Earth's structure at spatial scales that cannot be resolved by traditional tomographic methods. In addition, scattered seismic waves are extremely sensitive to slight temporal changes of the medium. Using the interferometric approach, multiply-scattered coda waves can be used to monitor the evolution of the dynamic Earth.

Bibliography

Abubakirov, I., and Gusev, A., 1990. Estimation of scattering properties of lithosphere of Kamchatka based on Monte-Carlo simulation of record envelope of a near earthquake. *Physics of the Earth and Planetary Interiors*, **64**, 52–67.

Aki, K., 1969. Analysis of the seismic coda of local earthquakes as scattered waves. *Journal of Geophysical Research*, **74**, 615–618.

Aki, K., 1973. Scattering of P waves under the Montana LASA. *Journal of Geophysical Research*, **78**, 1334–1346.

Aki, K., and Chouet, B., 1975. Origin of coda waves, sources and attenuation. *Journal of Geophysical Research*, **80**, 3322–3342.

Brenguier, F., Shapiro, N. M., Campillo, M., Ferrazzini, V., Duputel, Z., Coutant, O., and Nercessian, A., 2008. Toward forecasting volcanic eruptions using seismic noise. *Nature Geoscience*, **1**, 126–130.

Campillo, M., 2006. Phase and correlation in 'Random' seismic fields and the reconstruction of the green function. *Pure and Applied Geophysics*, **163**, 475–502.

Campillo, M., and Paul, A., 2003. Long range correlations in the diffuse seismic coda. *Science*, **299**, 547–549.

Chen, X., and Aki, K., 1991. General coherence functions for amplitude and phase fluctuations in a randomly heterogeneous medium. *Geophysical Journal International*, **105**, 155–162.

Cleary, J., and Haddon, R., 1972. Seismic wave scattering near the Core-Mantle Boundary: a new interpretation of precursors to PKP. *Nature*, **240**, 549–551.

Del Pezzo, E., 2008. Seismic wave scattering in volcanoes. In Sato, H., Fehler, M., and Dmowska, R. (eds.), *Earth Heterogeneity and Scattering Effect on Seismic Waves*. New York: Academic. Advances in Geophysics, Vol. 50, pp. 353–369.

Derode, A., Tourin, A., and Fink, M., 2001. Random multiple scattering of ultrasound.I. Coherent and ballistic waves. *Physical Review E*, **64**, 36–605.

Derode, A., Larose, E., Tanter, M., de Rosny, J., Tourin, A., Campillo, M., and Fink, M., 2003. Recovering the Green's function from field-field correlations in an open scattering medium. *The Journal of the Acoustical Society of America*, **113**, 2973–2976.

Dziewonski, A., and Anderson, D., 1981. Preliminary reference Earth model. *Physics of the Earth and Planetary Interiors*, **25**, 297–356.

Earle, P., and Shearer, P., 1997. Observations of PKKP precursors used to estimate small-scale topography on the core-mantle boundary. *Science*, **277**, 667–670.

Fehler, M., Hoshiba, M., Sato, H., and Obara, K., 1992. Separation of scattering and intrinsic attenuation for the Kanto-Tokai region, Japan, using measurements of S-wave energy versus hypocentral distance. *Geophysical Journal International*, **108**, 787–800.

Flatté, S., and Wu, R., 1988. Small-scale structure in the lithosphere and asthenosphere deduced from arrival time and amplitude fluctuations at NOR-SAR. *Journal of Geophysical Research*, **93**, 6601–6614.

Furumura, T., and Kennett, B., 2005. Subduction zone guided waves and the heterogeneity structure of the subducted plate: Intensity anomalies in northern Japan. *Journal of Geophysical Research*, **110**, B10302.

Gusev, A., and Abubakirov, I., 1987. Monte-Carlo simulation of record envelope of a near earthquake. *Physics of the Earth and Planetary Interiors*, **49**, 30–36.

Gusev, A., and Abubakirov, I., 1996. Simulated envelopes of non-isotropically scattered body waves as compared to observed ones: another manifestation of fractal heterogeneity. *Geophysical Journal International*, **127**, 49–60.

Gusev, A., and Abubakirov, I., 1999a. Vertical profile of effective turbidity reconstructed from broadening of incoherent body-wave pulses. I- General approach and the inversion procedure. *Geophysical Journal International*, **136**, 295–308.

Gusev, A., and Abubakirov, I., 1999b. Vertical profile of effective turbidity reconstructed from broadening of incoherent body-wave pulses. II- Application to Kamchatka data. *Geophysical Journal International*, **136**, 309–323.

Hedlin, M., Shearer, P., and Earle, P., 1997. Seismic evidence for small-scale heterogeneity throughout the Earth's mantle. *Nature*, **387**, 145–150.

Hennino, R., Trégourès, N., Shapiro, N. M., Margerin, L., Campillo, M., van Tiggelen, B. A., and Weaver, R. L., 2001. Observation of equipartition of seismic waves. *Physical Review Letters*, **86**, 3447–3450.

Hoshiba, M., 1991. Simulation of multiple scattered coda wave excitation based on the energy conservation law. *Physics of the Earth and Planetary Interiors*, **67**, 123–136.

Hoshiba, M., 1993. Separation of scattering attenuation and intrinsic absorption in Japan using the multiple lapse time window analysis of full seismogram envelope. *Journal of Geophysical Research*, **98**, 15809–15824.

Hoshiba, M., 1995. Estimation of nonisotropic scattering in western Japan using coda wave envelopes: application of a multiple nonisotropic scattering model. *Journal of Geophysical Research*, **100**, 645–657.

Hoshiba, M., 1997. Seismic coda wave envelope in depth-dependent S wave velocity structure. *Physics of the Earth and Planetary Interiors*, **104**, 15–22.

King, D., Haddon, R., and Husebye, E., 1975. Precursors to PP. *Physics of the Earth and Planetary Interiors*, **10**, 103–127.

Klimeš, L., 2002. Correlation functions of random media. *Pure and Applied Geophysics*, **159**, 1811–1831.

Koper, K., Franks, J., and Dombrovskaya, M., 2004. Evidence for small-scale heterogeneity in Earth's inner core from a global study of PKiKP coda waves. *Earth and Planetary Science Letters*, **228**, 227–241.

Korn, M., 1990. A modified energy flux model for lithospheric scattering of teleseismic body waves. *Geophysical Journal International*, **102**, 165–175.

Kravtsov, Y., Müller, T., Shapiro, S., and Buske, S., 2003. Statistical properties of reflection traveltimes in 3-D randomly inhomogeneous and anisomeric media. *Geophysical Journal International*, **154**, 841–851.

Kubanza, M., Nishimura, T., and Sato, H., 2007. Evaluation of strength of heterogeneity in the lithosphere from peak amplitude analyses of teleseismic short-period vector P waves. *Geophysical Journal International*, **171**, 390–398.

Leyton, F., and Koper, K., 2007. Using PKiKP coda to determine inner core structure: 2. Determination of QC. *Journal of Geophysical Research*, **112**, B05317.

Lobkis, O. I., and Weaver, R. L., 2001. On the emergence of the Green's function in the correlations of a diffuse field. *The Journal of the Acoustical Society of America*, **110**, 3011–3017.

Maeda, T., Sato, H., and Ohtake, M., 2006. Constituents of vertical-component coda waves at long periods. *Pure and Applied Geophysics*, **163**, 549–566.

Maeda, T., Sato, H., and Nishimura, T., 2007. Synthesis of coda wave envelopes in randomly inhomogeneous elastic media in a half-space: single scattering model including Rayleigh waves. *Geophysical Journal International*, **172**, 130–154.

Margerin, L., 2005. Introduction to radiative transfer of seismic waves. In Nolet, G., and Levander, A. (eds.), *Seismic Earth: Analysis of Broadband Seismograms*. Washington: American Geophysical Union. Geophysical Monograph, Vol. 157, pp. 229–252.

Margerin, L., 2006. Attenuation, transport and diffusion of scalar waves in textured random media. *Tectonophysics*, **416**, 229–244.

Margerin, L., and Nolet, G., 2003a. Multiple scattering of high-frequency seismic waves in the deep earth: modeling and numerical examples. *Journal of Geophysical Research*, **108**, 2234.

Margerin, L., and Nolet, G., 2003b. Multiple scattering of high-frequency seismic waves in the deep earth: PKP precursor analysis and inversion for mantle granularity. *Journal of Geophysical Research*, **108**, 2514.

Margerin, L., Campillo, M., and van Tiggelen, B. A., 1998. Radiative tranfer and diffusion of waves in a layered medium: new insight into coda Q. *Geophysical Journal International*, **134**, 596–612.

Margerin, L., Campillo, M., Shapiro, N., and van Tiggelen, B. A., 1999. Residence time of diffuse waves in the crust as a physical interpretation of coda Q: application to seismograms recorded in mexico. *Geophysical Journal International*, **138**, 343–352.

Margerin, L., Campillo, M., and van Tiggelen, B. A., 2000. Monte Carlo simulation of multiple scatterring of elastic waves. *Journal of Geophysical Research*, **105**, 7873–7892.

Nakahara, H., Nishimura, T., Sato, H., and Ohtake, M., 1998. Seismogram envelope inversion for the spatial distribution of high-frequency energy radiation from the earthquake fault: application to the 1994 far east off Sanriku earthquake, Japan. *Journal of Geophysical Research*, **103**, 855–867.

Obara, K., and Sato, H., 1995. Regional differences of random inhomogeneities around the volcanic front in the Kanto-Tokai area, Japan, revealed from the broadening of *S* wave seismogram envelopes. *Journal of Geophysical Research*, **100**, 2103–2121.

Papanicolaou, G., Bal, G., and Ryzhik, L., 2000. Probabilistic theory of transport processes with polarization. *SIAM Journal on Applied Mathematics*, **60**, 1639–1666.

Paul, A., Campillo, M., Margerin, L., Larose, E., and Derode, A., 2005. Empirical synthesis of time-asymmetrical Green functions from the correlation of coda waves. *Journal of Geophysical Research*, **110**, B08302.

Poupinet, G., and Kennett, B., 2004. On the observation of high frequency PKiKP and its coda in Australia. *Physics of the Earth and Planetary Interiors*, **146**, 497–511.

Poupinet, G., Ellsworth, W., and Frechet, J., 1984. Monitoring velocity variations in the crust using earthquake doublets: an application to the Calaveras fault, California. *Journal of Geophysical Research*, **89**, 5719–5731.

Poupinet, G., Got, J.-L., and Brenguier, F., 2008. Monitoring temporal variations of physical properties in the crust by cross-correlating the waveforms of seismic doublets. In Sato, H., Fehler, M. and Dmowska, R. (eds.), *Advances in Geophysics : Earth Heterogeneity and Scattering Effect on Seismic Waves*. New York: Academic Press, Vol. 50, pp. 374–399.

Przybilla, J., and Korn, M., 2008. Monte Carlo simulation of radiative energy transfer in continuous elastic random media-three-component envelopes and numerical validation. *Geophysical Journal International*, **173**, 566–576.

Ratdomopurbo, A., and Poupinet, G., 1995. Monitoring a temporal change of seismic velocity in a volcano: application to the 1992 eruption of Mt. Merapi (Indonesia). *Geophysical Research Letters*, **22**, 775–778.

Rytov, S. M., Kravtsov, Y. A., and Tatarskii, V. I., 1989a. *Principles of Statistical Radiophysics*. New York: Springer. Wave Propagation through Random Media, Vol. 4.

Rytov, S. M., Kravtsov, Y. A., and Tatarskii, V. I., 1989b. *Principles of Statistical Radiophysics*. New York: Springer. Elements of Random Fields, Vol. 1.

Ryzhik, L., Papanicolaou, G., and Keller, J. B., 1996. Transport equation for elastic and other waves in random media. *Wave Motion*, **24**, 327–370.

Saito, T., Sato, H., and Ohtake, M., 2002. Envelope broadening of spherically outgoing waves in three-dimensional random media having power law spectra. *Journal of Geophysical Research*, **107**, 2089.

Saito, T., Sato, H., Ohtake, M., and Obara, K., 2005. Unified explanation of envelope broadening and maximum-amplitude decay of high-frequency seismograms based on the envelope simulation using the Markov approximation: forearc side of the volcanic front in northeastern Honshu, Japan. *Journal of Geophysical Research*, **110**, B01304.

Sánchez-Sesma, F. J., and Campillo, M., 2006. Retrieval of the green function from cross correlation: the canonical elastic problem. *Bulletin Seismological Society of America*, **96**, 1182–1191.

Sato, H., 1982. Amplitude attenuation of impulsive waves in random media based on travel time corrected mean wave formalism. *The Journal of the Acoustical Society of America*, **71**, 559–564.

Sato, H., 1984. Attenuation and envelope formation of three-component seismograms of small local earthquakes in randomly inhomogeneous lithosphere. *Journal of Geophysical Research*, **89**, 1221–1241.

Sato, H., 1989. Broadening of seismogram envelopes in the randomly inhomogeneous lithosphere based on the parabolic approximation: southeastern Honshu, Japan. *Journal of Geophysical Research*, **94**, 17735–17747.

Sato, H., 1994. Multiple isotropic scattering model including P-S conversion for the seismogram envelope formation. *Geophysical Journal International*, **117**, 487–494.

Sato, H., 2006. Synthesis of vector wave envelopes in three-dimensional random elastic media characterized by a Gaussian autocorrelation function based on the Markov approximation: plane wave case. *Journal of Geophysical Research*, **111**, B06306.

Sato, H., 2007. Synthesis of vector wave envelopes in three-dimensional random elastic media characterized by a Gaussian autocorrelation function based on the Markov approximation: spherical wave case. *Journal of Geophysical Research*, **112**, B01301.

Sato, H., 2008. Synthesis of vector-wave envelopes in 3-D random media characterized by a nonisotropic Gaussian ACF based on the Markov approximation. *Journal of Geophysical Research*, **113**, B08304.

Sato, H., 2009. Green's function retrieval from the CCF of coda waves in a scattering medium. *Geophysical Journal International*, **179**, 1580–1583.

Sato, H., and Fehler, M., 1998. *Seismic Wave Propagation and Scattering in the Heterogeneous Earth*. New York: Springer.
Sato, H., and Korn, M., 2009. Synthesis of vector-wave envelopes in random elastic media on the basis of the markov approximation. In Sato, H., Fehler, M., and Dmowska, R. (eds.), *Earth Heterogeneity and Scattering Effect on Seismic Waves*. New York: Academic. Advances in Geophysics, Vol. 50, pp. 43–93.
Sato, H., and Nishino, M., 2002. Multiple isotropic-scattering model on the spherical Earth for the synthesis of Rayleigh-wave envelopes. *Journal of Geophysical Research*, **107**, 2343.
Sato, H., Fehler, M., and Dmowska, R. (eds.), 2008. *Earth Heterogeneity and Scattering Effects on Seismic Waves*. Amsterdam: Academic. Advances in Geophysics, Vol. 50.
Sawazaki, K., Sato, H., Nakahara, H., and Nishimura, T., 2009. Time-Lapse changes of seismic velocity in the shallow ground caused by strong ground motion shock of the 2000 Western-Tottori Earthquake, Japan, as revealed from coda deconvolution analysis. *Bulletin Seismological Society of America*, **99**, 352–366.
Scherbaum, F., and Sato, H., 1991. Inversion of full seismogram envelopes based on the parabolic approximation: estimation of randomness and attenuation in southeast Honshu, Japan. *Journal of Geophysical Research*, **96**, 2223–2232.
Sens-Schönfelder, C., and Wegler, U., 2006a. Radiative transfer theory for estimation of the seismic moment. *Geophysical Journal International*, **167**, 1363–1372.
Sens-Schönfelder, C., and Wegler, U. C., 2006b. Passive image interferometry and seasonal variations of seismic velocities at merapi volcano, indonesia. *Geophysical Research Letters*, **33**, L21302.
Shapiro, S. A., and Kneib, G., 1993. Seismic attenuation by scattering: theory and numerical results. *Geophysical Journal International*, **114**, 373–391.
Shearer, P., and Earle, P., 2004. The global short-period wavefield modelled with a Monte Carlo seismic phonon method. *Geophysical Journal International*, **158**, 1103–1117.
Shearer, P., and Earle, P., 2008. Observing and modeling elastic scattering in the deep Earth. In Sato, H., Fehler, M., and Dmowska, R. (eds.), *Earth Heterogeneity and Scattering Effect on Seismic Waves*. New York: Academic. Advances in Geophysics, Vol. 50, pp. 167–193.
Sheng, P., 1995. *Introduction to Wave Scattering, Localization and Mesoscopic Phenomena*. San Diego: Academic.
Shiomi, K., Sato, H., and Ohtake, M., 1997. Broad-band power-law spectra of well-log data in Japan. *Geophysical Journal International*, **130**, 57–64.
Snieder, R., 2002. Coda wave interferometry and the equilibration of energy in elastic media. *Physical Review E*, **66**, 46615.
Snieder, R., 2004a. Coda wave interferometry. In McGraw-Hill Yearbook of Science & Technology. New York: McGraw-Hill, pp. 54–56.
Snieder, R., 2004b. Extracting the Greens function from the correlation of coda waves: a derivation based on stationary phase. *Physical Review E*, **69**, 46610.
Snieder, R., 2006. The theory of coda wave interferometry. *Pure and Applied Geophysics*, **163**, 455–473.
Snieder, R., and Page, J., 2007. Multiple scattering in evolving media. *Physics Today*, **60**, 49–55.
Snieder, R., Grêt, A., Douma, H., and Scales, J., 2002. Coda wave interferometry for estimating nonlinear behavior in seismic velocity. *Science*, **295**, 2253–2255.
Takahashi, T., Sato, H., Nishimura, T., and Obara, K., 2007. Strong inhomogeneity beneath Quaternary volcanoes revealed from the peak delay analysis of S-wave seismograms of microearthquakes in northeastern Japan. *Geophysical Journal International*, **168**, 90–99.
Takahashi, T., Sato, H., Nishimura, T., and Obara, K., 2009. Tomographic inversion of the peak delay times to reveal random velocity fluctuations in the lithosphere: method and application to northeastern Japan. *Geophysical Journal International*, **178**, 1437–1455.
Vidale, J., and Earle, P., 2000. Fine-scale heterogeneity in the Earth's inner core. *Nature*, **404**, 273–275.
Wegler, U., 2004. Diffusion of seismic waves in a thick layer: theory and application to Vesuvius volcano. *Journal of Geophysical Research*, **109**, 07303.
Wegler, U., and Lühr, B., 2001. Scattering behaviour at Merapi volcano (Java) revealed from an active seismic experiment. *Geophysical Journal International*, **145**, 579–592.
Wegler, U., Korn, M., and Przybilla, J., 2006a. Modeling full seismogram envelopes using radiative transfer theory with Born scattering coefficients. *Pure and Applied Geophysics*, **163**, 503–531.
Wegler, U., Lühr, B., Snieder, R., and Ratdomopurbo, A., 2006b. Increase of shear wave velocity before the 1998 eruption of Merapi volcano (Indonesia). *Geophysical Research Letters*, **33**, L09303.
Wu, R. S., 1982. Attenuation of short period seismic waves due to scattering. *Geophysical Research Letters*, **9**, 9–12.
Wu, R. S., 1985. Multiple scattering and energy transfer of seismic waves-separation of scattering effect from intrinsic attenuation I. Theoretical modelling. *Geophysical Journal International*, **82**, 57–80.
Wu, R. S., 1989. Seismic wave scattering. In James, D. E. (ed.), *The Encyclopedia of Solid Earth Geophysics*. New York: Van Nostrand Reinhold, pp. 1166–1187.
Wu, R. S., and Aki, K., 1988. Multiple scattering and energy transfer of seismic waves-separation of scattering effect from intrinsic attenuation II. Application of the theory to Hindu Kush region. *Pure and Applied Geophysics*, **128**, 49–80.
Wu, R., and Flatté, S., 1990. Transmission fluctuations across an array and heterogeneities in the crust and upper mantle. *Pure and Applied Geophysics*, **132**, 175–196.
Wu, R., Xu, Z., and Li, X., 1994. Heterogeneity spectrum and scale-anisotropy in the upper crust revealed by the German continental deep-drilling (KTB) holes. *Geophysical Research Letters*, **21**, 911–914.
Yoshimoto, K., 2000. Monte Carlo simulation of seismogram envelopes in scattering media. *Journal of Geophysical Research*, **105**, 6153–6161.
Zeng, Y., 1993. Theory of scattered P-and S-wave energy in a random isotropic scattering medium. *Bulletin Seismological Society of America*, **83**, 1264–1276.
Zheng, Y., and Wu, R., 2005. Measurement of phase fluctuations for transmitted waves in random media. *Geophysical Research Letters*, **32**, L14314.
Zheng, Y., and Wu, R., 2008. Theory of transmission fluctuations in random media with a depth-dependent background velocity structure. In Sato, H., Fehler, M., and Dmowska, R. (eds.), *Earth Heterogeneity and Scattering Effect on Seismic Waves*. New York: Academic. Advances in Geophysics, Vol. 50, pp. 21–41.
Zheng, Y., Wu, R., and Lay, T., 2007. Inverting the power spectrum for a heterogeneous medium. *Geophysical Journal International*, **168**, 1005–1010.

Cross-references

Energy Partitioning of Seismic Waves
Propagation of Elastic Waves: Fundamentals
Seismic Diffraction
Seismic, Ambient Noise Correlation

SEISMIC ZONATION

Yanxiang Yu, Mengtan Gao, Guangyin Xu
Institute of Geophysics, China Earthquake Administration, Haidian District, Beijing, China

Definition

Seismic zonation. The process of subdividing the territory into regions with respect to the level of seismic hazard. The result of seismic zonation is usually presented as a map, which is based on seismic hazard map.

Introduction

Seismic zonation is useful for hazard reduction such as earthquake-resistant design of structures, risk analysis, land-use planning, etc. Many earthquake-prone countries developed seismic zonation maps. Seismic zonation map is usually revised or updated periodically with the progress in methodology and accumulation of new data. Seismic intensity (see *Earthquakes, Intensity*) or ground motion parameters such as peak ground acceleration (PGA), peak ground velocity (PGV), and spectral acceleration at specific natural period are mostly adopted in seismic zonation map. In the early stage, most seismic zonation maps were in terms of intensity, but since the 1980s, ground motion parameters have become popular. More commonly, such maps take the exceeding probability of 10% within 50 years (return period 475 years) as standard.

Methodology

The basic method to develop seismic zonation map is the seismic hazard analysis (see *Seismic Hazard*) approach. Both deterministic and probabilistic approaches are adopted. Although some new deterministic approaches based on the computation of synthetic seismograms are used in recent years in developing seismic zonation maps (Parvez et al., 2003), nevertheless probabilistic approaches are more popular. The seismic zonation maps, based on seismic hazard maps (see *Seismology, Global Earthquake Model*), of many countries are developed by the use of probabilistic approach. The well-known USGS National Seismic Hazard Maps (Algermissen and Perkins, 1976; Algermissen et al., 1990; Frankel et al., 1996; Frankel et al., 2002; Petersen et al., 2008) and the Global Seismic Hazard Assessment Programme (GSHAP) are the typical (see *Seismic Hazard*) ones. In this article, only probabilistic method is briefly introduced.

The probabilistic method (usually referred to as PSHA – probabilistic seismic hazard analysis) was first introduced by Cornell (1968) and since then widely adopted and modified (McGuire, 1978; Bender and Perkins, 1982). There are four basic steps for assessment of PSHA (Figure 1):

Step 1: Definition of seismic sources. Sources may range from small faults to large seismotectonic provinces with uniform seismicity. The type of seismic source can be both line or area sources.

Step 2: Definition of seismicity recurrence characteristic for the sources, where each source is described by an earthquake probability distribution, or recurrence relationship. A recurrence relationship indicates the chance of an earthquake of a given magnitude to occur anywhere inside the source during a specified period of time. An upper bound earthquake is chosen for each source, which represents the maximum event to be considered.

Step 3: Development of ground motion attenuation relationships. This is usually done empirically from strong motion records (see *Earthquakes, Strong-Ground Motion*).

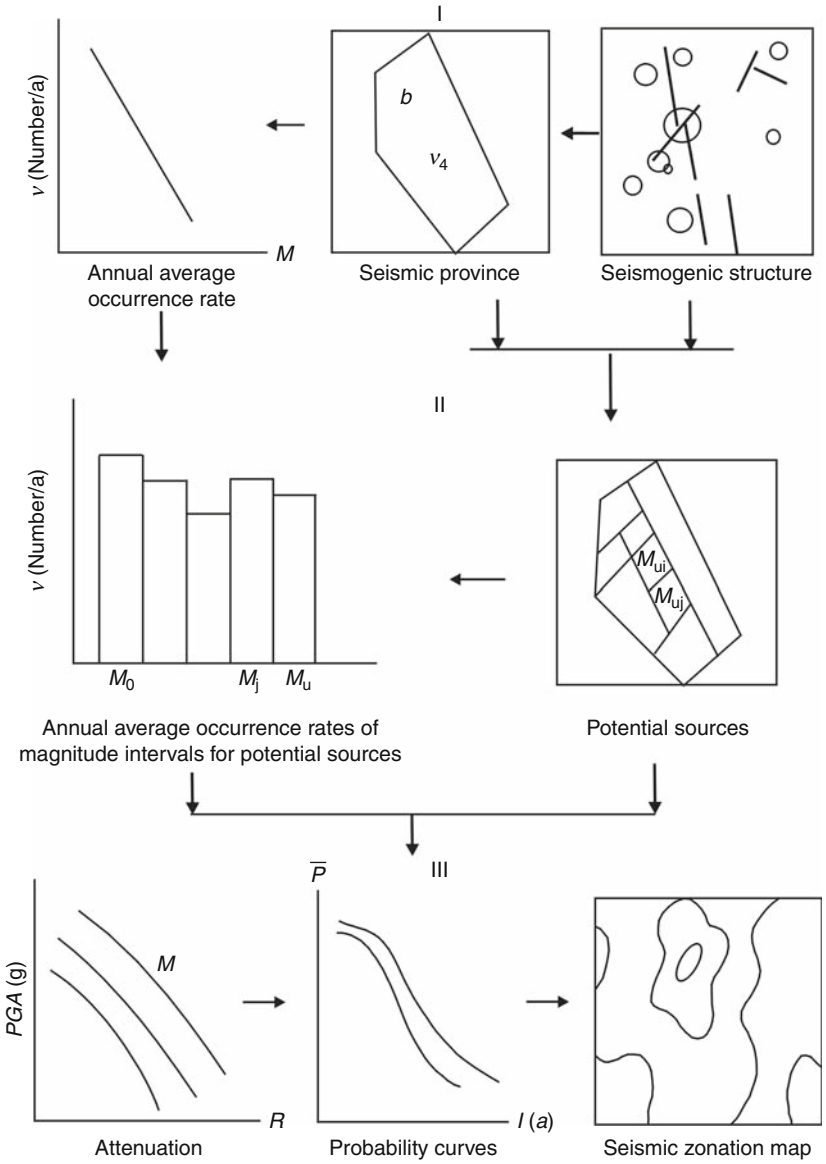
Step 4: Determination of the hazard at each site. In this case, the effects of all the earthquakes of different sizes occurring at different locations in different earthquake sources at different probabilities of occurrence are integrated into one curve that shows the probability of exceeding different levels of ground motion level (such as PGA) at the site during a specified period of time.

There are two basic assumptions in the seismic hazard analysis method at present: (a) Seismicity in the region around the site in the past indicates that in future, the recurrence rate of given site is the same as that of historic period. (b) Seismicity of the region can be expressed by tectonic earthquakes in the region, i.e., the seismic activities distribute homogeneously in a certain tectonic area or an active fault.

The two assumptions accord with the two principles of historic earthquake repeatness and tectonic extrapolation. The difference is that the PSHA adds a new concept of magnitude interval recurrent rate and the hazard is evaluated with probabilistic analysis method. The method can provide the exceedence probability of different ground motion extent (intensity, acceleration, etc.) at the site in specific time intervals, so that the earthquake resistant parameters can be selected with different exposure period, risk level, and various engineering structures.

Example: seismic zonation map of china (2001)

The first seismic zonation map of China was compiled by Wong (1921) after the Haiyuan earthquake, which occurred in 1920. After that, three versions of seismic zonation map were developed in 1957, 1977, and 1990 (Shi et al., 1992). The 1957 version demonstrated the maximum affected intensity of China. The 1977 version was provided by using the methodology of long-term and middle-term earthquake prediction. This version of seismic zonation map demonstrated the maximum encountered intensity in the forthcoming 100 years. This map was adopted by the building code. The 1990 version used probabilistic method. The seismic intensity with exceeding probability of 10% within 50 years was given in the map. It was used in the building code and other regulations related with seismic design. In 2001, a new seismic zonation map of China was issued. This map also used probabilistic method. The PGA and characteristic



Seismic Zonation, Figure 1 Approaches of probabilistic seismic hazard analysis.

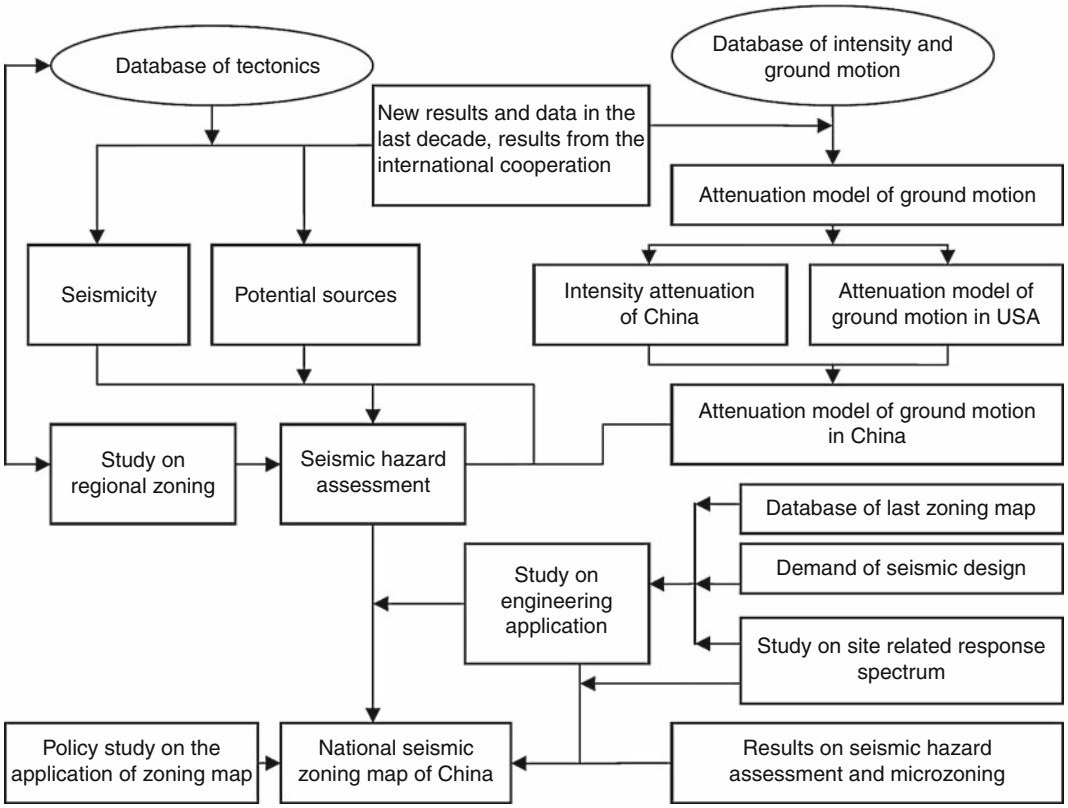
period of response spectrum Tg with exceeding probability of 10% within 50 years were given in the map. The probabilistic method used in developing seismic zonation of China (2001) was a little bit different from PSHA.

Most earthquakes in China are intraplate earthquakes (see *Seismicity, Intraplate*). Their non-homogeneity in space, non-stationarity in time were shown in the historical earthquake catalogs. The PSHA method applied in China can reveal non-homogeneity in space, non-stationarity in time. The approaches applied differ from PSHA in two aspects:

1. Evaluating seismicity and determining magnitude distribution, total annual occurrence rate of the province are on the basis of the seismic tendency estimation and the seismic characteristics analysis.
2. Annual rates of sources in all magnitude intervals are determined by spatial distribution functions, which describe relative risk among sources in the province.

The technical approach developing the national seismic zonation map of China (2001) is shown in Figure 2 (Gao, 2003).

The key scientific problems in compiling this map are: (a) how to treat the uncertainties in the evaluation of seismicity parameters and the delineation of potential sources as well as the attenuation relationship; (b) how to select the suitable parameters in the zonation map to fit the need of seismic design; and (c) how to use the domestic data and the data from the world to get the attenuation of the ground motion parameters.



Seismic Zonation, Figure 2 Technical approach of developing national seismic zonation map of China (2001).

The logic tree method is adopted to treat the uncertainties in seismic hazard analysis. The multi-set of potential source delineations was used. Four groups of scientists participated in the work of potential sources delineations based on the independent background and database. There are four sets of potential sources adopted in the logic tree analysis. Figure 3 is one of the potential source set.

Site-related seismic response spectrum is the base of seismic design for ordinary industry and civil construction. The site-related response spectrum is not only closely related to the earthquake environment, but also related to the soil condition. It is very difficult to decide site-related response spectrum by a single parameter such as intensity or peak acceleration. According to the results form seismic hazard assessment and the building code, the basic parameters used in the seismic zoning map to provide site-related response spectrum are PGA and the characteristic period of the response spectrum.

The attenuation relationships for the acceleration response spectrum platform value Sa and the velocity response spectrum platform value Sv were developed by modifying corresponding attenuation relations in the western United States according to the differences of intensity attenuation relations. Then the PGA and the characteristic period of response spectrum of acceleration Tg were defined as:

$$PGA = \frac{Sa}{2.5}$$
$$Tg = 2\pi \frac{Sv}{Sa} \quad (1)$$

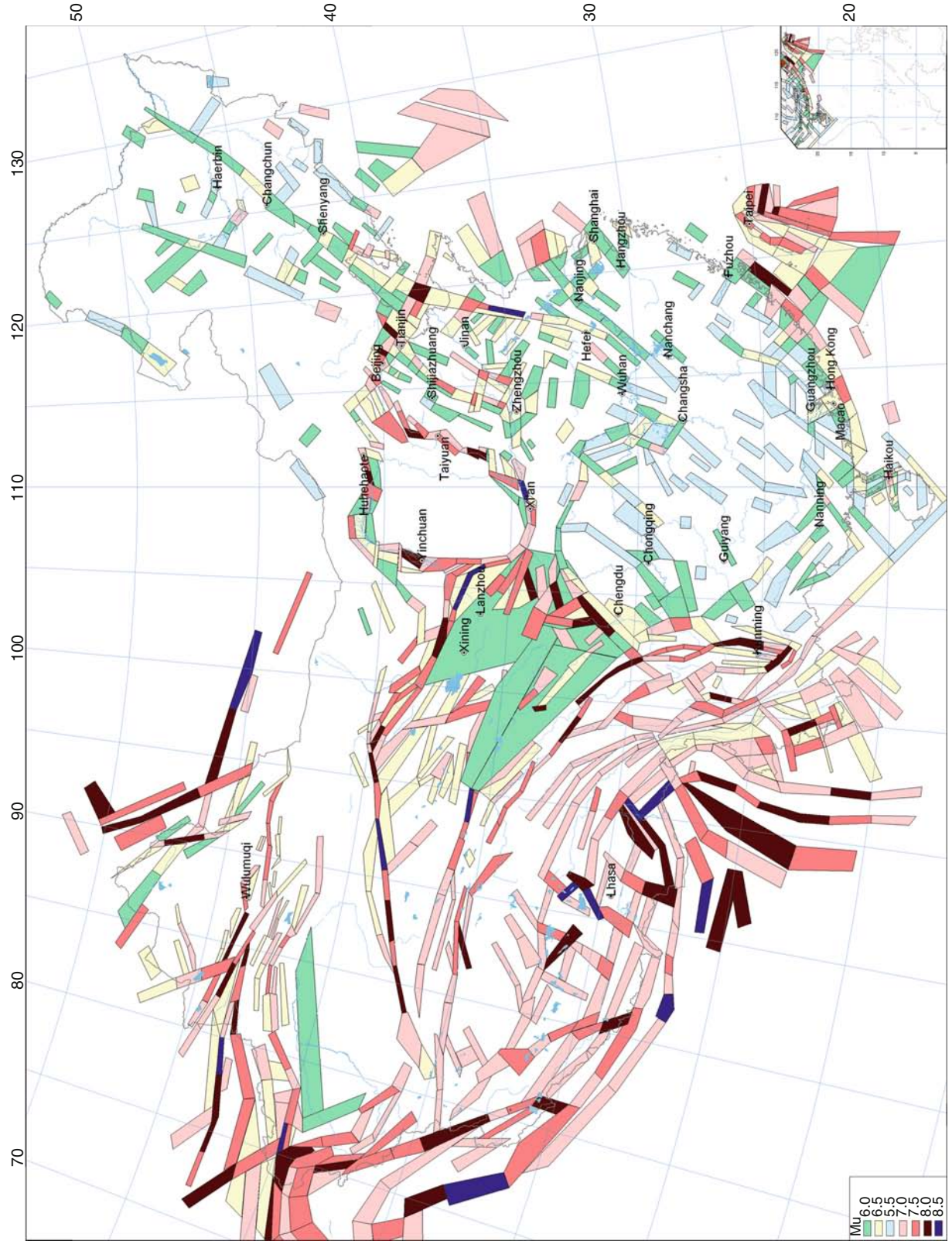
The country was divided into 40,000 grids. The probabilistic seismic hazard analysis for every grid was performed. The ground motion parameters with exceeding probability 10% within 50 years (return period 475 years) were determined.

The new zoning map includes two specific maps. One is PGA in gravity unit g (Figure 4) and the other is Tg in second (Figure 5). The scale of the maps are 1:4,000,000. The soil condition is medium hard soil. From these two parameters, the design response spectrum can be determined easily by (Equation 2):

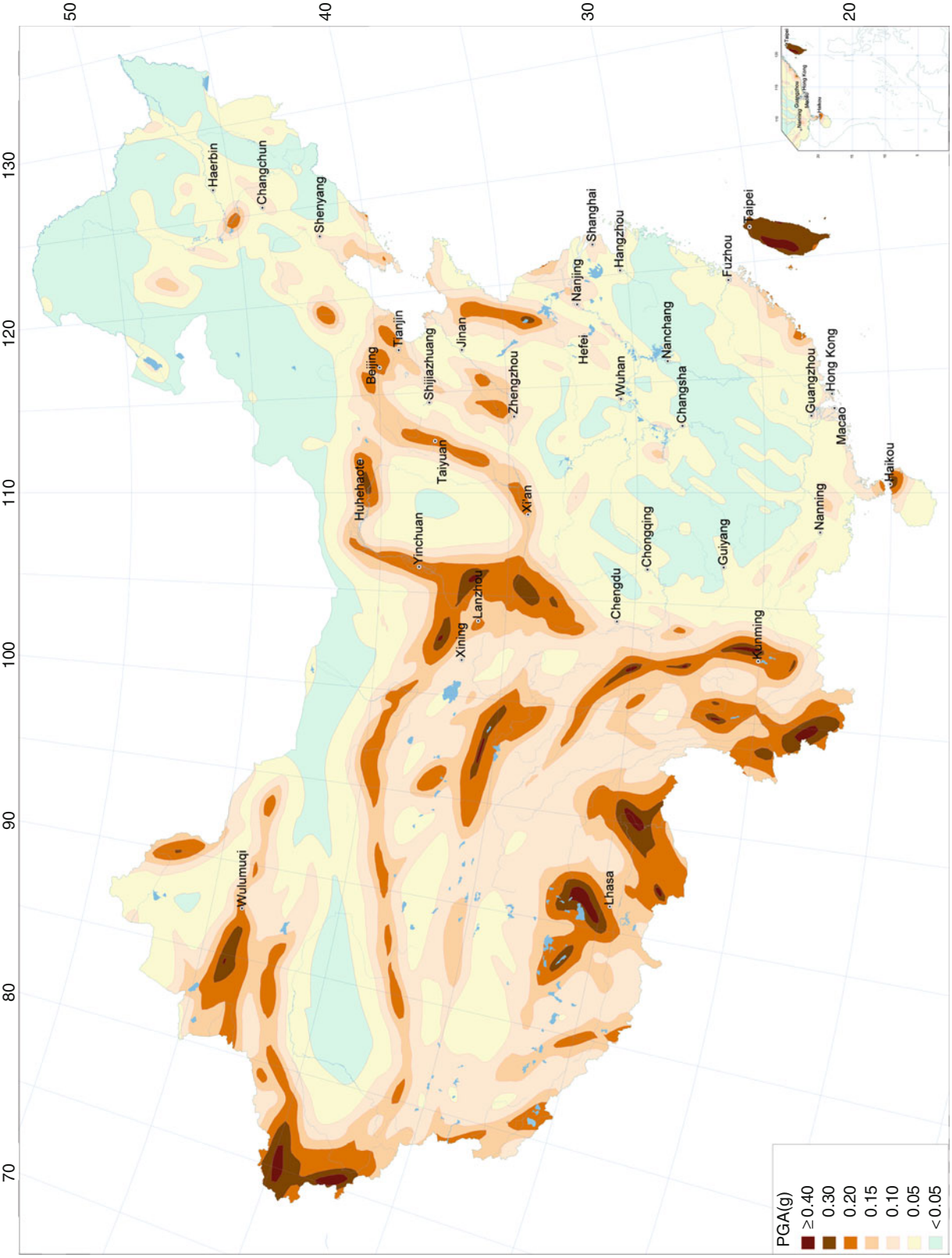
$$Sa(T) = 2.5a \begin{cases} 6T + 0.4 & 0 \le T < 0.1 \\ 1 & 0.1 \le T < Tg \\ Tg/T & T \ge Tg \end{cases} \quad (2)$$

where a is PGA in g, T is the natural period in second.

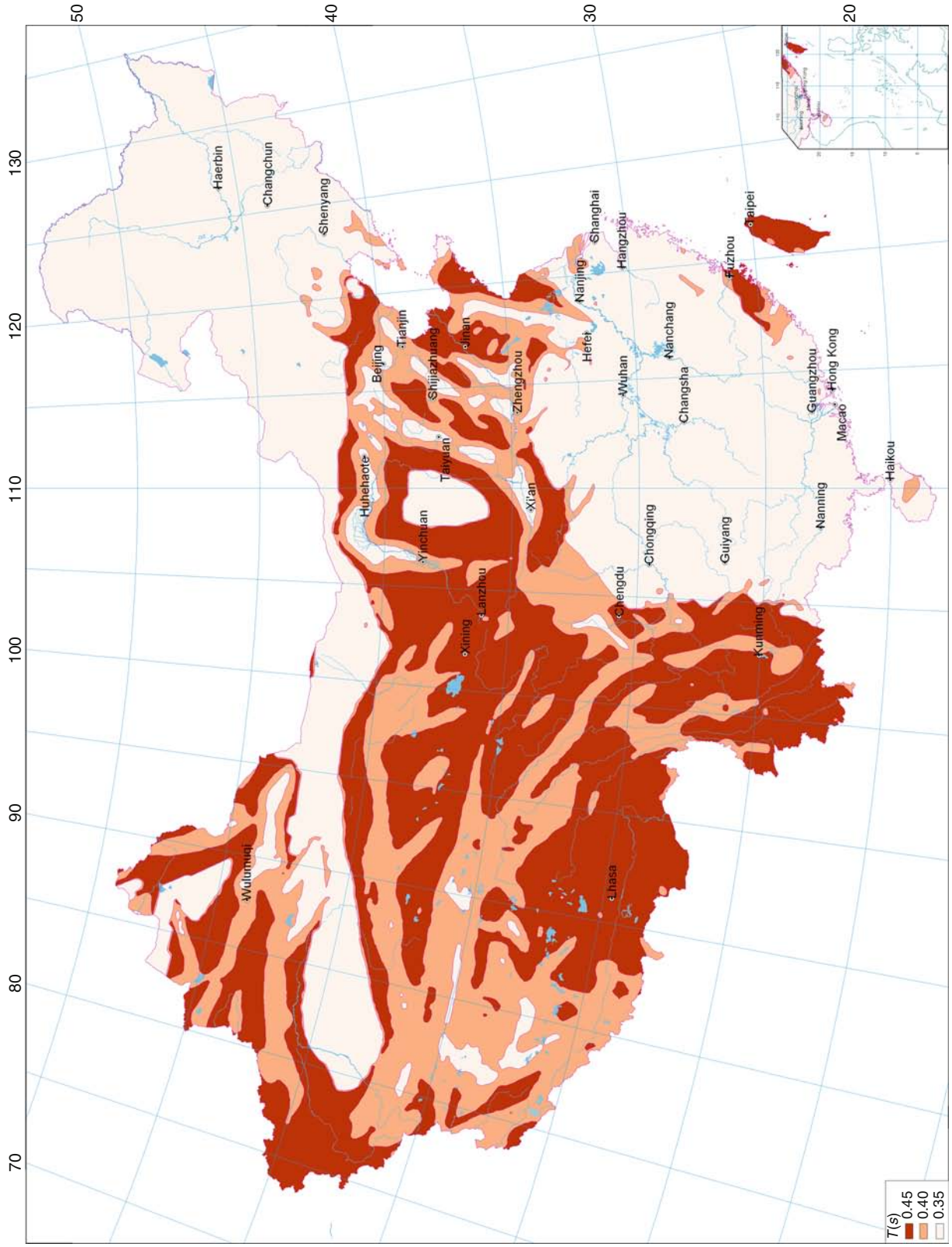
In the acceleration map (Figure 4), the territory is divided into seven zones. The acceleration for the seven zones are 0.05, 0.05, 0.10, 0.15, 0.20, 0.30, and $\geq$0.40 g, respectively.



Seismic Zonation, Figure 3 One of the seismic potential source set used in developing national seismic zonation of China (2001).



Seismic Zonation, Figure 4 Illustration of acceleration zonation map of China.



Seismic Zonation, Figure 5 Illustration of characteristic period of seismic response spectrum zonation map of China.

Seismic Zonation, Table 1 Tg value for different soil conditions (unit: second)

Zone	Soil conditions			
	Hard	Medium hard	Medium soft	Soft
1	0.25	0.35	0.45	0.65
2	0.30	0.40	0.55	0.75
3	0.35	0.45	0.65	0.90

In the characteristic period zoning map (Figure 5), the territory of China is divided into three zones. Zone 1 is the region with Tg = 0.35 s, zone 2 with Tg = 0.40 s, and zone 3 with Tg = 0.45 s.

For different soil conditions, the Tg value should be adjusted as Table 1, but the PGA value does not change with soil conditions.

The new national seismic zonation map of China serves as the obliged state standard, which took into effect in August 1, 2001. For the ordinary new constructed buildings, the standard must be followed.

Summary

Seismic zonation is a process of estimation of the seismic hazard in terms of parameters of ground motion for a certain area. Assessment results in seismic zonation map compilation, which reflects territorial distribution of the seismic hazard (see *Seismic Hazard*). Seismic zonation map is useful for hazard reduction such as earthquake-resistant design of structures, risk analysis, land-use planning, etc. Many countries apply seismic hazard maps in anti-seismic codes. Recently, the probabilistic seismic hazard analysis method is more commonly used in compiling seismic zonation map. The seismic zonataion maps take the exceeding probability of 10% within 50 years (return period 475 years) as standard. Seismic zonation will develop with the development of seismic hazard assessment methodology and anti-seismic policy.

Bibliography

Algermissen, S. T., and Perkins, D. M., 1976. A probabilistic estimate of the maximum acceleration in rock in the contiguous United States: *U.S. Geological Survey Open-File Report* 76–416, 45 pp, 2 pls., scale 1:7,500,000.

Algermissen, S. T., Perkins, D. M., Thenhaus, P. C., Hanson, S. L., and Bender, B. L., 1990. Probabilistic earthquake acceleration and velocity maps for the United States and Puerto Rico: *U.S. Geological Survey Miscellaneous Field Studies Map MF-2120*, 2 sheets, scale 1:7,500,000.

Bender, B. K., and Perkins, D. M., 1982. SEISRISK II: a computer program for seismic hazard estimation. *U.S. Geological Survey Open-File Report,* 82–293.

Cornell, C. A., 1968. Engineering seismic risk analysis. *Bulletin. Seismological Society of America*, **58**, 1583–1606.

Frankel, A. D., Mueller, C., Barnhard, T., Perkins, D., Leyendecker, E., Dickman, N., Hanson, S., and Hopper, M., 1996. National Seismic Hazard Maps – Documentation June 1996: *U.S. Geological Survey Open-File Report,* 96–532.

Frankel, A. D., Petersen, M. D., Mueller, C. S., Haller, K. M., Wheeler, R. L., Leyendecker, E. V., Wesson, R. L., Harmsen, S. C., Cramer, C. H., Perkins, D. M., and Rukstales, K. S., 2002. Documentation for the 2002 update of the National Seismic Hazard Maps: *U.S. Geological Survey Open-File Report* 2002–420.

Gao, M., 2003. New national seismic zoning map of China. *Acta Seismologica Sinica*, **16**, 639–645.

McGuire, R. K., 1978. FRISK: computer program for seismic risk analysis using faults as earthquake sources. *U.S. Geological Survey Open File-Report* 78–1007.

Parvez, I. A., Vaccari, F., and Panza, G. F., 2003. A deterministic seismic hazard map of India and adjacent areas. *Geophysical Journal International*, **155**, 489–508.

Petersen, M. D., Cramer, C. H., Reichle, M. S., Frankel, A. D., and Hanks, T. C., 2008. Discrepancy between earthquake rates implied by historic earthquakes and a consensus geologic source model for California. *Bulletin. Seismological Society of America*, **90**, 1117–1132.

Shi, Z., Yan, J., and Gao, M., 1992. Research on the principle and methodology of seismic zonation – results of the trials in North China. *Acta Seismologica Sinica*, **5**, 305–314.

Wong, W. H., 1921. Zhuizhi Collections. Beijing: The Commercial Press. (in Chinese)

Cross-references

Earthquakes, Intensity
Earthquakes, Strong-Ground Motion
Seismic Hazard
Seismic Microzonation
Seismicity, Intraplate
Seismology, Global Earthquake Model

SEISMIC, AMBIENT NOISE CORRELATION

Michel Campillo[1], Philippe Roux[1], Nikolai M. Shapiro[2]
[1]Observatoire de Grenoble, Université Joseph Fourier and CNRS, Grenoble, France
[2]Institut de Physique du Globe de Paris, Paris, France

Definition

Seismic noise: permanent motion of the Earth surface that is not related to earthquakes or specific controlled sources.

Introduction

Traditional observational methods in seismology are based on earthquake records. It results in two main shortcomings. First, most techniques are based on waves emitted by earthquakes that occurred only in geologically active areas, mainly plate boundaries. This results in a limited resolution in all other areas where earthquakes are not present. Second, the repetition of earthquakes is rare, preventing the study of continuous changes within active structures such as volcanoes or faults.

Also at smaller scales in the context of geophysics prospecting, the resolution is limited by the number and power of sources, making it difficult to image large areas and/or deep structures. Similarly, reproducible sources

are necessary for time-lapse monitoring leading to long-duration surveys that are difficult to achieve.

Nowadays, the seismic networks are producing continuous recordings of the ground motion. These huge amounts of data consist mostly of so called seismic noise, a permanent vibration of the Earth due to natural or industrial sources. Passive seismic tomography is based on the extraction of the coherent contribution to the seismic field from the cross-correlation of seismic noise between station pairs.

As described in many studies where noise has been used to obtain the Green's function between receivers, coherent waves are extracted from noise signals even if, at first sight, this coherent signal appears deeply buried in the local incoherent seismic noise. Recent studies on passive seismic processing have focused on two applications, the noise-extracted Green's functions associated to surface waves leads to subsurface imaging on scales ranging from thousands of kilometers to very short distances; on the other hand, even when the Green's function is not satisfactorily reconstructed from seismic ambient noise, it has been shown that seismic monitoring is feasible using the scattered waves of the noise-correlation function.

Theoretical basis for the interpretation of noise records at two stations

Passive seismology is an alternative way of probing the Earth's interior using noise records only. The main idea is to consider seismic noise as a wave field produced by randomly and homogeneously distributed sources when averaged over long time series. In this particular case, cross-correlation between two stations yields the Green's function between these two points. In the case of a uniform spatial distribution of noise sources, the cross-correlation of noise records converges to the complete Green's function of the medium, including all reflection, scattering, and propagation modes. However, in the case of the Earth, most of ambient seismic noise is generated by atmospheric and oceanic forcing at the surface. Therefore, the surface wave part of the Green's function is most easily extracted from the noise cross-correlations. Note that the surface waves are the largest contribution of the Earth response between two points at the surface.

Historically speaking, helioseismology was the first field where ambient-noise cross-correlation performed from recordings of the Sun's surface random motion was used to retrieve time-distance information on the solar surface. More recently, a seminal paper was published by Weaver and Lobkis (2001) that showed how, at the laboratory scale, diffuse thermal noise recorded and cross-correlated at two transducers fastened to one face of an aluminum sample provided the complete Green's function between these two points. This result was generalized to the case where randomization is not produced by the distribution of sources, but is provided by multiple scattering that takes place in heterogeneous media.

By summing the contributions of all sources to the correlation, it has been shown numerically that the correlation contains the causal and acausal Green's function of the medium. Cases of non-reciprocal (e.g., in the presence of a flow) or inelastic media have also been theoretically investigated. Derode et al. (2003) proposed to interpret the Green's function reconstruction in terms of a time-reversal analogy that makes it clear that the convergence of the noise-correlation function towards the Green's function is bonded to the stationary phase theorem. For the more general problem of elastic waves, one could summarize that the Green's function reconstruction depends on the equipartition condition of the different components of the elastic field. In other words, the emergence of the Green's function is effective after a sufficient self-averaging process that is provided by random spatial distribution of the noise sources when considering long-time series as well as scattering (e.g., Gouédard et al., 2008 and references herein).

Applications in seismology

For the first time, Shapiro and Campillo (2004) reconstructed the surface wave part of the Earth response by correlating seismic noise at stations separated by distances of hundreds to thousands of kilometers, and measured their dispersion curves at periods ranging from 5 s to about 150 s. Then, a first application of passive seismic imaging in California (e.g., Shapiro et al., 2005; Sabra et al., 2005) appeared to provide a much greater spatial accuracy than for usual active techniques. More recently, the feasibility of using the noise cross-correlations to monitor continuous changes within volcanoes and active faults was demonstrated (e.g., Brenguier, 2008a, b). These results demonstrated a great potential of using seismic noise to study the Earth interior at different scales in space and time. At the same time, the feasibility of both noise-based seismic imaging and monitoring in every particular case depends on spatio-temporal properties of the available noise wavefield. Therefore, a logical initial step for most of noise-based studies is to characterize the distribution of noise sources. Also, in many cases, knowledge of the distribution of the noise sources can bring very important information about the coupling between the Solid Earth with the Ocean and the Atmosphere. So far, we can identify three main types of existing seismological applications related to noise correlations: (1) studies of spatio-temporal distribution of seismic noise sources, (2) noise-based seismic imaging, and (3) noise-based seismic monitoring.

Noise source origin and distribution

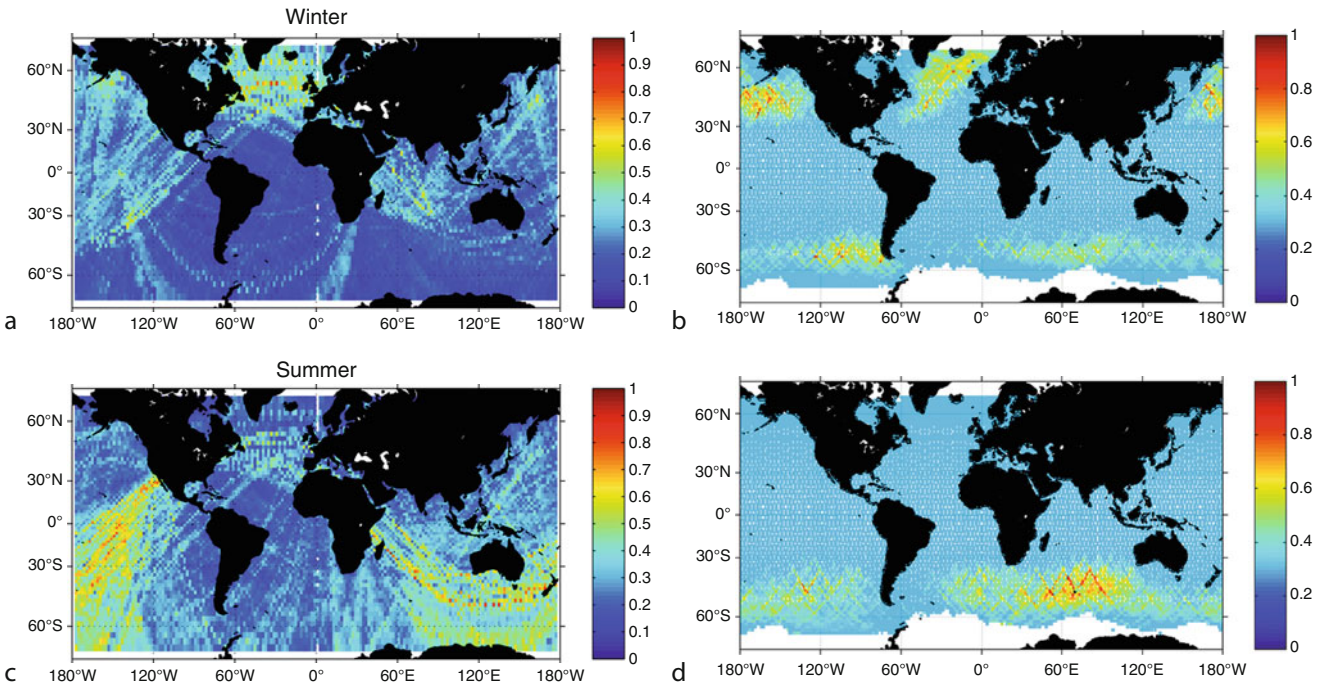
Distribution of noise sources strongly depends on the spectral range under consideration. At high frequencies (> 1 Hz), the noise is strongly dominated by local sources that may have very different origins and are often anthropogenic. At these scales, the properties of the noise wavefield should be studied separately for every particular case and no reasonable generalization can be done. At

longer periods, noise is dominated by natural sources. In particular, it is well established that two main peaks in the seismic noise spectra in so-called microseismic band (1–20 s) are related to forcing from oceanic gravity waves. It has been also argued that at periods longer than 20 s, the oceanic gravity and infragravity waves play a major role in the seismic noise excitation. The interaction between these oceanic waves and the solid Earth is governed by a complex non-linear mechanism (Longuet-Higgins, 1950) and, as a result, the noise excitation depends on many factors such as the intensity of the oceanic waves but also the intensity of their interferences as well as the seafloor topography (e.g., Kedar et al., 2008). Overall, the generation of seismic noise is expected to be strongly modulated by strong oceanic storms and, therefore, to have a clear seasonal and non-random pattern.

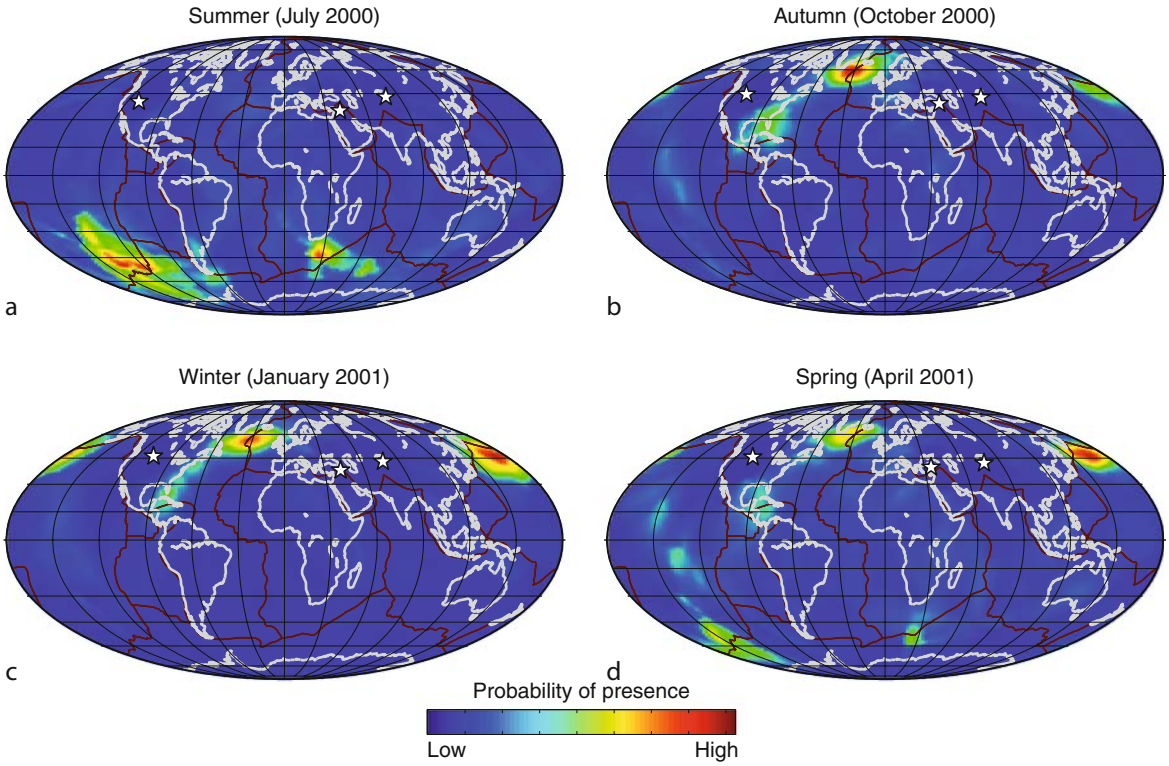
Seismic noise in the microseismic spectral band is dominated by fundamental mode surface waves. It is currently debated whether the surface wave component of microseisms is generated primarily along coastlines or if it is also generated in deep-sea areas. Inhomogeneous distribution and seasonality of microseismic noise sources is clearly revealed by the amplitude of the Rayleigh wave reconstructed in noise cross-correlations (e.g., Stehly et al., 2006) as shown in Figure 1. At the same time, body waves were detected in the secondary microseismic band and can be sometimes associated with specific storms. Figure 2 shows that sources of microseismic P waves are located in specific areas in deep ocean and exhibit strong seasonality as determined from the analysis of records by dense seismic networks (Landes et al., 2010).

Noise-based seismic imaging

Numerous studies has demonstrated that, when considered over sufficiently long times, the noise sources become sufficiently well distributed over the Earth's surface and that dispersion curves of fundamental mode surface waves can be reliably measured from correlations of seismic noise at periods between 5 and 50 s for most of interstation directions. This led to the fast development during recent years of the ambient-noise surface wave tomography. It consists of computing cross-correlations between vertical and horizontal components for all available station pairs followed by measuring group and phase velocity dispersion curves of Rayleigh and Love waves (e.g., Bensen et al., 2007). This dispersion curves are then regionalized (e.g., Lin et al., 2009) and inverted to obtain three-dimensional distribution of shear velocities in the crust and the uppermost mantle. After first results obtained in southern California (Shapiro et al., 2005; Sabra et al., 2005), this method has been applied with many regional seismological networks (e.g., Yao et al., 2006; Lin et al., 2007; Yang et al., 2008a). At smaller scales, it can be used



Seismic, Ambient Noise Correlation, Figure 1 Comparison between seasonal variations of the location of seismic noise sources and significant wave height. (**a**) and (**c**) Geographical distribution of the apparent source of the Rayleigh waves detected in the 10–20 s noise cross correlations during the winter and the summer, respectively. (**b**) and (**d**) Global distribution of the square of wave height measured by TOPEX/Poseidon during the winter and the summer, respectively (From Stehly et al., 2006).

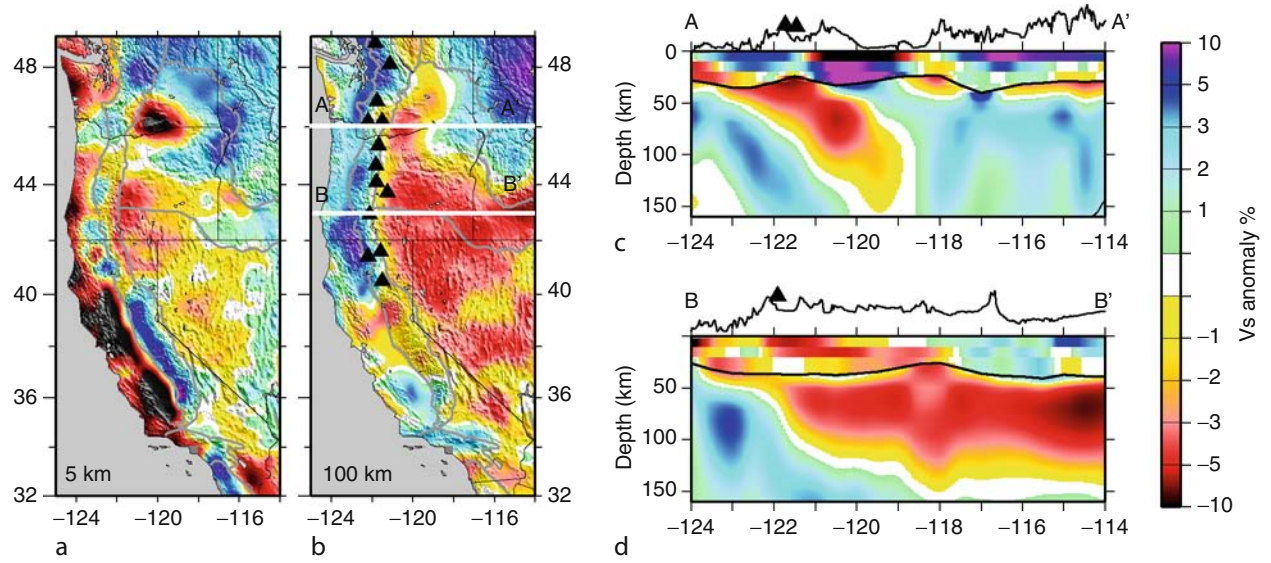


Seismic, Ambient Noise Correlation, Figure 2 Seasonal variation of the location of P-wave seismic noise sources in the secondary microseismic band (0.1–0.3 Hz) determined from the analysis of records at the three seismic networks indicated with white stars (From Landes et al., 2010).

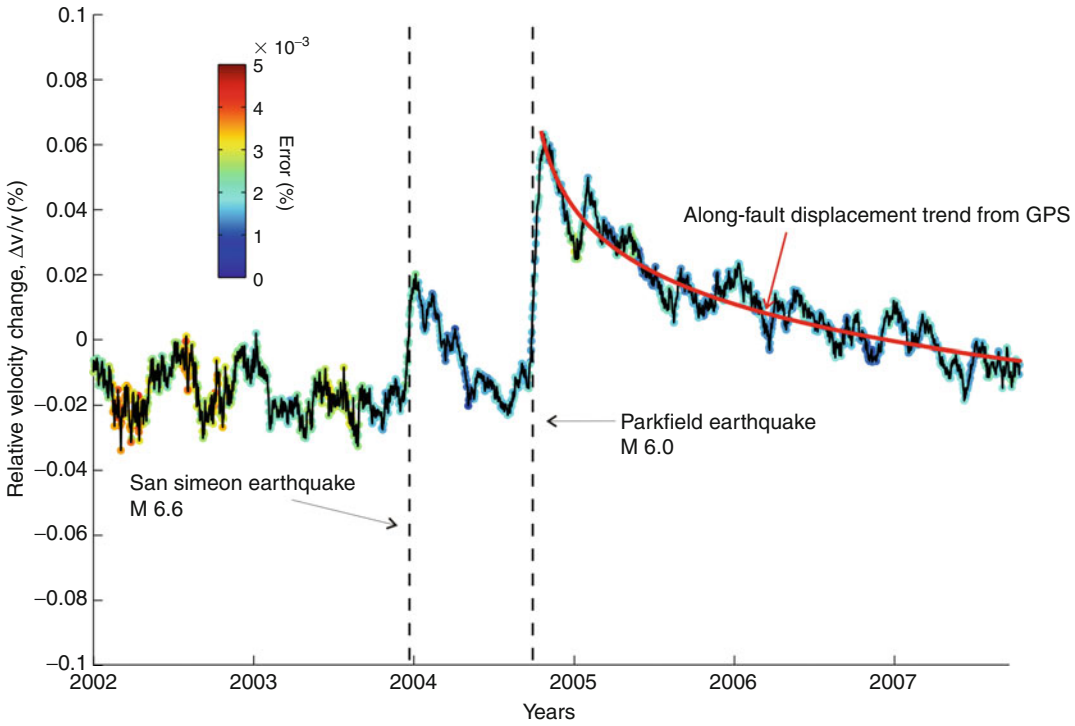
to study shallow parts of volcanic complexes (e.g., Brenguier et al., 2007). The ambient-noise surface wave tomography is especially advantageous in context of dense continent-scale broadband seismic networks such as available in USA (e.g., Moschetti et al., 2007; Yang et al., 2008b) and Europe (e.g., Stehly et al., 2009). At these scales, noise-based imaging can be used to obtain high-resolution information about the crustal and the upper mantle structure including seismic anisotropy (e.g., Moschetti et al., 2010) and can be easily combined with earthquake-based measurements to extend the resolution to larger depths (e.g., Yang et al., 2008b). An example of results obtained from combined noise and earthquakes based surface wave tomography in western USA is shown in Figure 3.

Noise-based monitoring

One of the advantages of using continuous noise records to characterize the earth materials is that a measurement can easily be repeated. This led recently to the idea of a continuous monitoring of the crust based on the measurements of wave speed variations. The principle is to apply a differential measurement to correlation functions, considered as virtual seismograms. The technique developed for repeated earthquakes (doublets), proposed by Poupinet et al., 1984, can be used with correlation functions. In a seismogram, or a correlation function, the delay accumulates linearly with the lapse time when the medium undergoes a homogeneous wave speed change, and a slight change can be detected more easily when considering late arrivals. It was therefore reasonable, and often necessary, to use coda waves for the measurements of temporal changes. Noise-based monitoring relies on the autocorrelation or cross-correlation of seismic noise records (Sens-Schönfelder and Wegler, 2006; Brenguier et al., 2008a, b). When data from a network are available, using cross-correlation take advantage of the number of pairs with respect to the number of stations. It is worth noting that the use of the coda of the correlation functions is also justified by the fact that its sensitivity to changes in the origin of the seismic noise is much smaller than the sensitivity of the direct waves. Several authors noted that an anisotropic distribution of sources leads to small errors in the arrival time of the direct waves, which can be evaluated quantitatively (e.g., Weaver et al., 2009). While in most of the cases, they are acceptable for imaging, they can be larger than the level of precision required when investigating temporal changes. The issue of the nature of the tail (coda) of the cross-correlation function is therefore fundamental and was analyzed by Stehly et al. (2008). These authors showed that it contains at least partially the coda of the Green function, i.e., physical arrivals which



Seismic, Ambient Noise Correlation, Figure 3 Shear-velocity structure of the crust and the upper mantle obtained from the inversion of the USArray data. (**a**) and (**b**) Horizontal cross-sections at depths of 5 and 100 km. (**c**) and (**d**) Vertical cross-sections along profiles delineated by the white lines in (**b**). Black lines outline the Moho. Topography is superimposed above individual cross sections. The black triangles represent active volcanoes in the Cascade Range (From Yang et al., 2008b).



Seismic, Ambient Noise Correlation, Figure 4 Relative seismic velocity change during 6 years measured from continuous noise correlations in Parkfield. The dashed lines indicated two major earthquakes: the San Simeon event that occurred 80 km from Parkfield and the local Parkfield event (Modified from Brenguier et al., 2008b).

kinematics is controlled by the wave speeds of the medium. It can therefore be used for monitoring temporal changes. As an illustration of the capability of this approach, we present in Figure 4 a measure of the average wave speed change during a period of 6 years in the region of Parkfield, California. Two main events occurred in this region during the period of study: the 2003 San Simeon and 2004 Parkfield earthquakes. In both cases, noise-based monitoring indicates a co-seismic speed drop. The measured relative variations of velocity before de San Simeon earthquake are as small as 10^{-4}. The changes of velocity associated with earthquakes are associated with at least two different physical mechanisms: (1) the damage induced by the strong ground motions in shallow layers and fault zone, as illustrated by the co-seismic effect of the distant San Simeon event, and (2) co-seismic bulk stress change followed by the post-seismic relaxation, as shown with the long-term evolution after the local Parkfield event, similar in shape to the deformation measured with GPS.

Summary

Continuous recordings of the Earth surface motion by modern seismological networks contain a wealth of information on the structure of the planet and on its temporal evolution. Recent developments shown here make it possible to image the lithosphere with noise only and to detect temporal changes related to inner deformations.

Bibliography

Bensen, G. D., Ritzwoller, M. H., Barmin, M. P., Levshin, A. L., Lin, F., Moschetti, M. P., Shapiro, N. M., and Yang, Y., 2007. Processing seismic ambient noise data to obtain reliable broad-band surface wave dispersion measurements. *Geophysical Journal International*, **169**, 1239–1260, doi:10.1111/j.1365-246X.2007.03374.x, 2007.

Brenguier, F., Shapiro, N. M., Campillo, M., Nercessian, A., and Ferrazzini, V., 2007. 3-D surface wave tomography of the Piton de la Fournaise volcano using seismic noise correlations. *Geophysical Research Letters*, **34**, L02305, doi:10.1029/2006GL028586.

Brenguier, F., Shapiro, N., Campillo, M., Ferrazzini, V., Duputel, Z., Coutant, O., and Nercessian, A., 2008a. Toward forecasting volcanic eruptions using seismic noise. *Nature Geoscience*, **1**(2), 126–130.

Brenguier, F., Campillo, M., Hadziioannou, C., Shapiro, N. M., Nadeau, R. M., and Larose, E., 2008b. Postseismic relaxation along the San Andreas fault in the Parkfield area investigated with continuous seismological observations. *Science*, **321**(5895), 1478–1481.

Derode, A., Larose, E., Tanter, M., de Rosny, J., Tourin, A., Campillo, M., and Fink, M., 2003. Recovering the Green's function from field-field correlations in an open scattering medium. *The Journal of the Acoustical Society of America*, **113**, 2973–2976.

Gouédard, P., Stehly, L., Brenguier, F., Campillo, M., de Verdière Colin, Y., Larose, E., Margerin, L., Roux, P., Sanchez-Sesma, F. J., Shapiro, N. M., and Weaver, R. L., 2008. Cross-correlation of random fields: mathematical approach and applications. *Geophysical Prospecting*, **56**, 375–393.

Kedar, S., Longuet-Higgins, M., Webb, F., Graham, N., Clayton, R., and Jones, C., 2008. The origin of deep ocean microseisms in the North Atlantic Ocean. *Royal Society of London Proceedings Series A*, **464**, 777–793, doi:10.1098/rspa.2007.0277.

Landes, M., Hubans, F., Shapiro, N. M., Paul, A., and Campillo, M., 2010. Origin of deep ocean microseims by using teleseismic body waves. *Journal of Geophysical Research*, doi:10.1029/2009JB006918.

Lin, F., Ritzwoller, M. H., Townend, J., Savage, M., and Bannister, S., 2007. Ambient noise Rayleigh wave tomography of New Zealand. *Geophysical Journal International*, doi:10.1111/j.1365-246X.2007.03414.x.

Lin, F.-C., Ritzwoller, M. H., and Snieder, R., 2009. Eikonal tomography: surface wave tomography by phase-front tracking across a regional broad-band seismic array. *Geophysical Journal International*, **177**(3), 1091–1110.

Longuet-Higgins, M. S., 1950. A theory of the origin of microseisms. *Philosophical Transactions of the Royal Society of London Series A*, **243**, 1–35.

Moschetti, M. P., Ritzwoller, M. H., and Shapiro, N. M., 2007. Surface wave tomography of the western United States from ambient seismic noise: Rayleigh wave group velocity maps. *Geochemistry, Geophysics, Geosystems*, **8**, Q08010, doi:10.1029/2007GC001655.

Moschetti, M. P., Ritzwoller, M. H., and Lin, F. C., 2010. Seismic evidence for widespread crustal deformation caused by extension in the western USA. *Nature*, **464**, 885–889, doi:10.1038/nature08951.

Poupinet, G., Ellsworth, W. L., and Frechet, J., 1984. Monitoring velocity variations in the crust using earthquake doublets: an application to the Calaveras Fault, California. *Journal of Geophysical Research*, **89**, 5719–5731.

Sabra, K. G., Gerstoft, P., Roux, P., Kuperman, W. A., and Fehler, M. C., 2005. Extracting time domain Green's function estimates from ambient seismic noise. *Geophysical Research Letters*, **32**, L03310.

Sens-Schönfelder, C., and Wegler, U., 2006. Passive image interferometry and seasonal variations of seismic velocities at Merapi Volcano, Indonesia. *Geophysical Research Letters*, **33**, L21302.

Shapiro, N. M., and Campillo, M., 2004. Emergence of broadband Rayleigh waves from correlations of the ambient seismic noise. *Geophysical Research Letters*, **31**, L07614, doi:10.1029/2004GL019491.

Shapiro, N. M., Campillo, M., Stehly, L., and Ritzwoller, M., 2005. High resolution surface wave tomography from ambient seismic noise. *Science*, **307**, 1615–1618.

Stehly, L., Campillo, M., and Shapiro, N., 2006. A Study of the seismic noise from its long range correlation properties. *Journal of Geophysical research*, **111**, B10306.

Stehly, L., Campillo, M., Froment, B., and Weaver, R. L., 2008. Reconstructing Green's function by correlation of the coda of the correlation (C3) of ambient seismic noise. *Journal of Geophysical Research*, **113**, B11306.

Stehly, L., Fry, B., Campillo, M., Shapiro, N. M., Guilbert, J., Boschi, L., and Giardini, D., 2009. Tomography of the Alpine region from observations of seismic ambient noise. *Geophysical Journal International*, **178**, 338–350.

Weaver, R. L., and Lobkis, O. I., 2001. Ultrasonics without a source: thermal fluctuation correlations at MHz frequencies. *Physical Review Letters*, **87**(13), 134301, doi:10.1103/PhysRevLett.87.134301.

Weaver, R. L., Froment, B., and Campillo, M., 2009. On the correlation of non-isotropically distributed ballistic scalar diffuse waves. *Journal of the Acoustical Society of America*, 1817–1826.

Yang, Y., Li, A., and Ritzwoller, M. H., 2008a. Crustal and uppermost mantle structure in southern Africa revealed from ambient

noise and teleseismic tomography. *Geophysical Journal International*, doi:10.1111/j.1365-246X.2008.03779.x.
Yang, Y., Ritzwoller, M. H., Lin, F.-C., Moschetti, M. P., and Shapiro, N. M., 2008b. The structure of the crust and uppermost mantle beneath the western US revealed by ambient noise and earthquake tomography. *Journal of Geophysical Research*, **113**, B12310, doi:10.1029/2008JB005833.
Yao, H., van der Hilst, R. D., and de Hoop, M. V., 2006. Surface-wave array tomography in SE Tibet from ambiant seismic noise and two-station analysis – I. Phase velocity maps. *Geophysical Journal International*, **166**, 732–744.

Cross-references

Body Waves
Earthquakes and Crustal Deformation
Seismic Noise
Seismic Tomography
Seismic Waves, Scattering
Surface Waves

SEISMIC, MIGRATION

Samuel H. Gray
CGGVeritas, Calgary, AB, Canada

Synonyms

Seismic imaging

Definition

Seismic migration. A set of techniques for transforming recorded (elastic-wave) seismic reflection data into an image of reflecting boundaries in the earth's interior. In simplest form, these are intended to correct certain distortions present in recorded wavefields. The distortions can be caused by: diffractors inside the earth, which scatter the incident energy to a range of receiver locations; geometric effects caused by dipping reflectors; and velocity effects, which cause the seismic waves to change direction as they propagate from source to reflector to receiver. Migration is also used to estimate seismic velocity and to provide amplitudes for rock property analysis.

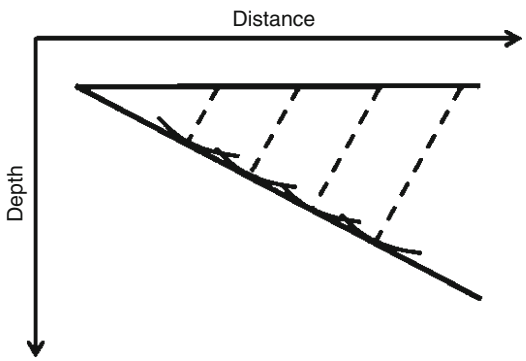
The migration heuristic assumes that subsurface reflectors are made up of point diffractors (Trorey, 1970). The response of a reflector to input seismic energy is the sum of responses to all the diffractors, and migration transforms this response to the actual reflector shape by collapsing each diffractor response to a point (diffraction stack).

History

Mapping using pencil and paper; mechanical migration

The first migrations were performed in the 1920s as a manual operation. A single-fold reflection profile was acquired with sources and receivers on the surface of the earth, and two-way travel times for events from a shallow dipping reflector were picked on the recorded traces. The extent of the lateral offset between source and receiver locations was neglected, which is equivalent to assuming that sources and receivers were colocated, or at zero offset. Then circular arc segments were drawn, centered at the source-receiver midpoints and with radii equal to one-half the product of the picked travel times and the velocity of propagation in the overburden. The envelope of these mapped arcs gave a rough picture of the reflecting interface (Figure 1).

This technique reveals principles and problems of migration. First, reflection events move sideways, or *migrate*, from their picked time locations on the recording surface (the source-receiver midpoints) to subsurface locations at the reflection interface. When reflectors in the earth are not flat, the lateral position of a subsurface reflector is different from that of the recording location of the same event. Second, the assumption of colocated sources and receivers is problematic. As the use of seismic data increased, reflection profiles were typically acquired using higher fold, or multiplicity: for each source, a spread of many receivers recorded reflection data in order to provide a greater multiplicity of observations. Some of the receivers were a considerable distance from the source locations, violating the zero-offset assumption. Later, the common-midpoint (CMP) stack (Mayne, 1962) performed approximate corrections intended to shift each event on a nonzero-offset recorded trace to its zero-offset time, once again allowing the use of the zero-offset assumption in migration (poststack migration). In areas of complicated geologic structure, however, the time corrections used in the CMP stacking process are not accurate, requiring that each recorded trace be migrated using its actual source and receiver location (prestack migration). Third, if more than one reflector exists inside the earth, the various migrated events can interfere with one another as the circular arcs that build up their envelopes overlap. This problem is solved by the more recent application of wave theory to migration. This allows the



Seismic, Migration, Figure 1 Schematic illustration of early pencil-and-paper migration. Assuming constant velocity and zero-offset source and receiver, an event on a seismic trace is swung out along an arc of a circle. Repeating this for the corresponding event on all traces produces an image of a reflector, which lies along the envelope of all the arcs.

rigorous migration of waveforms, with energy buildup near reflector locations, and energy cancellation away from reflector locations. Fourth, if the velocity used in the migration process is incorrect, the image of a subsurface reflector will be located incorrectly. This fundamental velocity problem persists today.

The process of swinging reflection events through arcs to construct approximate reflector segments spawned a number of mechanical devices that migrated reflection records. They allowed the migration of picked events from nonzero-offset traces. However, they were all based on the same principle of swinging events over a restricted range of angles. This process was formalized by Hagedoorn (1954).

Digital migration

Seismic recording, processing, and migration became digital in the 1960s. Digital processing emphasized the discrete sampling of signals, in both time and space. It also allowed reflection records to be treated as wavefields, and wavefield algorithms could be applied to them. Wave-equation migration methods arose, based on differential equations, integral solutions to the wave equation, or mapping in the Fourier domain (Berkhout, 1982; Yilmaz, 1987). These methods are in use today. They have different realms of validity and different efficiencies. One of the methods, integral (Kirchhoff) migration, deserves mention. This technique is the direct descendant of pencil-and-paper migration. In Kirchhoff migration, samples on the recorded input traces are swung out over a range of angles. The earlier techniques could consider only isolated events, because constructing the envelopes corresponding to all reflection events of a record would produce an uninterpretable mess. However, including the wave equation as an essential part of the process allowed all samples to be swung out. Constructive and destructive interference, a natural consequence of treating the samples as components of a wavefield, caused reflection events to appear on the image, not as envelopes, but as waveforms moved to reflector locations.

Poststack migration

The first digital migrations were performed on computers with limited memory and processing capability. Migrating all traces of a two-dimensional (2-D) reflection survey involving hundreds of source locations, each with tens of receiver locations, would have taken months on early computers. So it became necessary to make use of the intermediate process of CMP stack to reduce the number of traces to be migrated. As mentioned above, this process performs time ("normal moveout," or NMO) corrections to traces with nonzero source-receiver offset. These corrections are time shifts, different for each sample, intended to produce equivalent zero-offset traces to be summed, or stacked, together. The quality of the normal-moveout process depends on the stacking velocity, which determines the time shift to be applied to each sample of each nonzero-offset trace. In order to stack together traces in a domain where the velocity varies as little as possible, it is desirable to sort the input traces from their original recorded common-shot configuration to a set of records where each record has a common source-receiver midpoint – thus the term common-midpoint stack. The stacking velocity for a common-midpoint gather is taken to be the velocity at the midpoint location. This velocity can vary with time down the zero-offset trace of the record. For each value of time, it is a type of average velocity. When all the traces in a CMP gather have been NMO-corrected and summed (stacked) together, the CMP gather has been replaced with a single, zero-offset, trace. The NMO/stack process is illustrated on a single reflection event in Figure 2. The ensemble of zero-offset traces, from all CMP gathers indexed by their midpoint locations, is a stacked record ready to be (poststack) migrated.

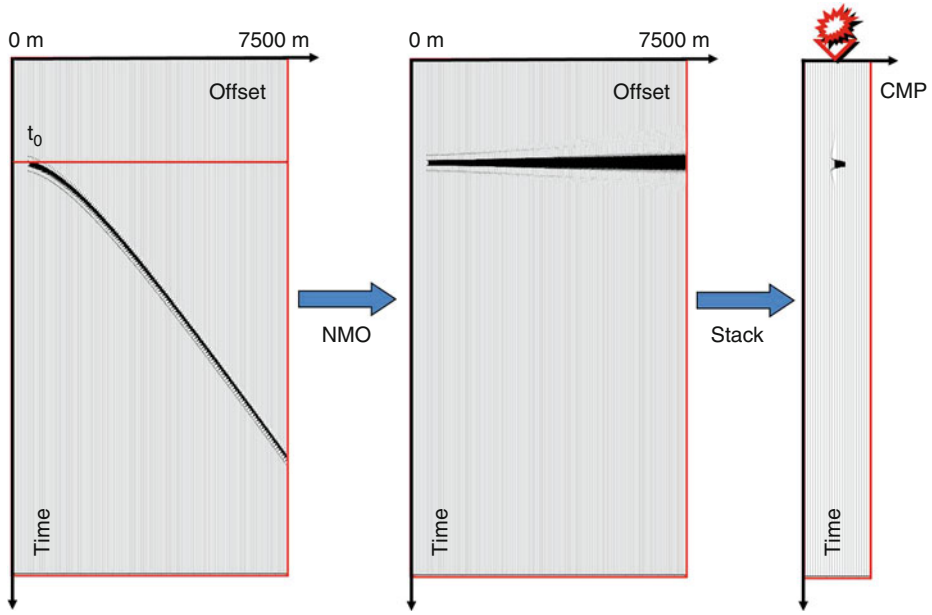
The CMP stacking process makes several approximations that are incompatible with the physics of wave propagation inside the earth. For example, sorting into CMP records and assuming that the seismic velocity does not change for all traces within a record ignores the lateral variability of the actual velocity, and it assumes that the reflectors inside the earth are locally flat below each midpoint location. These approximations were not usually damaging in earlier times, when drilling prospects were in shallow, relatively simple geology. As shallow targets were drilled, leaving deeper prospects under complex overburden with complex velocity behavior, the approximations broke down. Since about 2000, migration performed without CMP stacking (prestack migration) has become commonplace.

Prestack migration

Dropping the NMO/stack process from the migration flow eliminated some problems, but it made the velocity estimation problem more complicated. NMO velocities were useful for migration as well as stacking. These velocities were not necessarily accurate, but they provided a guide for the velocities to be used in migration. Using a rough estimate of velocity, poststack migration generates plausible migrated images, even if they lack the resolution required for interpretation of subtle targets. Accurate prestack migration, on the other hand, requires the velocity to be estimated without a prior stacking velocity analysis. New tools, consistent with wave propagation inside a geologically complicated earth, were needed for this task. These tools have been under development since the 1980s, but are not yet completely satisfactory. Unlike NMO/stack, they allow arbitrary velocity structure inside the earth, and there is no "locally flat" assumption.

2-D, 3-D, wide-azimuth 3-D acquisition and migration

Early seismic reflection profiles were acquired along 2-D lines, both land and marine. On land, a line of receivers was laid out, and shots were fired into the earth at locations along the line. As the shot locations moved along the line,



Seismic, Migration, Figure 2 NMO and CMP stack illustrated on a single reflection event.

receivers were picked up from the back end of the spread and placed at the front. Marine data were acquired using a vessel towing a linear array of receivers (a streamer), firing at regular locations along the line. 2-D processing assumes that the subsurface of the earth is invariant in the direction perpendicular to the line direction. When this assumption is violated, "out-of-plane" reflectors (not vertically below the acquisition line) contribute energy to the reflection records; 2-D migration cannot map this energy to its correct locations. For this reason, 3-D acquisition became commonplace in the 1970s and 1980s. On land, the receivers were laid out in a 3-D pattern (patch) that could be repeated as in 2-D. Gradually, the receiver patch became larger (thousands of meters on a side), with a wider range of azimuths, or compass directions of vectors joining the source and receiver locations. Today, the recording patch often contains several thousand receivers. Wide-azimuth marine acquisition is more difficult than land because it is physically impossible for a single vessel to tow the large number of streamers required. However, adding a second vessel as a source vessel alongside the first vessel at a large lateral distance produced data over a wide range of azimuths as for land.

Moving from 2-D to 3-D enabled the accurate migration of geologically complex structures such as salt bodies and overthrust structures, and it altered the economics of migration. In the initial days of 3-D seismic data, 2-D prestack migration was available but not commonly applied because it was still too computationally expensive. A fortiori, 3-D prestack migration was not yet feasible. In 3-D, the numbers of input and output locations are increased by an order of magnitude over 2-D, increasing migration cost by two orders of magnitude.

Deep crustal imaging

Most migration projects are performed by the oil and gas industry for hydrocarbon exploration, whose target depths rarely exceed 10 km. However, solid-earth geophysicists acquire and migrate seismic data to map crustal structures tens of kilometers deep. Imaging very deep targets with cumulative overburden effects such as wavefield scattering and absorption presents added challenges, and migration cannot be expected to be successful as often as it is for hydrocarbon exploration (Warner, 1987).

Time migration, depth migration

Seismic traces are recorded in time, with a particular event occurring at different times on traces with different lateral positions. The CMP stack produces a composite section, with a trace at each position. Although migration is used to image reflectors inside the earth, the CMP stack itself provides a crude image. With its collapse of many seismic records into one, and its use of redundancy (fold) to suppress noise, the CMP stack produces images that can sometimes be interpreted without migration. If the CMP stack is migrated with the depth axis on the migrated image replaced by time, events on the unmigrated stack and the migrated stack can be compared directly, helping with interpretation. Such a migration, called "time migration," is opposed to the more natural "depth migration."

This fact has caused a diversity of approaches to migration. Geophysical interpreters often prefer time migration, but structural geologists and engineers usually prefer depth migration. Structural geology deals with space (depth) and how space is filled with material; geologists and engineers are concerned with the physical dimensions

of a potential hydrocarbon trap. Even if a migrated image is interpreted in time, in principle it is preferable to perform the migration and display the final image in depth. In practice, however, there are valid arguments against this statement.

First, time migration is sometimes more robust than depth migration; that is, it is often easier to produce an interpretable image with time migration than with depth migration, even if the image is known not to be accurate. This is because time and depth migration treat velocity differently. Time migration, following procedures developed for NMO/stack, uses what is best called an imaging velocity field, i.e., one that best focuses the migrated image. At each image location, this velocity is an average of the seismic velocity values in the overburden above the image depth. The imaging velocity is free to vary spatially. In essence, time migration performs a constant-velocity migration at each image location, where the constant can change from point to point. The imaging velocity field that produces the best image need not be consistent with any possible field of actual velocities inside the earth, and any attempt to convert the imaging velocities to a field of geologically possible earth velocities can fail. For example, assuming that the imaging velocities are root-mean-square velocities and using the Dix (1955) equation to invert these to interval velocities often produces physically impossible values. This inconsistent handling of velocity allows time migration to be robust. Second, the precision implied by depth migration is not always justified. A migrated image purports to show locations of reflectors and diffractors inside the earth, but often the locations are incorrect. It is easy to see that this can happen for time migration, which is not precise because of its inconsistent handling of velocity. For depth migration, the implied accurate positioning of reflectors assumes a correct migration velocity, which is an interval velocity field, i.e., a field of local velocities. Using accurate interval velocities allows depth migration to image more accurately than time migration can, but using inaccurate interval velocities can degrade image coherence that the less precise velocities of time migration often preserve.

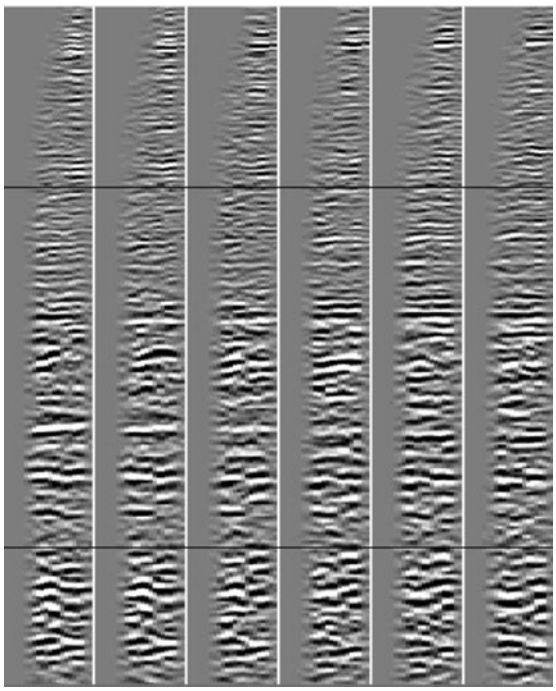
Purposes of migration

Strucural imaging

The primary goal of seismic migration is to correct distortions present in recorded seismic records, due to uncollapsed diffraction energy, geometric effects, or velocity effects. Even in areas of fairly flat geology, diffracted energy from abrupt fault terminations can obscure deeper reflection events, making their interpretation difficult. Therefore, collapsing unwanted diffraction energy to scatterer locations is important. Geometric and velocity effects are caused by complicated geologic structures. Familiar examples are "bowtie" signatures on stacked records converted by migration into synclines, and broad anticlinal structures converted by migration into narrower, steeper structures. Other examples include complicated water bottoms, complicated salt geometry, thrusted structures, heavily faulted areas, and gas-charged sediments. In these, reflector geometries or lateral velocity variations cause wavefield distortions that must be corrected before the image can be interpreted.

Velocity estimation

Poststack migration is performed using a velocity field that is specified beforehand. The velocity can be derived from NMO velocity analysis or from an assumed geologic model. It is difficult to analyze a poststack migrated image for velocity errors. By contrast, we can analyze prestack migration for velocity errors. To do this, one performs prestack migration separately on all the input records, with individual images that overlap considerably. In the overlap areas, the different images can be checked for consistency by sorting the migrated traces into common-image-point gathers (CIGs), which are the migration analog of NMO-corrected unmigrated CMP gathers. Within a CIG, each trace refers to a single surface location, imaged from a different input record. As an example, Figure 3 shows six CIG's from a land survey, referring to images below six different locations on the earth's surface. Each CIG displays migrated traces from different source-receiver offsets. If all the individual migrations have imaged one particular event at the same depth, the images are



Seismic, Migration, Figure 3 Prestack depth migration common-image gathers from six different locations on a survey from Western Canada. Each of the gathers contains traces from migrations of data with offsets ranging from 0 m to 4,500 m. The events on the gathers are flat except at the greatest depths, indicating that the velocity field is largely correct.

consistent at that location. That is the case in this example: except at the greatest depths, most of the events on each CIG appear flat. If the different migrations image a particular event at different depths, the images will be inconsistent, and the event will not appear flat on a CIG. In order to make the images consistent for such an event, the migration velocity will need to be adjusted in the overburden above the event location.

For prestack depth migration, the velocity estimation problem is to build an interval velocity field that makes the migrated events appear flat on all the CIGs. This is usually posed as an algebraic inverse problem, called seismic tomography, involving the interval velocity field, the reflector locations, and the raypaths from the reflectors to the source and receiver locations. This problem is solved iteratively for velocity model refinements. In order to avoid converging to an unrealistic model, tomography usually needs to be constrained, using partial knowledge of the seismic velocities and reflecting horizon locations. This biases the solution, and it can result in incomplete event flattening on the CIGs. Tomography is widely used, but it usually produces velocity models with limited spatial resolution

Increasingly, seismic velocities are recognized as anisotropic: at each location, seismic velocity varies with direction. Including anisotropy makes velocity estimation more difficult, but anisotropy is usually needed to produce accurate images. The most common assumption for anisotropy is transverse isotropy, often with a tilted axis of symmetry (TTI).

Currently, almost all migrations are performed before stack, and migration is used to estimate velocity (imaging velocity for time migration, interval velocity for depth migration).

Migrated amplitude analysis for rock property determination

In the 1980s, researchers observed that careful handling of amplitude terms within migration algorithms produced migrated amplitudes that are proportional to reflection coefficient values at rock boundaries. This observation is the basis of true-amplitude migration (Bleistein, 1987). Migrated amplitudes are now often analyzed in terms of angle-dependent reflection coefficients at rock boundaries. The reflection coefficients can be used to estimate the types of rock above and below each interface, and whether the rocks can bear hydrocarbons. This endeavor has had some success in areas of moderate velocity and structural complexity, but less success below areas of high complexity, such as below salt.

Imaging conditions

Isochron imaging

Pencil-and-paper migrations were performed by swinging reflection events to their migrated locations, which are the envelopes of all possible swing locations. Kirchhoff migration can be performed the same way, with the wave equation canceling the energy away from the envelopes and leaving only images of reflectors. The time of a sample on an input trace is the sum of time from the source to a reflector plus time from the reflector to the receiver; Kirchhoff migration acts by placing the sample at all subsurface locations (isochrons) that share the same two-way travel time.

Isochron imaging, useful in describing Kirchhoff migration, is plausible and correct, but it is not a wavefield concept, and cannot be applied to all migrations based on the wave equation. These require more general treatment, which is different for poststack and prestack migration.

Poststack migration – exploding reflector model

The CMP stack simulates a zero-offset record, where energy has been excited and recorded at the same location. It can also be considered as a single wavefield from a far-fetched experiment. Suppose all the reflectors and diffractors inside the earth exploded at the same instant (time zero), sending energy in all directions; some of the energy would emerge at the earth's surface at later times to be recorded at the receiver locations. Suppose also that the propagation velocity of the wavefield from the exploding reflectors is exactly one-half the actual propagation velocity inside the earth. Then the recording time for an event from an exploding reflector is the same as the two-way zero-offset reflection time, with the actual velocity, from the same reflector. This thought experiment (Loewenthal et al., 1976) provides the basis for a poststack migration imaging principle, breaking migration into two parts: downward continuation and imaging. The first part expresses the wavefield at one depth in terms of the wavefield at a shallower depth. As the wavefield moves down, time decreases towards zero (the time of the explosion). The second part picks off the downward-continued wavefield at each location when the time reaches zero. The result of the process is a snapshot of the reflectors in the act of explosion.

Reflected wavefield amplitude normalized by source wavefield amplitude

For prestack migration, downward continuation is useful, but exploding reflectors and zero-time imaging are not. As in poststack migration, the wavefield from the many receiver locations is downward continued into the earth and backwards in time. Likewise, the wavefield from the source location is downward continued, but forward in time so that it can interact with the receiver wavefield at actual reflection times. If the two downward-continued wavefields intersect at a location at a particular time, the process has captured the instant of reflection at that point. At that location and time, the receiver wavefield amplitude equals the source wavefield amplitude times the value of the reflection coefficient at the location (possibly zero, if the location is not on an actual reflector). That is, the migrated image at that location is the receiver wavefield divided by the source wavefield.

Deconvolution and cross correlation imaging conditions

Some migration techniques are applied in the time domain, others in the frequency (ω) domain. For frequency-domain methods, a precise statement of the prestack migration imaging principle is that reflectors exist inside the earth where downward continued source and receiver wavefields are in phase, and the reflection strength is the ratio of the two wavefields. For a single frequency, there is complete uncertainty about the location of a reflector; only by adding the results from many frequencies can reflector positions be precisely determined. Thus, the prestack-migration imaging principle can be written as

$$R(\mathbf{x}) = \int \frac{U(\mathbf{x};\omega)}{D(\mathbf{x};\omega)} \mathrm{d}\omega, \tag{1}$$

where $R(\mathbf{x})$ is reflection strength at location $\mathbf{x}$, U is the downward-continued receiver wavefield, and D is the downward-continued source wavefield. The quotient can cause stability problems when the source wavefield D becomes weak. To overcome instability, note that if both numerator and denominator are multiplied by the complex conjugate of D, denoted D^*, the phase of the denominator will disappear (the denominator becomes a positive real number). Instability, caused by the nearly vanishing denominator DD^*, can be eliminated by replacing the denominator with a constant value, leaving the modified imaging condition

$$R(\mathbf{x}) = \int U(\mathbf{x};\omega) D*(\mathbf{x};\omega) \mathrm{d}\omega. \tag{2}$$

Equations 1 and 2 migrate events to the same locations, but migrated amplitudes obtained using Equation 2 are different from those obtained using Equation 1. Imaging conditions (1) (*deconvolution*) and (2) (*crosscorrelation*) were introduced by Claerbout (1971).

Migration techniques

Integral (Kirchhoff) migration

Kirchhoff migration is the digital embodiment of the earliest migration heuristics, and it has a basis in the downward continuation of wavefields: the wavefield at depth is an integral of the recorded wavefield (Schneider, 1978). The mathematical physics of this operation requires amplitude and phase manipulations of the recorded wavefield, yielding various expressions for poststack and prestack migration with different recording configurations. For example, shot-record Kirchhoff migration is expressed as

$$I(\mathbf{x};\mathbf{x_s}) = \int \mathrm{d}\mathbf{x_r} \int \mathrm{d}t W \frac{\partial U(\mathbf{x_r};\mathbf{x_s};t)}{\partial t} \delta[t-(t_s+t_r)], \tag{3}$$

where $I(\mathbf{x}, \mathbf{x_s})$ is the migrated image at location $\mathbf{x}$ due to a source at $\mathbf{x_s}$, $\mathbf{x_r}$ are the receiver locations, W is a weight function, U is the recorded wavefield, δ is the Dirac delta function, and t_s and t_r are travel times from $\mathbf{x_s}$ and $\mathbf{x_r}$ to $\mathbf{x}$. The integral accumulates data samples from the receiver locations into the image at $\mathbf{x}$. Alternatively, Equation 3 allows an input sample at time $t_s + t_r$ from receiver location $\mathbf{x_r}$ to be smeared out into the isochron surface.

Kirchhoff migration is flexible, allowing the accumulation of trace values from any subvolume of the recorded traces into any subvolume of the image. This has made Kirchhoff migration very popular. However, complicated velocity models used in many depth migration projects have shown its accuracy limits, which are due mostly to a reliance on asymptotic ray theory to provide travel times and amplitudes.

Beam migration

In its simplest form, beam migration is a Kirchhoff migration with the integral over the recording surface broken into two stages: first, the domain of integration is divided into a number of overlapping regions; second, a directional decomposition of the data within each region is performed. This decomposition produces a set of local plane waves, over a range of angles, emerging within each region. Each of these local-plane wave data components is mapped back into the earth using raypaths sent in the directions of the emerging plane waves. A more complete description (Hill, 2001) uses Green's identity to downward continue the wavefields, with Gaussian beams as Green's functions. Gray et al. (2009) provide an elementary physical explanation of this complicated method, which is generally more accurate than Kirchhoff migration because it naturally allows several arrivals from each wavefield to accumulate at any image location (which is difficult for Kirchhoff migration). It has many of the flexibilities of Kirchhoff migration, but, like Kirchhoff migration, relies on asymptotic ray theory.

One-way wave-equation migration (OWEM)

Strictly speaking, applying the term "wave-equation migration" solely to migration methods that downward-continue wavefields recursively one depth step at a time is an abuse of terminology. This usage excludes Kirchhoff and beam migrations, implying that they do not derive from the wave equation, which is not true. This usage has become entrenched.

Downward continuing a wavefield from one depth to the next requires that the wavefield is propagating either up or down, but not both at once. (The wavefield can propagate laterally, but up or down is a distinguished direction.) Further, it requires an appropriate (upgoing or downgoing) wave equation. In the simple constant-velocity case, the one-way solution can be obtained easily. The two-way wave equation for wavefield P (acoustic, ignoring elastic effects such as mode conversion) is

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right) P = 0. \tag{4}$$

When this equation is expressed in the frequency/horizontal wave number domain, it factors into separate equations for upgoing and downgoing wavefields. Analytical solutions of these equations provide phase shifts that perform downward continuation. Once the downward-continued wavefields have been inverse Fourier transformed from wave number back to space, they can be combined and summed over frequency using Equation 1 or 2 to produce the migrated image (Gazdag, 1978). This method, phase-shift migration, can accurately image dips up to 90°. Its only limitation is the lateral invariance of the velocity field. It has proven successful in imaging salt flanks in marine sedimentary basins. Beneath the salt flanks, where significant velocity contrast has occurred, the phase-shift image is inaccurate.

Factoring the wave equation exactly into one-way equations works only for constant velocity. When velocity varies laterally, the factoring, again with z as a distinguished direction, is approximate, incurring kinematic (time) and dynamic (amplitude) errors. A number of methods downward continue wavefields using approximate one-way wave equations. The earliest of these used finite differences to discretize the equations, using low-angle approximations (e.g., Claerbout, 1970). Later methods generalized phase-shift migration to handle lateral velocity variations, and other methods combine aspects of finite-difference and phase-shift approaches. Development of these methods continues today. Because of the errors inherent in even the best one-way wave equations, none of the methods can be completely accurate.

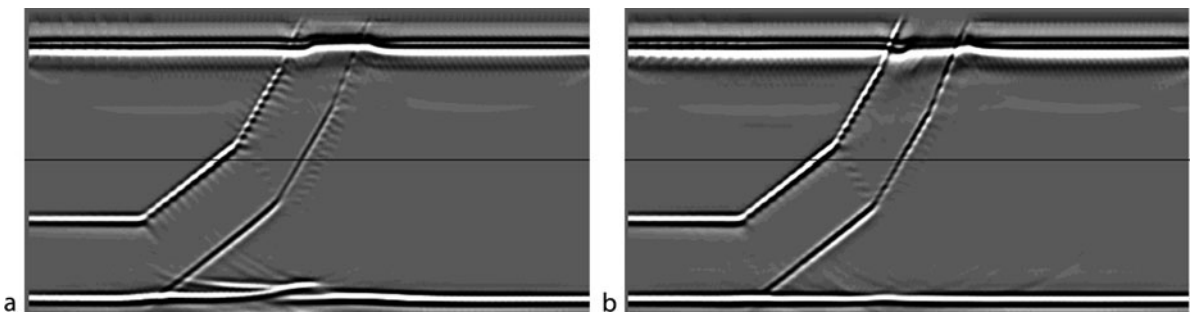
Two-way wave-equation migration (reverse-time migration)

Reverse-time migration (RTM) uses the full two-way wave equation 4. It does not operate by downward continuation of downgoing or upgoing wavefields. Instead, it takes the source and receiver wavefields at the recording surface and injects them into the earth by treating them as boundary values for the wave equation (McMechan, 1983), computing full wavefields inside the earth at each time step. The source wavefield is computed forwards in time, so successive wavefield snapshots show an expanding wavefront moving into the earth and interacting with reflectors and diffractors. The receiver wavefield is computed backwards in time (beginning with the final time sample on each trace of the recorded data), so successive snapshots show reflection events moving generally downward. Imaging is performed using a temporal version of the imaging condition (1) or (2). In principle, RTM is the most accurate migration method: it does not rely on asymptotic ray theory, and it suffers from neither the steep-dip nor the kinematic and dynamic errors of OWEM. Until recently, its use has been limited because it tends to be slower than other migration methods, and numerical approximations to the wave equation (e.g., low-order finite-difference approximations) produced errors (e.g., numerical dispersion) that other methods do not incur. However, modern computational hardware has both the speed and the memory availability to allow very good approximations to the wave equation, and high-quality RTM images are now the norm.

Migration examples

There are many reasons to migrate seismic data; we will need to be content with a few illustrations.

The first illustrates structural imaging in the presence of anisotropy. Figure 4 shows two images of a cross section designed to mimic the structural geology of a dipping thrust sheet, simplified to isolate the effect of anisotropy on image positioning and quality. The velocity is generally simple, with a constant velocity of 2,740 m/s outside the thrust sheet. The velocity complication lies in the thrust sheet, which is anisotropic (TTI), with a symmetry axis parallel to the boundaries of the sheet. In the thrust sheet, the velocity is 2,925 m/s in the direction of the symmetry axis, and higher in all other directions. Both images in Figure 4 are from prestack beam migration of the same input data set, which was generated by finite-difference elastic modeling (courtesy BP). In Figure 4a, anisotropy was ignored, and in Figure 4b, anisotropy was taken into account. The reflector below the dipping thrust sheet should be flat, as it is in Figure 4b. In Figure 4a, the flat reflector is pulled up beneath the thrust sheet, and is poorly imaged. The high-amplitude flat event cutting across the



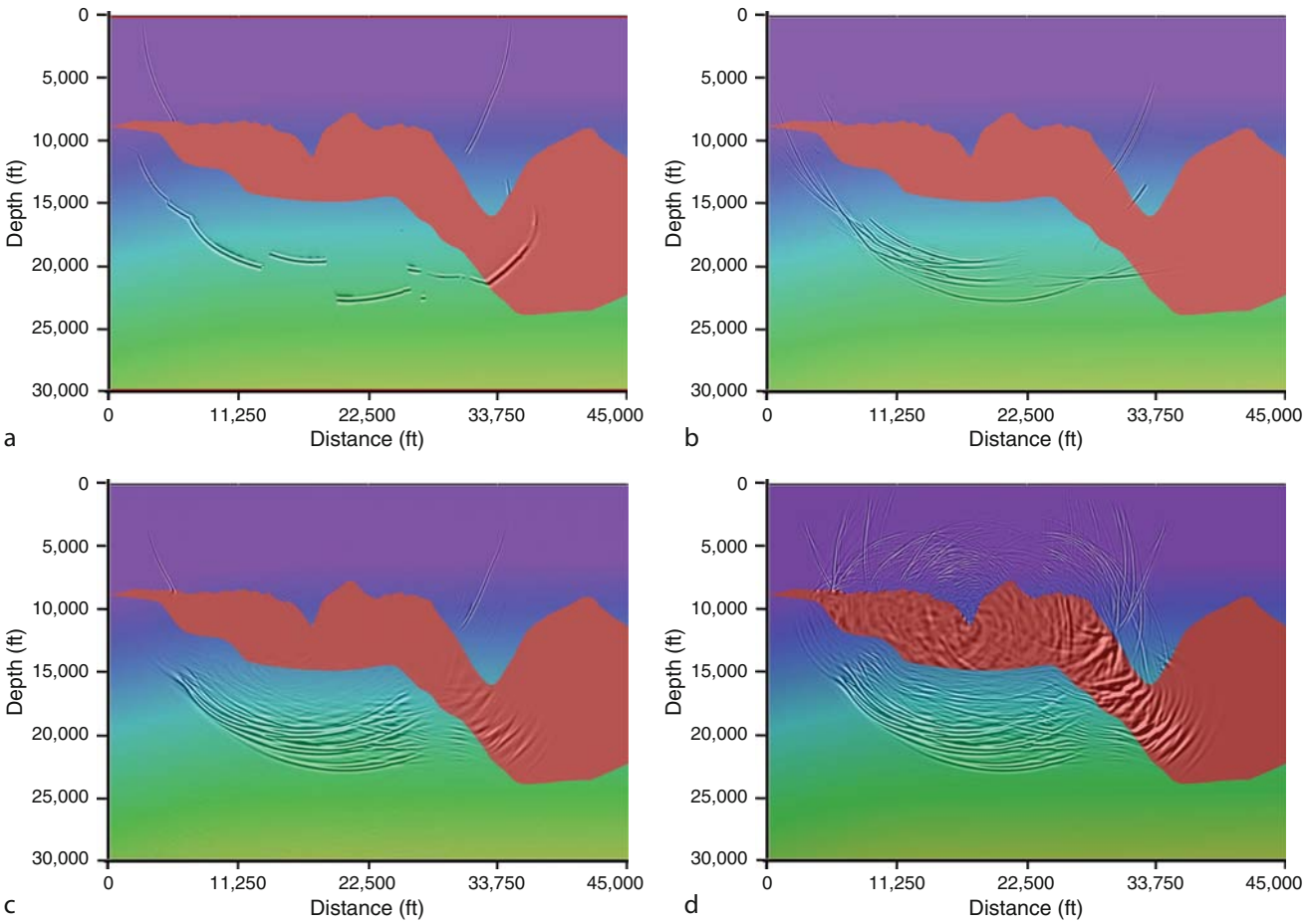
Seismic, Migration, Figure 4 Prestack depth migrated images from a synthetic thrust model data set. In (**a**), the anisotropy in the dipping thrust sheet is ignored by the migration, leading to inaccurate imaging below the thrust sheet. In (**b**), the migration has accounted for the anisotropy, leading to correct structural imaging.

entire section at shallow depths is a shear-wave artifact that has been mis-migrated by the P-wave migration program.

The second example illustrates the action of various migration methods: Kirchhoff, beam, OWEM, and RTM. Figure 5 shows four images overlaid on a geologic cross section, each obtained by migrating a single event on a single zero-offset trace through the complicated geology/velocity model shown (called Sigsbee2a). A high-velocity salt body of complex shape has been inserted into a set of lower-velocity sedimentary layers, and we note the effect of the geometry and velocity on the migration algorithms. The migrated images are, in effect, wavefronts modeled by the migration methods; full migrated images are formed by adding together such partial images from all samples on all traces. The Kirchhoff migrated image (Figure 5a) shows the effect of the assumption, made by most Kirchhoff migration programs, that at most one travel path for seismic energy connects a given upper-surface location with each subsurface location. The wavefield discontinuities in the subsalt area are physically impossible, and will cause inaccurate, noisy images in areas of structural complexity. The remaining images show wavefields of greater continuity and complexity in the subsalt area. The most complicated wavefield, produced by RTM, should produce the most accurate image. This happens in practice when the velocity field is accurate and the input traces are free from unwanted noise such as elastic-wave mode-converted events.

The evolving role of migration in seismic data processing

Before seismic processing became digital, migration was a standalone process. Pencil-and-paper or mechanical migration provided a cross section that showed where to drill. Later, poststack migration was the final seismic processing step; all other processing was intended to prepare a stacked data set for migration. With prestack migration, migration has moved to a central location in the processing flow. In tomography, it is common to process the signals on a migrated stack or in a set of CIG's. In postmigration amplitude analysis, it is common to analyze and alter the migrated waveforms. In the future, when the



Seismic, Migration, Figure 5 The action of migration, illustrated on four different migration methods: (**a**) Kirchhoff; (**b**) beam; (**c**) one-way wave-equation; (**d**) reverse-time. In each, a single event from a single zero-offset trace was migrated. The migration methods have interacted differently with the velocity model, producing different wavefronts. (From Etgen et al., 2009).

wave equation used for migration can model earth effects more completely than at present, seismic data will undergo less processing before migration, and the migrated data will predict rock properties directly.

Summary

Seismic migration has been used in prospecting for oil and natural gas since the 1920s. Since the 1970s, migration has been applied as a wave-equation process, explicitly recognizing that reflections recorded at the earth's surface are the result of wavefields propagating and reflecting inside the earth. Wavefield migration has taken several forms (time-domain, frequency-domain; finite-difference, integral) because of the variety of ways for expressing the propagation of wavefields. All these different forms of migration are still in use because each of them has features that the others lack: flexibility, steep-dip capability, etc. When computer power was limited, migration was usually performed after NMO/stack, allowing a reduction of the size of the data volume input to migration. Nowadays most migrations are performed before stack in 3-D. Another shift has been from time to depth, bringing migration closer to the problem of estimating seismic velocities inside the earth. A separate development has been the capability of migration to preserve amplitudes for estimating rock properties near reflector locations. With migration velocity estimation and amplitude analysis, migration has moved from the final step of the seismic processing flow to a more central role. This means that an increasing amount of processing and analysis is performed on migrated gathers before they are stacked to form the final structural image.

Bibliography

Berkhout, A. J., 1982. *Seismic Migration – Imaging of Acoustic Energy by Wavefield Extrapolation*. Amsterdam: Elsevier.
Bleistein, N., 1987. On the imaging of reflectors in the earth. *Geophysics*, **52**, 931–942.
Claerbout, J., 1970. Coarse-grid calculations of waves in inhomogeneous media with application to delineation of complicated seismic structure. *Geophysics*, **35**, 407–418.
Claerbout, J., 1971. Toward a unified theory of reflector mapping. *Geophysics*, **36**, 467–481.
Dix, C. H., 1955. Seismic velocities from surface measurements. *Geophysics*, **20**, 68–86.
Etgen, J., Gray, S. H., and Zhang, Y., 2009. An overview of depth migration in exploration geophysics. *Geophysics*, **74**, WCA5–WCA18.
Gazdag, J., 1978. Wave equation migration with the phase-shift method. *Geophysics*, **43**, 1342–1351.
Gray, S. H., Xie, Y., Notfors, C., Zhu, T., Wang, D., and Ting, C.-O., 2009. Taking apart beam migration. *The Leading Edge*, **28**, 1098–1109.
Hagedoorn, J. G., 1954. A process of seismic reflection interpretation. *Geophysical Prospecting*, **6**, 449–453.
Hill, N. R., 2001. Prestack Gaussian-beam depth migration. *Geophysics*, **66**, 1240–1250.
Loewenthal, D., Roberson, R., Sherwood, J., and Lu, L., 1976. The wave equation applied to migration. *Geophysical Prospecting*, **24**, 380–399.
Mayne, W. H., 1962. Common reflection point horizontal data stacking techniques. *Geophysics*, **27**, 927–938.
McMechan, G. A., 1983. Migration by extrapolation of time-dependent boundary values. *Geophysical Prospecting*, **31**, 413–420.
Schneider, W. A., 1978. Integral formulation for migration in two and three dimensions. *Geophysics*, **43**, 49–76.
Trorey, A. W., 1970. A simple theory of seismic diffractions. *Geophysics*, **35**, 762–784.
Warner, M., 1987. Migration – why doesn't it work for deep continental data. *Geophysical Journal of the Royal Astronomical Society*, **89**, 21–26.
Yilmaz, O., 1987. *Seismic Data Processing*. Tulsa: SEG.

Cross-references

Seismic Anisotropy
Seismic Data Acquisition and Processing
Seismic Diffraction
Seismic Imaging, Overview
Seismic Properties of Rocks
Seismic Waves, Scattering
Seismic, Ray Theory
Seismic, Reflectivity Method
Seismic, Waveform Modeling and Tomography
Traveltime Tomography Using Controlled-Source Seismic Data

SEISMIC, RAY THEORY

Vlastislav Červený[1], Ivan Pšenčík[2]
[1]Department of Geophysics, Mathematics and Physics, Charles University, Praha, Czech Republic
[2]Institute of Geophysics, Academy of Sciences of Czech Republic, Praha, Czech Republic

Synonyms

Asymptotic ray theory; Ray series method; Seismic ray method

Definition

Seismic ray theory. High-frequency asymptotic method of study of seismic wavefields in complex inhomogeneous isotropic or anisotropic media with curved structural interfaces.

Introduction

The ray theory belongs to the methods most frequently used in seismology and seismic exploration for forward and inverse modeling of high-frequency seismic body waves. In smoothly varying media with smooth interfaces, it can provide useful approximate solutions of the elastodynamic equation of satisfactory accuracy. Starting from an intuitive description of the propagation of seismic waves along special trajectories - rays, it has developed into a highly sophisticated method, described briefly in this review paper.

The ray method has its advantages and disadvantages. The basic advantages are its applicability to complex,

isotropic and anisotropic, laterally varying layered media and its numerical efficiency in such computations. It provides a physical insight into the wave propagation process by separating the wavefield into individual elementary waves and by allowing their identification. In addition, it makes possible to track the paths in the medium along which energy of individual waves propagates, an aspect very important in tomography. The ray method also represents an important basis for other related, more sophisticated methods, such as the paraxial ray method, the Gaussian beam summation method, the Maslov method, the asymptotic diffraction theory, etc. The ray method also has some limitations. As mentioned above, it is approximate. It is applicable only to smooth media with smooth interfaces, in which the characteristic dimensions of inhomogeneities are considerably larger than the prevailing wavelength of the considered waves. The ray method can yield distorted results and may even fail in some special regions called *singular regions*.

The seismic ray method owes a lot to optics and radiophysics. Although the techniques used in different branches of physics are very similar, there are some substantial differences. The ray method in seismology is usually applied to more complicated structures than in optics or radiophysics. There are also different numbers and types of waves considered in different branches of physics.

The first seismological applications of ray concepts date back to the end of the 19th century. Then, only *kinematics*, specifically travel times, were used. Probably the first attempts to use also *dynamics* (amplitudes and waveforms) were made by Sir H. Jeffreys. The ray series solutions of elastodynamic equation with variable coefficients were first suggested by Babich (1956) and Karal and Keller (1959) for inhomogeneous isotropic media, and by Babich (1961) for inhomogeneous anisotropic media.

The Earth's interior is anisotropic or weakly anisotropic in some of its parts. Seismic anisotropy and its effects on wave propagation play an important role in contemporary seismology and seismic exploration. Consequently, it has also been necessary to develop the ray theory for elastic anisotropic media. It is important to emphasize that, for S waves, the ray theory for anisotropic media does not yield the ray theory for isotropic media in the zero anisotropy limit. For this reason, we describe systematically the ray theory for anisotropic media and also present corresponding formulae for isotropic media, and explain the differences between both of them.

S waves require generally a special attention. Well understood phenomenon is propagation of two separate shear waves in anisotropic media. Less understood and an underestimated phenomenon is shear-wave coupling, which occurs in weakly anisotropic media or in vicinities of shear-wave singularities. In such regions, standard ray theories for anisotropic as well as isotropic media do not work properly. Therefore, we also briefly describe the coupling ray theory for S waves, which fills the gap between ray theories for isotropic and anisotropic media.

We give here neither a detailed derivation of ray-theoretical expressions nor a relevant systematic bibliography. This would extend the text considerably. We refer, however, to several textbooks, in which the ray theory is treated in a considerably greater detail (Červený et al., 1977; Kravtsov and Orlov, 1990; Červený, 2001; Chapman, 2004). The reader may also find useful information in several review papers devoted to seismic ray theory and its various aspects (Červený et al., 1988; Virieux, 1996; Chapman, 2002; Červený et al., 2007). Examples of computations based on the ray theory can be found, for example, in Červený et al. (1977) and Gjøystdal et al. (2002). Here we refer only to papers, in which the relevant methods and procedures were first proposed, and/or which give a useful more recent treatment of the subject.

We use the following notation. We denote Cartesian coordinates x_i and time t. The dots above letters denote partial derivatives with respect to time ($\ddot{u}_i = \partial^2 u_i/\partial t^2$) and the index following the comma in the subscript indicates the partial derivative with respect to the relevant Cartesian coordinate ($u_{i,j} = \partial u_i/\partial x_j$). We consider high-frequency time-harmonic seismic body waves, with the exponential factor $\exp(-\mathrm{i}\omega t)$, where ω is fixed, positive, real-valued circular frequency. The lower-case Roman indices take the values 1, 2, 3, the upper-case indices 1, 2. Hats over bold symbols indicate 3×3 matrices, bold symbols without hats denote 2×2 matrices. The Einstein summation convention over repeating Roman indices is used, with exception of indices in parentheses.

Basic equations of the seismic ray method

For smoothly varying elastic media, the source-free *equation of motion* reads

$$\tau_{ij,j} - \rho\,\ddot{u}_i = 0. \tag{1}$$

Here $\tau_{ij}(x_n, t)$, and $u_i(x_n, t)$ are Cartesian components of stress tensor and displacement vector, respectively, and ρ is the density. In anisotropic media, the stress tensor τ_{ij} and the infinitesimal strain tensor $e_{ij} = \frac{1}{2}(u_{i,j} + u_{j,i})$ are related by Hooke's law:

$$\tau_{ij} = c_{ijkl} e_{kl} = c_{ijkl} u_{k,l}. \tag{2}$$

$c_{ijkl}(x_n)$ is a tensor of elastic moduli (stiffness tensor), satisfying symmetry relations $c_{ijkl} = c_{jikl} = c_{ijlk} = c_{klij}$. There are, at the most, 21 independent elastic moduli. Inserting Equation 2 into Equation 1, we get the *elastodynamic equation*

$$(c_{ijkl} u_{k,l})_{,j} - \rho\,\ddot{u}_i = 0. \tag{3}$$

In the seismic ray method, high-frequency seismic body waves propagating in smoothly varying, isotropic or anisotropic, media are studied. The formal ray series solution of the elastodynamic equation (3) for the displacement vector

$\mathbf{u}(x_n,t)$ is sought in the form of an asymptotic series in inverse powers of circular frequency ω,

$$\mathbf{u}(x_n,t) = \exp[-\mathrm{i}\omega(t-T(x_n))] \left[\mathbf{U}^{(0)}(x_n) + \frac{\mathbf{U}^{(1)}(x_n)}{(-\mathrm{i}\omega)} + \frac{\mathbf{U}^{(2)}(x_n)}{(-\mathrm{i}\omega)^2} + \ldots\right]. \quad (4)$$

Here $T(x_n)$ is the real-valued travel time, $\mathbf{U}^{(k)}$, $k = 0,1,2,\ldots$ are complex-valued *vectorial amplitude coefficients*. Surfaces $T(x_i) = const.$ are called *wavefronts* (or phase fronts). In perfectly elastic media, functions $T(x_n)$, and $\mathbf{U}^{(k)}(x_n)$ are frequency independent.

Also other forms of the ray series have been used in the seismic ray method. For example, Chapman (2004) developed the seismic ray method using the ray series for particle velocity and traction. Such a formal ray series has certain advantages with respect to Equation 4. Here, however, we consider systematically the traditional ray series (4) for the displacement vector.

Inserting Equation 4 into elastodynamic equation (3), we obtain a series in inverse powers of ω, which equals zero. Consequently, the coefficients of the individual powers of ω must also equal zero. This yields a system of equations called the basic recurrence system of equations of the ray method. This system can be used to determine the eikonal equations for travel times $T(x_n)$ and, successively the equations for the amplitude coefficients $\mathbf{U}^{(0)}(x_n)$, $\mathbf{U}^{(1)}(x_n)$, $\mathbf{U}^{(2)}(x_n)$,.... The equations for $\mathbf{U}^{(k)}(x_n)$ yield, among others, transport equations. For a detailed derivation of the basic system of equations of the ray method see Červený (2001, sect. 5.7).

The vectorial amplitude coefficients $\mathbf{U}^{(k)}(x_n)$, $k = 1,2,\ldots$, can be expressed as a sum of the *principal component* and *additional component*. The principal component of $\mathbf{U}^{(k)}(x_n)$ is the projection of $\mathbf{U}^{(k)}(x_n)$ into the unit vector parallel to the zero-order amplitude coefficient $\mathbf{U}^{(0)}(x_n)$, the additional component of $\mathbf{U}^{(k)}(x_n)$ is the remaining part of $\mathbf{U}^{(k)}(x_n)$. In this way, the additional component of the zero-order amplitude coefficient $\mathbf{U}^{(0)}(x_n)$ is zero. The complexity of the equations for higher-order amplitude coefficients $\mathbf{U}^{(k)}$ increases rapidly with increasing k. Moreover, the higher-order amplitude coefficients are inaccurate and unstable, as they are very sensitive to fine details of the medium. The instability of the amplitude coefficients increases with increasing k. For these reasons, only the zero-order coefficient $\mathbf{U}^{(0)}(x_n)$, at the most with the additional component of $\mathbf{U}^{(1)}(x_n)$, has been used in seismological applications. In the following, we shall concentrate on the zero-order ray approximation only.

The zero-order approximation of the ray method reads:

$$\mathbf{u}(x_n,t) = \mathbf{U}(x_n)\exp[-\mathrm{i}\omega(t-T(x_n))]. \quad (5)$$

In Equation 5, we have dropped the superscript (0) of $\mathbf{U}(x_n)$. We call $\mathbf{U}(x_n)$ the complex-valued *vectorial amplitude*. In smooth, laterally varying media, containing smooth structural interfaces, the zero-order approximation (5) of the ray method usually offers sufficiently accurate results, particularly for travel time $T(x_n)$. Its great advantage is that it allows one to work with frequency-independent travel time and amplitude. However, if the medium under consideration becomes more and more complex (less smooth), vectorial amplitude $\mathbf{U}(x_n)$ becomes less accurate. In structures exceeding a certain degree of complexity, the ray method may yield inaccurate results or even fail.

The first equation of the basic system of equations of the ray method reads:

$$(\Gamma_{ik} - \delta_{ik})U_k = 0, \quad i = 1,2,3. \quad (6)$$

Here $\mathbf{\Gamma}$ is the 3×3 *generalized Christoffel matrix* with elements given by the relation:

$$\Gamma_{ik} = a_{ijkl}p_j p_l. \quad (7)$$

In Equation 7, p_i are the Cartesian components of the slowness vector $\mathbf{p}$,

$$p_i = \partial T/\partial x_i \quad (8)$$

and $a_{ijkl} = c_{ijkl}/\rho$ are density-normalized elastic moduli. Note that the classical Christoffel matrix, with elements $a_{ijkl}n_j n_l$, contains components of the real-valued unit vector $\mathbf{n}$ (perpendicular to the wavefront) instead of $\mathbf{p}$. For this reason, we call $\mathbf{\Gamma}$the "generalized" Christoffel matrix. The relation between p_i and n_i is $p_i = n_i/\mathcal{C}$, where $\mathcal{C}$ is the phase velocity.

The generalized 3×3 Christoffel matrix in solid media is symmetric ($\Gamma_{ik} = \Gamma_{ki}$), positive definite ($\Gamma_{ik}a_i a_k > 0$, where a_i are components of any non-vanishing real-valued vector) and homogeneous function of the second degree in p_i ($\Gamma_{ik}(x_n, ap_j) = a^2\Gamma_{ik}(x_n,p_j)$ for any non-vanishing constant a). It has three real-valued positive eigenvalues $G_m(x_n,p_j)$, and three corresponding real-valued unit eigenvectors $\mathbf{g}^{(m)}(x_n,p_j)$, $m = 1,2,3$. G_m and $\mathbf{g}^{(m)}$ are solutions of the eigenvalue equation

$$(\Gamma_{ik} - \delta_{ik}G_m)g_k^{(m)} = 0, \quad i = 1,2,3. \quad (9)$$

Eigenvectors $\mathbf{g}^{(1)}, \mathbf{g}^{(2)}, \mathbf{g}^{(3)}$ are mutually perpendicular. Eigenvalue G_m and the relevant eigenvector $\mathbf{g}^{(m)}$ are mutually related as follows:

$$G_m = \Gamma_{ik}g_i^{(m)}g_k^{(m)} = a_{ijkl}p_j p_l g_i^{(m)}g_k^{(m)}. \quad (10)$$

For isotropic media, it is sufficient to specify elastic moduli $c_{ijkl}(x_n)$ in terms of Lamé's elastic moduli $\lambda(x_n)$ and $\mu(x_n)$, describing isotropic media, as follows:

$$c_{ijkl} = \lambda\delta_{ij}\delta_{kl} + \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}). \quad (11)$$

Elements of the generalized Christoffel matrix are then given by the relation:

$$\Gamma_{ik} = \frac{\lambda+\mu}{\rho}p_i p_k + \frac{\mu}{\rho}\delta_{ik}p_n p_n. \quad (12)$$

In isotropic media, the expressions for eigenvalues and eigenvectors of the generalized Christoffel matrix can be determined analytically:

$$G_1 = G_2 = \beta^2 p_k p_k, \qquad G_3 = \alpha^2 p_k p_k. \tag{13}$$

Here

$$\alpha^2 = (\lambda + 2\mu)/\rho, \qquad \beta^2 = \mu/\rho. \tag{14}$$

The eigenvector relevant to the eigenvalue G_3 equals $\mathbf{n}$, the unit vector perpendicular to the wavefront. The eigenvectors relevant to coinciding eigenvalues G_1 and G_2 are mutually perpendicular unit vectors situated arbitrarily in the plane perpendicular to $\mathbf{n}$.

Eikonal equation. Polarization vector

The comparison of the basic equation of the ray method (6) with the eigenvalue equation (9) for the 3×3 generalized Christoffel matrix shows that Equation 6 is satisfied, if the eigenvalue G_m of the generalized Christoffel matrix satisfies the relation

$$G_m(x_i, p_j) = 1, \tag{15}$$

and if the complex-valued vectorial amplitude $\mathbf{U}$ of the wave under consideration is related to eigenvector $\mathbf{g}^{(m)}$ as follows:

$$\mathbf{U} = A\mathbf{g}^{(m)}. \tag{16}$$

Equation 15 is the important *eikonal equation*. It is a nonlinear, first-order partial differential equation for travel time $T(x_n)$. Equation 16 shows that displacement vector $\mathbf{U}$ is parallel to the appropriate eigenvector $\mathbf{g}^{(m)}$. For this reason, we call $\mathbf{g}^{(m)}$ the *polarization vector*. Symbol $A(x_n)$ denotes the complex-valued, frequency-independent, *scalar amplitude*.

Taking into account that G_m is a homogeneous function of the second degree in p_i, where $\mathbf{p} = \mathcal{C}^{-1}\mathbf{n}$, we obtain $G_m(x_i, p_j) = \mathcal{C}^{-2} G_m(x_i, n_j)$. This, Equations 15 and 10 yield

$$\mathcal{C}^2(x_i, n_j) = G_m(x_i, n_j) = a_{ijkl} n_j n_l g_i^{(m)} g_k^{(m)}. \tag{17}$$

Phase velocity $\mathcal{C}$ is the velocity of the wavefront in direction $\mathbf{n}$. The phase-velocity vector $\boldsymbol{\mathcal{C}} = \mathcal{C}(x_i, n_j)\mathbf{n}$ has the direction of $\mathbf{n}$, i.e., it is perpendicular to the wavefront. It follows from Equation 17 that the squares of phase velocity $\mathcal{C}$ are eigenvalues $G_m(x_i, n_j)$ of the classical Christoffel matrix with elements $a_{ijkl} n_j n_l$.

Generally, eigenvalues G_m, $m = 1, 2, 3$, of the generalized Christoffel matrix are mutually different. They correspond to three high-frequency body waves propagating in inhomogeneous anisotropic media. We assign G_1 and G_2 to S1 and S2 waves and G_3 to P wave. If the eigenvalues are different, their polarization vectors can be determined uniquely.

If two eigenvalues coincide, we speak of the *degenerate case* of the eigenvalue problem. The corresponding eigenvectors can then be chosen as mutually perpendicular vectors situated arbitrarily in the plane perpendicular to the third eigenvector. Eigenvalues G_m may coincide *locally*, along certain lines or at certain points, which correspond to the so-called *S-wave singular directions*, or may be close to one another *globally* in a vicinity of singular directions or in *weakly anisotropic media*. The approximate but unique determination of polarization vectors in the latter situations is possible using perturbation approach (Jech and Pšenčík, 1989).

In isotropic media, the S-wave eigenvalues G_1 and G_2 coincide *globally*, see Equation 13. Consequently, in isotropic media, the S waves are controlled by a single eikonal equation and we have thus only two different eikonal equations corresponding to P and S waves. As the equations for the eigenvalues in isotropic media can be determined analytically, we can express the eikonal equations for P and S waves explicitly:

$$\alpha^2 p_k p_k = 1 \quad \text{for P waves}, \tag{18}$$

$$\beta^2 p_k p_k = 1 \quad \text{for S waves}. \tag{19}$$

In isotropic media, the generally complex-valued amplitude vector $\mathbf{U}$ can be expressed in the simple form (16) only for P waves. In this case the polarization vector $\mathbf{g}^{(3)} = \mathbf{n}$, i.e., it is perpendicular to the wavefront. For S waves, $\mathbf{U}$ must be considered in the following form:

$$\mathbf{U} = B\mathbf{g}^{(1)} + C\mathbf{g}^{(2)}. \tag{20}$$

Here $\mathbf{g}^{(1)}$ and $\mathbf{g}^{(2)}$ are two mutually perpendicular unit vectors in the plane tangent to the wavefront, i.e., perpendicular to the vector $\mathbf{n}$. The computation of $\mathbf{g}^{(1)}$ and $\mathbf{g}^{(2)}$ along the ray is explained later, see Equation 37. Symbols $B(x_n)$ and $C(x_n)$ are the corresponding, generally complex-valued scalar amplitudes.

In the seismic ray method, it is common to express the eikonal equation (15) in Hamiltonian form. Hamiltonian $\mathcal{H}(x_i, p_j)$ may be introduced in various ways. We shall consider the Hamiltonian, which is a homogeneous function of the second degree in p_i. For inhomogeneous anisotropic media, we can introduce the Hamiltonian expressed in terms of $G_m(x_i, p_j)$, see Equation 10:

$$\mathcal{H}(x_i, p_j) = \frac{1}{2} G_m(x_i, p_j) = \frac{1}{2} a_{ijkl} p_j p_l g_i^{(m)} g_k^{(m)}. \tag{21}$$

The eikonal equation (15) then yields:

$$\mathcal{H}(x_i, p_j) = \frac{1}{2}. \tag{22}$$

It holds for anisotropic as well as isotropic media.

From Equations 13 and 21, we get for isotropic inhomogeneous media:

$$\mathcal{H}(x_i, p_j) = \frac{1}{2} V^2(x_i) p_k p_k, \tag{23}$$

where $V = \alpha$ for P waves and $V = \beta$ for S waves.

Ray tracing and travel-time computation

The eikonal equation in Hamiltonian form (22), with $p_j = \partial T/\partial x_j$, is a non-linear partial differential equation of the first order for travel time $T(x_i)$. It can be solved by the method of characteristics. The characteristics of eikonal equation (22) are spatial trajectories, along which Equation 22 is satisfied, and along which travel time T can be computed by quadratures. The characteristics of the eikonal equation represent *rays*.

The characteristics of the eikonal equation expressed in general Hamiltonian form are described by a system of non-linear, ordinary differential equations of the first order:

$$\frac{\mathrm{d}x_i}{\mathrm{d}u} = \frac{\partial \mathcal{H}}{\partial p_i}, \quad \frac{\mathrm{d}p_i}{\mathrm{d}u} = -\frac{\partial \mathcal{H}}{\partial x_i}, \quad \frac{\mathrm{d}T}{\mathrm{d}u} = p_k \frac{\partial \mathcal{H}}{\partial p_k}. \tag{24}$$

Here u is a real-valued parameter along the ray. The relation between parameter u and the travel time along the ray depends on the form of the Hamiltonian used, see the last equation in Equations 24. For Hamiltonians, which are homogeneous functions of the second degree in p_i, the Euler equation for homogeneous functions yields $p_k \partial \mathcal{H}/\partial p_k = 2\mathcal{H}$. If we consider Hamiltonian (21), we get $\mathrm{d}T/\mathrm{d}u = 1$ from Equations 24. For travel time T along the ray, denoted $\tau = T$, Equations 24 simplify to:

$$\frac{\mathrm{d}x_i}{\mathrm{d}\tau} = \frac{\partial \mathcal{H}}{\partial p_i}, \quad \frac{\mathrm{d}p_i}{\mathrm{d}\tau} = -\frac{\partial \mathcal{H}}{\partial x_i}. \tag{25}$$

This system of equations is usually called the *ray tracing system*. Solution of the ray tracing system (25) with appropriate initial conditions yields $x_i(\tau)$, the coordinates of points along the ray trajectory, and $p_i(\tau)$, the Cartesian components of the slowness vectors along the ray. The travel time T along the ray is obtained automatically, $T = \tau$.

Inserting Equation 21 in Equations 25, we obtain the ray tracing system for m-th wave in inhomogeneous anisotropic media:

$$\begin{aligned} \frac{\mathrm{d}x_i}{\mathrm{d}\tau} &= a_{ijkl} p_l g_j^{(m)} g_k^{(m)}, \\ \frac{\mathrm{d}p_i}{\mathrm{d}\tau} &= -\frac{1}{2} \frac{\partial a_{jkln}}{\partial x_i} p_k p_n g_j^{(m)} g_l^{(m)}. \end{aligned} \tag{26}$$

In the derivation of the first set of Equations 26 for $\partial \mathcal{H}/\partial p_i$, we took into account that $\Gamma_{ik} \partial (g_i^{(m)} g_k^{(m)})/\partial p_n = 0$. An alternative version of ray tracing equation (26) was derived by Červený (1972), in which the eigenvectors $\mathbf{g}^{(m)}$ are not used.

The initial conditions for the ray tracing system (26) are $x_i = x_{0i}$, $p_i = p_{0i}$, where x_{0i} and p_{0i} satisfy the eikonal equation (22), corresponding to the wave we wish to compute (P, S1 or S2). Components p_{0i} of the initial slowness vector $\mathbf{p}_0$ can be then expressed as $p_{0i} = n_{0i}/\mathcal{C}(x_{0i})$, where $\mathcal{C}$ is the relevant phase velocity. The eikonal equation (22) is then satisfied along the whole ray.

In inhomogeneous isotropic media, the ray tracing system (25) with Equation 23 yields

$$\frac{\mathrm{d}x_i}{\mathrm{d}\tau} = V^2 p_i, \quad \frac{\mathrm{d}p_i}{\mathrm{d}\tau} = -\frac{\partial \ln V}{\partial x_i}. \tag{27}$$

The initial conditions for the ray tracing system (27) are again $x_i = x_{0i}$, $p_i = p_{0i}$, where $p_{0i} = n_{0i}/V(x_{0i})$. Here $V = \alpha$ for P waves, and $V = \beta$ for S waves.

As τ is the travel time along the ray, $\mathrm{d}x_i/\mathrm{d}\tau$ represent the Cartesian components $\mathcal{U}_i$ of the *ray-velocity vector* $\boldsymbol{\mathcal{U}}$ of the m-th wave:

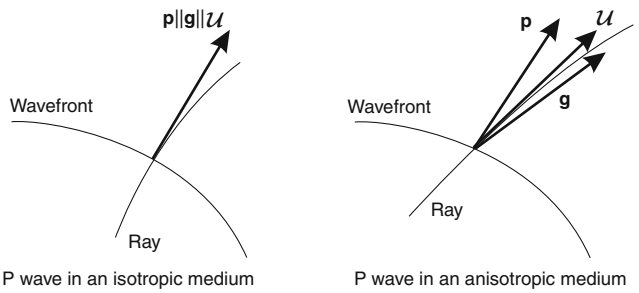
$$\mathcal{U}_i = a_{ijkl} p_l g_j^{(m)} g_k^{(m)}. \tag{28}$$

In non-dissipative anisotropic media, the ray-velocity vector $\boldsymbol{\mathcal{U}}$ is also called the group-velocity vector or the energy-velocity vector. As indicated by the name, the energy velocity vector $\boldsymbol{\mathcal{U}}$ represents the velocity of the energy propagation.

In anisotropic media, the ray-velocity vector $\boldsymbol{\mathcal{U}}$ must be strictly distinguished from the phase-velocity vector $\boldsymbol{\mathcal{C}}$. In inhomogeneous anisotropic media, the ray-velocity and phase-velocity vectors $\boldsymbol{\mathcal{U}}$ and $\boldsymbol{\mathcal{C}}$ are generally different, both in size and direction. Vector $\boldsymbol{\mathcal{U}}$ is always greater than $\boldsymbol{\mathcal{C}}$. The two vectors are equal (in size and direction) only in special directions, called *longitudinal directions*.

In inhomogeneous isotropic media, Equation 28 for the ray-velocity vector yields $\boldsymbol{\mathcal{U}} = V^2 \mathbf{p}$. For the phase-velocity vector, using Equation 17, we get $\boldsymbol{\mathcal{C}} = V^2 \mathbf{p}$. In both cases, $V = \alpha$ for P waves, and $V = \beta$ for S waves. Thus, the ray-velocity and phase-velocity vectors are identical in isotropic media.

Figure 1 shows mutual orientation of ray-velocity vector $\boldsymbol{\mathcal{U}}$, phase-velocity vector $\boldsymbol{\mathcal{C}}$ (parallel to slowness vector $\mathbf{p}$) and polarization vector $\mathbf{g}$ of a P wave propagating in an isotropic (left) and anisotropic (right) medium. While $\boldsymbol{\mathcal{U}}$, $\boldsymbol{\mathcal{C}}$ and $\mathbf{g}$ are parallel in isotropic media, they generally differ in anisotropic media. For S waves, the vectors $\boldsymbol{\mathcal{U}}$ and $\boldsymbol{\mathcal{C}}$ have similar orientation as in the case of P waves. The polarization vectors $\mathbf{g}$ are, however, perpendicular



Seismic, Ray Theory, Figure 1 Slowness vector $\mathbf{p}$ (perpendicular to the wavefront), ray-velocity vector $\boldsymbol{\mathcal{U}}$ (tangent to the ray) and polarization vector $\mathbf{g}$ of a P wave propagating in an isotropic (*left*) and anisotropic (*right*) medium. For simplicity, the three vectors in the right-hand plot are shown in one plane. In general, this is not the case in anisotropic media.

(isotropic medium) or nearly perpendicular (anisotropic medium) to the ray.

Ray tracing systems (26) and (27) can be simply solved if the initial values x_{0i} and p_{0i} are specified at some point S. We then speak of *initial-value ray tracing*. The standard numerical procedures of solving the system of ordinary differential equations of the first order with specified initial conditions can then be used (Runge-Kutta, etc.). A very important role in seismology is played by *boundary-value ray tracing*, in which we seek the ray, satisfying some boundary conditions. The typical boundary-value problem is *two-point ray tracing*, in which we seek the ray connecting two specified points. Mostly, the controlled initial-value ray tracing (controlled shooting method) is used to solve this problem (Červený et al., 2007). Boundary-value ray tracing is considerably more complicated than initial-value ray tracing.

There are four important differences between initial-value ray tracing in isotropic and anisotropic media. First: In anisotropic media, we deal with three waves, P, S1 and S2, in isotropic media with two waves, P and S, only. Second: In inhomogeneous anisotropic media, ray tracing system (26) is the same for all three waves. The wave under consideration is specified by the initial conditions, which must satisfy the eikonal equation of the considered wave. In isotropic inhomogeneous media, the ray tracing systems are different for P and S waves, see Equations 27 with $V = \alpha$ and $V = \beta$, respectively. Third: In isotropic media, the initial direction of the slowness vector specifies directly the initial direction of the ray (as the tangent to the ray and the slowness vector have the same directions). In anisotropic media, the direction of the ray is, generally, different from the direction of the slowness vector. Nevertheless, we have to use p_{0i} as the initial values for the ray tracing system. The ray-velocity vector $\boldsymbol{\mathcal{U}}$ can be simply calculated from slowness vector $\mathbf{p}$ at any point of the ray, including the initial point. Fourth: Ray tracing for P and S waves is regular everywhere in inhomogeneous isotropic media. In anisotropic media, problems arise with tracing S-wave rays in vicinities of singular directions, or if medium is nearly isotropic (quasi-isotropic).

The problem of ray tracing and travel-time computation in inhomogeneous media has been broadly discussed in the seismological literature; particularly for inhomogeneous isotropic media. Many ray tracing systems and many suitable numerical procedures for performing ray tracing have been proposed. For 1-D isotropic media (vertically inhomogeneous, radially symmetric), the ray tracing systems may be simplified so that they reduce to simple quadratures, well known from classical seismological textbooks (Aki and Richards, 1980). Standard programs for ray tracing and travel-time computations in laterally varying isotropic and anisotropic structures are available, see, for example, program packages SEIS (2D isotropic models), CRT and ANRAY (3D isotropic/anisotropic models) at http://sw3d.cz/. Programs for anisotropic media have, however, problems with S-wave computations in quasi-isotropic media and in the vicinities of shear-wave singularities. In such cases, the standard ray theory should be replaced by the coupling ray theory. Numerical procedures based on the coupling ray theory are, unfortunately, rare.

Ray tracing may also serve as a basis for the so-called *wavefront construction method* (Gjøystdal et al., 2002). In this case, for a selected wave, wavefronts with travel times $T = T_0 + k\Delta T$ are computed successively from the previous wavefronts with travel times $T = T_0 + (k-1)\Delta T$. The wavefront construction method has found broad applications in seismic exploration.

Let us consider a two-parametric system of rays, call it the *ray field*, and specify the individual rays in the ray field by *ray parameters* γ_1, γ_2. Ray parameters γ_1, γ_2 may represent, e.g., the take-off angles at a point source, or the curvilinear Gaussian coordinates of initial ray points along the initial surface. The family of rays with ray parameters within the limit $[\gamma_1, \gamma_1 + \mathrm{d}\gamma_1]$, $[\gamma_2, \gamma_2 + \mathrm{d}\gamma_2]$, is called the *elementary ray tube* or briefly the ray tube. We further introduce *ray coordinates* $\gamma_1, \gamma_2, \gamma_3$ in such a way that γ_1, γ_2 are ray parameters, and γ_3 is some monotonic parameter along a ray (arclength s, travel time τ, etc.). Here we consider $\gamma_3 = \tau$, but our results may be simply modified for any other monotonic parameter γ_3. We further introduce the 3×3 transformation matrix $\hat{\mathbf{Q}}$ from ray to Cartesian coordinates with elements $Q_{ij} = \partial x_i / \partial \gamma_j$. The Jacobian of transformation from ray to Cartesian coordinates, $\det \hat{\mathbf{Q}}$, can be expressed as follows:

$$\det \hat{\mathbf{Q}}(\tau) = (\partial \mathbf{x}(\tau)/\partial \gamma_1 \times \partial \mathbf{x}(\tau)/\partial \gamma_2)^T \boldsymbol{\mathcal{U}}(\tau). \qquad (29)$$

The vectorial product in Equation 29 has the direction of the normal to the wavefront, specified by $\mathbf{n} = \mathcal{C}\,\mathbf{p}$. As $\mathbf{p}(\tau) \cdot \boldsymbol{\mathcal{U}}(\tau) = 1$, see Equations 28, 10, and 15, we also obtain

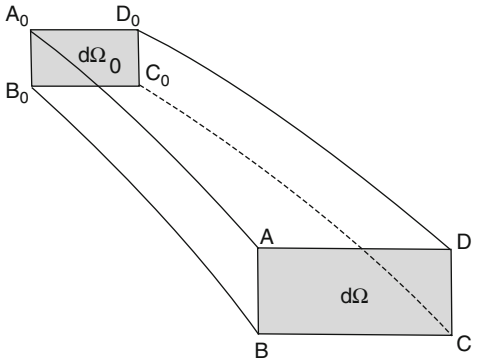
$$\det \hat{\mathbf{Q}}(\tau) = \pm \mathcal{C}(\tau) |(\partial \mathbf{x}(\tau)/\partial \gamma_1 \times \partial \mathbf{x}(\tau)/\partial \gamma_2)|. \qquad (30)$$

Thus Jacobian $\det \hat{\mathbf{Q}}(\tau)$ equals $\pm \mathcal{C}(\tau) \mathrm{d}\Omega(\tau)$, where $\mathrm{d}\Omega(\tau) = |(\partial \mathbf{x}(\tau)/\partial \gamma_1 \times \partial \mathbf{x}(\tau)/\partial \gamma_2)|$ is the scalar surface element cut out of the wavefront by the ray tube. It measures the expansion or contraction of the ray tube, see Figure 2. For this reason, the 3×3 matrix $\hat{\mathbf{Q}}(\tau)$ is also often called the *geometrical spreading matrix* and various quantities related to $\det \hat{\mathbf{Q}}(\tau)$ are called *geometrical spreading*. It plays an important role in the computation of the ray-theory amplitudes.

Transport equation. Computation of ray-theory amplitudes

The second equation of the basic system of equations of the ray method yields the transport equation for the scalar ray-theory amplitude $A(x_i)$. The transport equation is a partial differential equation of the first order. It can be expressed in several forms. One of them, valid both for isotropic and anisotropic media, reads

$$\nabla \cdot (\rho A^2 \boldsymbol{\mathcal{U}}) = 0. \qquad (31)$$



Seismic, Ray Theory, Figure 2 Elementary ray tube. $d\Omega_0$ and $d\Omega$ are scalar surface elements cut out of the wavefront by the ray tube. This means that in isotropic media, the normals to $d\Omega_0$ and $d\Omega$ are parallel to rays. In anisotropic media, they are not.

It is common to solve the transport equation along the ray. $\nabla \cdot \boldsymbol{\mathcal{U}}$ can then be expressed as follows:

$$\nabla \cdot \boldsymbol{\mathcal{U}} = \mathrm{d}[\ln(\det \hat{\mathbf{Q}})]/\mathrm{d}\tau \tag{32}$$

(Červený, 2001, Equation 3.10.24). Inserting Equation 32 into Equation 31 yields the transport equation in the form of the first-order ordinary differential equation along the ray:

$$\mathrm{d}\big(\rho(\tau)A^2(\tau)\det \hat{\mathbf{Q}}(\tau)\big)/\mathrm{d}\tau = 0. \tag{33}$$

This yields a simple form of the continuation relation for $A(\tau)$ along the ray:

$$A(\tau) = \left[\frac{\rho(\tau_0)\det \hat{\mathbf{Q}}(\tau_0)}{\rho(\tau)\det \hat{\mathbf{Q}}(\tau)}\right]^{1/2} A(\tau_0). \tag{34}$$

We obtain another suitable continuation relation for amplitudes along the ray by introducing a special local Cartesian coordinate system y_1, y_2, y_3, varying along the ray. We call it the *wavefront orthonormal coordinate system*. At any point of the ray specified by $\gamma_3 = \tau$, the y_3 axis is parallel to slowness vector $\mathbf{p}$, and the y_1, y_2 axes are confined to the plane tangential to the wavefront at $\gamma_3 = \tau$. Axes y_1 and y_2 are mutually perpendicular. If we denote the 3×3 transformation matrix from ray coordinates to wavefront orthonormal coordinates by $\hat{\mathbf{Q}}^{(y)}$, then

$$\det \hat{\mathbf{Q}}(\tau) = \det \hat{\mathbf{Q}}^{(y)}(\tau) = \mathcal{C}(\tau)\det \mathbf{Q}^{(y)}(\tau). \tag{35}$$

Here $\mathcal{C}(\tau)$ is the phase velocity, and $\mathbf{Q}^{(y)}(\tau)$ is the 2×2 upper-left submatrix of $\hat{\mathbf{Q}}^{(y)}(\tau)$. Using Equation 35 in Equation 34, we obtain the continuation relation in an alternative form:

$$A(\tau) = \left[\frac{\rho(\tau_0)\mathcal{C}(\tau_0)\det \mathbf{Q}^{(y)}(\tau_0)}{\rho(\tau)\mathcal{C}(\tau)\det \mathbf{Q}^{(y)}(\tau)}\right]^{1/2} A(\tau_0). \tag{36}$$

An important property of continuation relation (36) is that $\det \mathbf{Q}^{(y)}(\tau)$ is uniquely determined by coordinates y_1 and y_2, confined to the plane tangential to the wavefront at τ. Thus, Equation 36 remains valid for any coordinate systems q_i (even nonorthogonal), in which mutually perpendicular coordinate axes q_1 and q_2 are confined to the plane tangential to the wavefront, but the axis q_3 is taken in a different way than y_3, for example along the ray. This is, e.g., the case of the well-known *ray-centered coordinate system* q_1, q_2, q_3. We have $\det \mathbf{Q}^{(q)}(\tau) = \det \mathbf{Q}^{(y)}(\tau)$.

Transport equations for P and S waves in isotropic media may be also expressed in the form of Equation 31. The expression is straightforward for P waves. For S waves, transport equations for scalar amplitudes B and C in Equation 20 are generally coupled. They decouple only if the unit vectors $\mathbf{g}^{(1)}$ and $\mathbf{g}^{(2)}$ in Equation 20 satisfy the following relation along the ray:

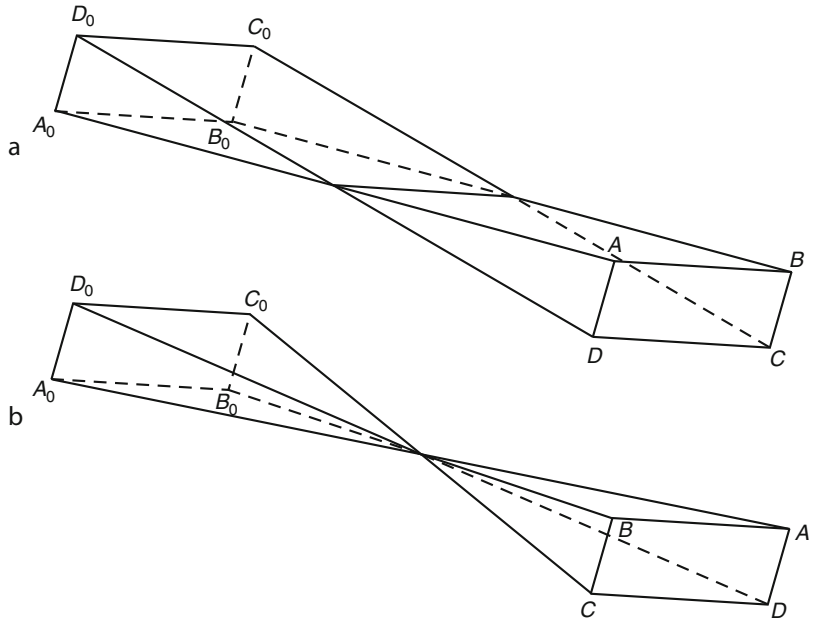
$$\mathrm{d}\mathbf{g}^{(M)}/\mathrm{d}\tau = (\mathbf{g}^{(M)} \cdot \nabla\beta)\mathbf{n}, \quad M = 1, 2. \tag{37}$$

In the terminology of the Riemanian geometry, vector $\mathbf{g}^{(M)}$ satisfying Equation 37 is transported parallelly along the ray. If $\mathbf{g}^{(1)}$ and $\mathbf{g}^{(2)}$ are chosen as mutually perpendicular and perpendicular to $\mathbf{n}$ at one point of the ray, Equation 37 guarantees that they have these properties at any point of the ray. Consequently, $\mathbf{g}^{(1)}$ and $\mathbf{g}^{(2)}$ are always perpendicular to the ray and do not rotate around it as the S wave progresses. As $\mathbf{g}^{(1)}$, $\mathbf{g}^{(2)}$ and $\mathbf{n}$ are always orthonormal, and $\mathbf{n}$ is known at any point of the ray, it is not necessary to use Equation 37 to compute both vectors $\mathbf{g}^{(M)}$. One of them can be determined from the orthonormality condition, once the other has been computed using Equation 37.

Quantity $\det \hat{\mathbf{Q}}(\tau)$ in Equation 34 may be zero at some point $\tau = \tau^C$. This means that the cross-sectional area of the ray tube shrinks to zero at $\tau = \tau^C$. The relevant point $\tau = \tau^C$ of the ray is called the *caustic point*. At the caustic point, the ray solution is singular and yields an infinite amplitude there. In passing through the caustic point τ^C along the ray, the argument of $[\det \hat{\mathbf{Q}}(\tau)]^{1/2}$ may change by $\pm\pi/2$ or $\pm\pi$ (Kravtsov and Orlov, 1999). The former case corresponds to the caustic point of the first order, see Figure 3a, during which the ray tube shrinks to an elementary arc, the latter case corresponds to the caustic point of the second order, see Figure 3b, during which the ray tube shrinks to a point. It is common to introduce the *phase shift due to caustic* $T^C(\tau, \tau_0)$ using the relation

$$\left[\frac{\det \hat{\mathbf{Q}}(\tau_0)}{\det \hat{\mathbf{Q}}(\tau)}\right]^{1/2} = \left|\frac{\det \hat{\mathbf{Q}}(\tau_0)}{\det \hat{\mathbf{Q}}(\tau)}\right|^{1/2} \exp[\mathrm{i}T^C(\tau, \tau_0)] \tag{38}$$

if caustic point τ^C is situated between τ_0 and τ. The phase shift due to the caustic is cumulative. If the ray passes through several caustic points along the ray between τ_0 and τ, the phase shift due to caustics is the sum of the individual phase shifts. It is often expressed in the form $T^C(\tau, \tau_0) = -\frac{1}{2}\pi k(\tau, \tau_0)$, where $k(\tau, \tau_0)$ is an integer,



Seismic, Ray Theory, Figure 3 Caustic points of (**a**) the first order and (**b**) second order. (Figure 3.13 of Červený, 2001.)

called the KMAH index (to acknowledge the work by - Keller, Maslov, Arnold and Hörmander in this field). The continuation relation for ray-theory amplitudes (34) can then be modified to read:

$$A(\tau) = \left(\frac{\rho(\tau_0)|\det \hat{\mathbf{Q}}(\tau_0)|}{\rho(\tau)|\det \hat{\mathbf{Q}}(\tau)|} \right)^{1/2} \exp[\mathrm{i}T^C(\tau, \tau_0)] A(\tau_0). \tag{39}$$

Equation 36 can be transformed to the analogous form as Equation 39 as the zeros of $\det \mathbf{Q}^{(\gamma)}(\tau)$ are situated at the same points τ^C on the ray as the zeros of $\det \hat{\mathbf{Q}}(\tau)$.

The KMAH index can be calculated along the ray as a byproduct of dynamic ray tracing. For detailed derivations and discussion see Bakker (1998) and Klimeš (2010).

There are some differences between the KMAH indices along the rays in isotropic and anisotropic media. In isotropic media, the KMAH index always increases when the ray passes through a new caustic point, either by one or two. In anisotropic media, however, it may also decrease by one or two at some caustic points. This happens only for S waves as a consequence of the concave form of the slowness surface of the corresponding S wave.

Dynamic ray tracing. Paraxial approximations

As we can see in Equation 34, the computation of the ray-theory amplitudes requires knowledge of $\det \hat{\mathbf{Q}}$, where $\hat{\mathbf{Q}}(\tau)$ characterizes the properties of the ray field in the vicinity of the ray under consideration. $\hat{\mathbf{Q}}(\tau)$ can be computed by the procedure called *dynamic* (or paraxial) *ray tracing*. In addition to $\hat{\mathbf{Q}}(\tau)$ with elements $Q_{ij}(\tau) = \partial x_i / \partial \gamma_j$, we also have to introduce a new 3×3 matrix $\hat{\mathbf{P}}(\tau)$ with elements $P_{ij}(\tau) = \partial p_i / \partial \gamma_j$. The equation for P_{ij} must be included to obtain the linear dynamic ray tracing system. Differentiating ray tracing equations (25) with respect to γ_j, we can easily obtain a system of linear ordinary differential equations of the first order for Q_{ij} and P_{ij},

$$\begin{aligned} \frac{\mathrm{d}Q_{ij}}{\mathrm{d}\tau} &= \frac{\partial^2 \mathcal{H}}{\partial p_i \partial x_k} Q_{kj} + \frac{\partial^2 \mathcal{H}}{\partial p_i \partial p_k} P_{kj}, \\ \frac{\mathrm{d}P_{ij}}{\mathrm{d}\tau} &= -\frac{\partial^2 \mathcal{H}}{\partial x_i \partial x_k} Q_{kj} - \frac{\partial^2 \mathcal{H}}{\partial x_i \partial p_k} P_{kj}, \end{aligned} \tag{40}$$

see Červený (1972). This system is usually called the *dynamic ray tracing system*, and the relevant procedure dynamic ray tracing. It can be solved along a given ray Ω, or together with it.

The dynamic ray tracing system (40) may be expressed in various forms. Instead of Cartesian coordinates x_i, we can use the wavefront orthonormal coordinates y_i, or the ray-centered coordinates q_i. Then, instead of the 3×3 matrices $\hat{\mathbf{Q}}$ and $\hat{\mathbf{P}}$, it is sufficient to seek the 2×2 matrices $\mathbf{Q}^{(y)}$, $\mathbf{P}^{(y)}$ or $\mathbf{Q}^{(q)}$, $\mathbf{P}^{(q)}$. This reduces the number of DRT equations, but complicates their right-hand sides (Červený 2001, sect. 4.2).

As the dynamic ray tracing system (40) is of the first order and linear, we can compute its fundamental matrix consisting of six linearly independent solutions. The 6×6 fundamental matrix of system (40) specified by the 6×6 identity matrix at an arbitrary point $\tau = \tau_0$ of the ray is called the *ray propagator matrix* and denoted by $\mathbf{\Pi}(\tau, \tau_0)$.

The 6×6 ray propagator matrix $\mathbf{\Pi}(\tau, \tau_0)$ is symplectic:

$$\mathbf{\Pi}^T(\tau, \tau_0)\mathbf{J}\mathbf{\Pi}(\tau, \tau_0) = \mathbf{J}, \quad \text{with } \mathbf{J} = \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{0} \end{pmatrix} \tag{41}$$

If we know the matrices $\hat{\mathbf{Q}}(\tau_0)$, $\hat{\mathbf{P}}(\tau_0)$, we can compute $\hat{\mathbf{Q}}(\tau)$, $\hat{\mathbf{P}}(\tau)$ at any point τ of the ray by a simple matrix multiplication

$$\begin{pmatrix} \hat{\mathbf{Q}}(\tau) \\ \hat{\mathbf{P}}(\tau) \end{pmatrix} = \mathbf{\Pi}(\tau, \tau_0) \begin{pmatrix} \hat{\mathbf{Q}}(\tau_0) \\ \hat{\mathbf{P}}(\tau_0) \end{pmatrix}. \tag{42}$$

The ray propagator matrix $\mathbf{\Pi}(\tau, \tau_0)$ satisfies the chain rule, $\mathbf{\Pi}(\tau, \tau_0) = \mathbf{\Pi}(\tau, \tau_1)\mathbf{\Pi}(\tau_1, \tau_0)$, where point τ_1 is situated arbitrarily on the ray. It is simple to compute the inverse of $\mathbf{\Pi}(\tau, \tau_0)$: $\mathbf{\Pi}^{-1}(\tau, \tau_0) = \mathbf{\Pi}(\tau_0, \tau)$. We can express $\mathbf{\Pi}(\tau, \tau_0)$ in the following way:

$$\mathbf{\Pi}(\tau, \tau_0) = \begin{pmatrix} \hat{\mathbf{Q}}_1(\tau, \tau_0) & \hat{\mathbf{Q}}_2(\tau, \tau_0) \\ \hat{\mathbf{P}}_1(\tau, \tau_0) & \hat{\mathbf{P}}_2(\tau, \tau_0) \end{pmatrix}, \tag{43}$$

where $\hat{\mathbf{Q}}_1(\tau, \tau_0), \hat{\mathbf{Q}}_2(\tau, \tau_0), \hat{\mathbf{P}}_1(\tau, \tau_0)$ and $\hat{\mathbf{P}}_2(\tau, \tau_0)$ are 3×3 matrices.

Equation 42 can be used to obtain a very important quantity – the 3×3 matrix $\hat{\mathbf{M}}(\tau)$ of second derivatives of the travel-time field with respect to Cartesian coordinates, with elements $M_{ij} = \partial^2 T / \partial x_i \partial x_j$:

$$\hat{\mathbf{M}}(\tau) = \hat{\mathbf{P}}(\tau)(\hat{\mathbf{Q}}(\tau))^{-1}. \tag{44}$$

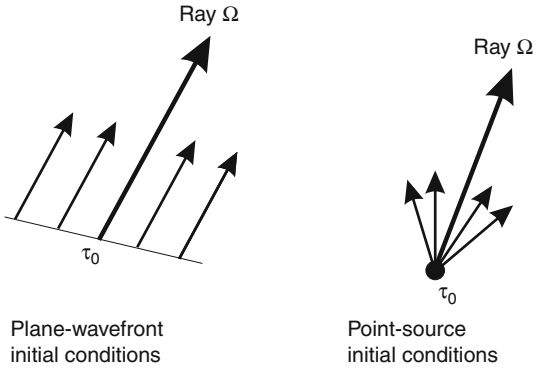
Matrix $\hat{\mathbf{M}}(\tau)$ plays an important role in the computation of travel time not only along the ray Ω, but also in its "quadratic" paraxial vicinity:

$$T(\mathbf{x}) = T(\mathbf{x}^\Omega) + (\mathbf{x} - \mathbf{x}^\Omega)^T \mathbf{p}(\tau) + \frac{1}{2}(\mathbf{x} - \mathbf{x}^\Omega)^T \hat{\mathbf{M}}(\tau)(\mathbf{x} - \mathbf{x}^\Omega). \tag{45}$$

In Equation 45, $\mathbf{x}$ denotes an arbitrary point in the paraxial vicinity of the ray Ω, close to point $\mathbf{x}^\Omega = \mathbf{x}^\Omega(\tau)$ on the ray Ω; slowness vector $\mathbf{p}(\tau)$ and the matrix $\hat{\mathbf{M}}(\tau)$ are given at $\mathbf{x}^\Omega$. The possibility of computing the travel time in the paraxial vicinity of the ray has many important applications.

The properties of the 6×6 ray propagator matrix $\mathbf{\Pi}(\tau, \tau_0)$ described above remain valid even for the 4×4 ray propagator matrices $\mathbf{\Pi}^{(y)}(\tau, \tau_0)$ or $\mathbf{\Pi}^{(q)}(\tau, \tau_0)$ expressed in wavefront orthonormal coordinates y_i or ray-centered coordinates q_i. The ray propagator matrices $\mathbf{\Pi}^{(y)}(\tau, \tau_0)$ and $\mathbf{\Pi}^{(q)}(\tau, \tau_0)$ are identical, therefore, they can be expressed in terms of the same 2×2 matrices $\mathbf{Q}_1(\tau, \tau_0)$, $\mathbf{Q}_2(\tau, \tau_0)$, $\mathbf{P}_1(\tau, \tau_0)$ and $\mathbf{P}_2(\tau, \tau_0)$. Matrices $\mathbf{Q}_1(\tau, \tau_0), \mathbf{P}_1(\tau, \tau_0)$ correspond to the plane-wavefront initial conditions at τ_0, and matrices $\mathbf{Q}_2(\tau, \tau_0), \mathbf{P}_2(\tau, \tau_0)$ to the point-source initial conditions at τ_0, see Figure 4. The 2×2 matrix $\mathbf{Q}_2(\tau, \tau_0)$ plays an important role in computing the ray-theory Green function. The quantity

$$\mathcal{L}(\tau, \tau_0) = |\det \mathbf{Q}_2(\tau, \tau_0)|^{1/2} \tag{46}$$



Seismic, Ray Theory, Figure 4 Plane-wavefront and point-source initial conditions for dynamic ray tracing. In anisotropic media, rays are not perpendicular to the wavefront.

is called the *relative geometrical spreading*. It corresponds to a point source.

As in Equation 44, we can define the 2×2 matrix of the second derivatives of the travel-time field with respect to y_1, y_2 or q_1, q_2 as follows:

$$\mathbf{M}(\tau) = \mathbf{P}(\tau)(\mathbf{Q}(\tau))^{-1}. \tag{47}$$

We will now briefly summarize several useful ray-theory quantities and applications, which rely fully or partly on dynamic ray tracing. For derivations and more detailed explanations, see Červený (2001, Chap. 4), where also many other applications and references can be found: (1) Paraxial travel times. (2) Paraxial slowness vectors. (3) Paraxial rays. (4) Curvature of the wavefront. (5) Matrix of geometrical spreading $\hat{\mathbf{Q}}$ and the relevant matrix $\hat{\mathbf{P}}$. (6) Continuation relations for ray-theory amplitudes along the ray. (7) Relative geometrical spreading. (8) Phase shift due to caustics. (9) Ray-theory elastodynamic Green function. (10) Higher-order spatial derivatives of the travel-time field. (11) Fresnel volumes and Fresnel zones. (12) Surface-to-surface propagator matrix. (13) Boundary-value problems in four-parametric system of paraxial rays, including two-point ray tracing. (14) Factorization of the geometrical spreading.

Dynamic ray tracing is also needed in the investigation of ray chaos and in computations of Lyapunov exponents, in the ray-perturbation methods and in modifications and extensions of the ray method such as Maslov method, Gaussian beam and Gaussian packet summation methods, in Kirchhoff-Helmholtz method and in various diffraction methods.

Coupling ray theory for S waves in anisotropic media

In inhomogeneous weakly anisotropic media, the standard ray theory described above yields distorted results since it is unable to describe the coupling of S1 and S2 waves propagating with approximately equal phase velocities. This problem can be removed by using the coupling ray

theory. In the coupling ray theory, the amplitudes of the two S waves can be computed along a trajectory called the *common ray* (Bakker, 2002; Klimeš, 2006). The closer the common ray approximates actual S-wave rays, the more accurate results the coupling ray theory yields. The common rays can be constructed in a reference isotropic medium or in the actual anisotropic medium. A convenient option is to compute common rays using ray tracing equations (25) with the Hamiltonian given as

$$\mathcal{H}(x_i, p_j) = \frac{1}{4}[G_1(x_i, p_j) + G_2(x_i, p_j)]. \tag{48}$$

In Equation 48, G_1 and G_2 are eigenvalues of the Christoffel matrix $\mathbf{\Gamma}$, see equation (7), corresponding to S1 and S2 waves.

The coupling ray theory solution is sought in the form (Coates and Chapman, 1990; Bulant and Klimeš, 2002):

$$\begin{aligned}\mathbf{u}(\tau, t) = A(\tau)[r_1(\tau)\mathbf{g}^{(1)}(\tau)\exp(\mathrm{i}\omega\tau_1) \\ + r_2(\tau)\mathbf{g}^{(2)}(\tau)exp(\mathrm{i}\omega\tau_2)]\exp(-\mathrm{i}\omega t).\end{aligned} \tag{49}$$

Here, $A(\tau)$ is the scalar amplitude (34) or (36) calculated along the common ray. The symbols $\mathbf{g}^{(1)}$ and $\mathbf{g}^{(2)}$ denote the S-wave eigenvectors of the generalized Christoffel matrix $\Gamma(x_i, p_j)$ calculated along the common ray. The travel times τ_1 and τ_2 are travel times corresponding to the above vectors $\mathbf{g}^{(1)}$ and $\mathbf{g}^{(2)}$. They can be obtained by quadratures along the common ray:

$$\begin{aligned}\mathrm{d}\tau_1/\mathrm{d}\tau &= [\Gamma_{ik} g_i^{(1)} g_k^{(1)}]^{-1/2}, \\ \mathrm{d}\tau_2/\mathrm{d}\tau &= [\Gamma_{ik} g_i^{(2)} g_k^{(2)}]^{-1/2}.\end{aligned} \tag{50}$$

The amplitude factors r_1 and r_2 are solutions of two coupled ordinary differential equations (Coates and Chapman, 1990):

$$\begin{pmatrix}\mathrm{d}r_1/\mathrm{d}\tau \\ \mathrm{d}r_2/\mathrm{d}\tau\end{pmatrix} = \frac{\mathrm{d}\varphi}{\mathrm{d}\tau}\begin{pmatrix}0 & \exp(\mathrm{i}\omega[\tau_2(\tau)-\tau_1(\tau)]) \\ -\exp(\mathrm{i}\omega[\tau_1(\tau)-\tau_2(\tau)]) & 0\end{pmatrix}\begin{pmatrix}r_1 \\ r_2\end{pmatrix}, \tag{51}$$

where the angular velocity $\mathrm{d}\varphi/\mathrm{d}\tau$ of the rotation of the eigenvectors $\mathbf{g}^{(1)}$ and $\mathbf{g}^{(2)}$ is given by

$$\frac{\mathrm{d}\varphi}{\mathrm{d}\tau} = \mathbf{g}^{(2)}\frac{\mathrm{d}\mathbf{g}^{(1)}}{\mathrm{d}\tau} = -\mathbf{g}^{(1)}\frac{\mathrm{d}\mathbf{g}^{(2)}}{\mathrm{d}\tau}. \tag{52}$$

For detailed description of the algorithm, see Bulant and Klimeš (2002).

There are many possible modifications and approximations of the coupling ray theory. In some of them, the amplitude vector $\mathbf{U}$ of coupled S waves is sought along the common ray in the form of Equation 20, in which the amplitude factors B and C can be expressed as

$$B(\tau) = A(\tau)\mathcal{B}(\tau) \qquad C(\tau) = A(\tau)\mathcal{C}(\tau). \tag{53}$$

In Equations 53, $A(\tau)$ is again the scalar ray amplitude, see equation (34) or (36), calculated along the common S-wave ray. There are many ways how to evaluate factors $\mathcal{B}$ and $\mathcal{C}$ (Kravtsov, 1968; Pšenčík, 1998; Červený et al., 2007). Here we present a combination of coupling ray theory and of the first-order ray tracing (Farra and Pšenčík, 2010). In the approximation of Farra and Pšenčík (2010), the common ray is obtained as the first-order ray, see section on ray perturbation methods. The vectors $\mathbf{g}^{(K)}$, appearing in Equation 20, specify the first-order approximation of the S-wave polarization plane. The factors $\mathcal{B}$ and $\mathcal{C}$ in Equations 53 are then obtained as a solution of two coupled ordinary differential equations, which result from the corresponding two coupled transport equations:

$$\begin{pmatrix}\mathrm{d}\mathcal{B}/\mathrm{d}\tau \\ \mathrm{d}\mathcal{C}/\mathrm{d}\tau\end{pmatrix} = -\frac{\mathrm{i}\omega}{2}\begin{pmatrix}\mathcal{M}_{11}-1 & \mathcal{M}_{12} \\ \mathcal{M}_{12} & \mathcal{M}_{22}-1\end{pmatrix}\begin{pmatrix}\mathcal{B} \\ \mathcal{C}\end{pmatrix}. \tag{54}$$

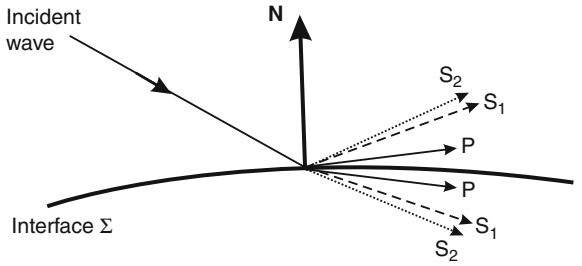
Evaluation of the matrix $\boldsymbol{\mathcal{M}}$ with elements $\mathcal{M}_{IJ}$ is simple (see Farra and Pšenčík (2010); Equations 20 and 7).

The resulting equations reduce to standard ray-theory equations in inhomogeneous isotropic media, they describe properly S-wave coupling in inhomogeneous weakly anisotropic media and even yield separate S waves when anisotropy is stronger. Common S-wave rays are regular everywhere. They do not suffer from problems well known from tracing rays of individual S waves in anisotropic media and are suitable for investigating shear-wave splitting.

Effects of structural interfaces

Assume that the ray is incident on a curved structural interface. If we wish to continue the ray computations for the reflected, transmitted, monotypic or converted waves, see Figure 5, we have to use relevant transformation relations for the ray tracing system, dynamic ray tracing system and for the ray theory amplitudes at the interface.

The transformation relations for ray tracing and dynamic ray tracing systems at interfaces are well known (Červený, 2001). For the sake of brevity, we do not present them here. We shall, however, discuss the transformation of amplitudes. In the zero-order ray approximation, the transformation of ray-theory amplitudes across an interface is described by plane-wave reflection/transmission coefficients. In other words, amplitudes of generated waves do not depend on the curvature of the wavefront of the incident wave and the curvature of the interface at the point of incidence Q. Neither do they depend on the gradients of the density and on gradients of the density-normalized elastic moduli at Q, on both sides of the interface. They depend only on the local values of the density and density-normalized elastic moduli at Q (on both sides of the interface) and on the angle of incidence (the acute angle between the slowness vector of the incident wave and the normal to the interface $\mathbf{N}$ at the point of incidence Q).



Seismic, Ray Theory, Figure 5 Slowness vectors of P, S1 and S2 waves generated at the point of incidence Q of a curved interface separating two inhomogeneous anisotropic media. All slowness vectors at Q are situated in the plane of incidence specified by the slowness vector of the incident wave and the normal to the interface **N** at Q. Ray-velocity vectors (tangent to rays) of individual waves at Q are, in general, not confined to the plane of incidence. In isotropic media, instead of reflected and transmitted S1 and S2 waves, single reflected and transmitted S waves are generated. In inhomogeneous weakly anisotropic media, single coupled S waves are generated. Ray-velocity vectors of individual waves at Q are situated in the plane of incidence in isotropic media.

Various types of R/T coefficients may be used. The displacement R/T coefficients are used most frequently (Aki and Richards, 1980; Červený et al., 1977 for isotropic media; Fedorov, 1968 for anisotropic media). Very useful are the energy R/T coefficients, as they are reciprocal. The relation between the energy R/T coefficient $\mathcal{R}(Q)$ and the displacement R/T coefficient $R(Q)$ is as follows:

$$\mathcal{R}(Q) = R(Q)\left[\frac{\rho(\tilde{Q})\mathcal{U}_n(\tilde{Q})}{\rho(Q)\mathcal{U}_n(Q)}\right]^{1/2} \tag{55}$$

(Červený 2001, sect. 5.4.3). Here Q is the point of incidence, and $\tilde{Q}$ the relevant initial point of the R/T wave, both points being, of course, identical. $\mathcal{U}_n$ is the normal component (perpendicular to the interface) of the ray-velocity vector. We further introduce the complete energy R/T coefficients $\mathcal{R}^C$ along the ray using the relation

$$\mathcal{R}^C = \prod_{k=1}^{N} \mathcal{R}(Q_k). \tag{56}$$

The complete energy R/T coefficient $\mathcal{R}^C$ corresponds to the ray which interacts N-times with interfaces (at points of incidence $Q_1, Q_2, ..., Q_N$) between the initial and end point of the ray.

Generalization of the continuation relation (36) for the ray-theory amplitudes along the ray situated in a laterally varying anisotropic medium containing curved interfaces then reads:

$$A(\tau) = \left(\frac{\rho(\tau_0)\mathcal{C}(\tau_0)|\det \mathbf{Q}^{(y)}(\tau_0)|}{\rho(\tau)\mathcal{C}(\tau)|\det \mathbf{Q}^{(y)}(\tau)|}\right)^{1/2} \mathcal{R}^C \exp[\mathrm{i}T^C(\tau,\tau_0)]A(\tau_0). \tag{57}$$

In seismic prospecting, in the technique called *amplitude variation with offset* (AVO), it is common to work with the so-called *weak-contrast R/T coefficients*. They are linearized versions of exact R/T displacement coefficients. Linearization is mostly made with respect to the contrasts of the density and elastic moduli across the interface. There is a great variety of linearized formulae depending on the type of media surrounding the interface (isotropic, anisotropic), strength of anisotropy (weak, strong), etc. The coefficients yield reasonable approximation in the vicinity of normal incidence. For increasing incidence angles, their accuracy decreases. The advantage of the weak-contrast coefficients is their simplicity and the possibility of expressing them in explicit form. The effects of the individual medium parameters on the coefficients can than be easily evaluated.

Ray-theory elastodynamic Green function

The elastodynamic Green function $G_{in}(R,t,S,t_0)$ represents the i-th Cartesian component of the displacement vector at location R and time t, due to a single-force point source situated at location S and oriented along the n-th Cartesian axis, with the time dependence $\delta(t-t_0)$. We introduce quite analogously the ray-theory elastodynamic Green function, with only two differences. The first difference is that ray-theory Green function is defined as a sum of *elementary ray-theory Green functions* computed along rays of selected elementary waves (direct, multiply reflected/transmitted, etc.). The second difference is that the elementary ray-theory Green functions are not exact, but only zero-order ray approximations.

In the frequency domain the elementary ray-theory elastodynamic Green function $G_{in}(R,S,\omega)$ for $t_0 = 0$ reads:

$$G_{in}(R,S,\omega) = \frac{g_n(S)g_i(R)\exp[\mathrm{i}T^G(R,S) + \mathrm{i}\omega T(R,S)]}{4\pi[\rho(S)\rho(R)\mathcal{C}(S)\mathcal{C}(R)]^{1/2}\mathcal{L}(R,S)}\,\mathcal{R}^C. \tag{58}$$

Here $\mathcal{L}(R,S)$ is the relative geometrical spreading, given by Equation 46, $g_i(R)$ and $g_n(S)$ are the eigenvectors of the generalized Christoffel matrix at R and S (polarization vectors corresponding to the considered elementary wave), T is the travel time along the ray from S to R, $\mathcal{R}^C$ the complete energy R/T coefficient resulting from interactions of the ray under consideration with interfaces between S and R, and $T^G(R,S)$ the complete phase shift due to caustics along the ray between S and R. The relevant KMAH index in anisotropic media may also include a contribution at a point source S (if the slowness surface of the considered wave is concave at S). In isotropic media, this contribution is always zero.

The complete energy R/T coefficient $\mathcal{R}^C$, the travel time $T(R,S)$, the relative geometrical spreading $\mathcal{L}(R,S)$ and the complete phase shift due to caustics are always reciprocal. Consequently, the elementary ray-theory

elastodynamic Green function satisfies a very important property of reciprocity:

$$G_{in}(R,S,\omega) = G_{ni}(S,R,\omega). \tag{59}$$

This relation is valid for any elementary seismic body wave generated by a point source.

For elementary ray-theory Green functions in inhomogeneous weakly anisotropic media see Pšenčík (1998).

Chaotic rays. Lyapunov exponents

In homogeneous media, geometrical spreading increases linearly with increasing length of the ray. In heterogeneous media, behavior of geometrical spreading is more complicated, and depends considerably on the degree of heterogeneity of the medium. In models, in which the heterogeneity exceeds certain degree, average geometrical spreading increases exponentially with increasing length of the ray. Rays in such a medium often exhibit chaotic behavior, which is characterized by a strong sensitivity of rays to the initial ray data (for example, to ray parameters). The rays with only slightly differing ray data at an initial point tend to diverge *exponentially* at large distances from the initial point. Consequently, the rays intersect many times and many rays pass through the same point. With such *chaotic rays*, two-point ray tracing is practically impossible, and the ray tubes are not narrow enough for travel time interpolation. The chaotic behavior of rays increases with increasing length of rays and prevents applicability of the ray theory.

The exponential divergence of chaotic rays in the phase space (the space formed by spatial coordinates x_i and slowness-vector components p_j) can be quantified by the so-called *Lyapunov exponents*. They may be introduced in several ways. It is common to express them in terms of characteristic values of the ray propagator matrix. The relevant expressions for the Lyapunov exponents and several numerical examples for 2D models without interfaces can be found in Klimeš (2002a). See also Červený et al. (2007), where other references can be found.

The estimate of the Lyapunov exponent of a single finite ray depends on its position and direction. The Lyapunov exponents associated with rays of different positions and directions can be used to calculate *average Lyapunov exponents* for the model. The average Lyapunov exponents play a very important role in smoothing the models so that they are suitable for ray tracing (Červený et al., 2007).

Ray perturbation methods

Ray perturbation methods represent an important part of the ray theory. They can be used for approximate but fast and transparent solutions of forward problems in complicated models. They play even more important role in the inverse problems.

Ray perturbation methods are useful everywhere, where we wish to compute the wavefield or its constituents (travel times, amplitudes, polarization) in complicated models, which deviate only little from simple, *reference models*, for which computations are simpler. The solutions for complicated models are then sought as perturbations of simpler solutions for the reference models. Examples are computations in weakly anisotropic media, which use an isotropic medium as reference, or in weakly dissipative media, which use a perfectly elastic medium as reference. Basic role in these approaches is played by *reference rays* traced in reference media. Solutions in perturbed media can be given in the form of a power series in the deviations of the perturbed and reference models. Mostly, the *first-order approximation*, i.e. the first term of the power series, is used.

The most frequent application of ray perturbation methods is in travel-time computations. First-order travel-time perturbation formulae for isotropic media are known and have been used (mostly in tomography) for several decades. Well known and broadly applied are also first-order travel-time formulae for anisotropic media (Červený and Jech, 1982; Hanyga, 1982; Červený, 2001, sect. 3.9). Travel-time perturbations are obtained by quadratures along reference rays. As integration parameters, the parameters along reference rays are used.

Recently, several procedures for computation of higher-order travel-time perturbations for weakly anisotropic media (note that anisotropy of the Earth is mostly weak) were proposed. The procedure based on the so-called perturbation Hamiltonians (Klimeš, 2002b; Červený et al., 2007) allows computation of highly accurate travel times along a fixed reference ray in a reference medium. Another procedure is based on the so-called first-order ray tracing described briefly below. In the latter method, second-order travel-time perturbations can be calculated along first-order rays.

Relatively recent is the use of ray perturbation methods in *first-order ray tracing* and *first-order dynamic ray tracing* (Pšenčík and Farra, 2007; Farra and Pšenčík, 2010). It allows to compute, approximately, not only rays and travel times, but whole wavefields. To derive first-order ray tracing and dynamic ray tracing, the perturbation approach is used in which deviations of anisotropy from isotropy are considered to be of the first order. Then it is just sufficient to use Equations 25 and 40 with Equation 21, in which the exact eigenvalue G_m is replaced by its first-order approximation. The resulting ray tracing provides first-order rays, first-order travel times and the first-order geometrical spreading. By simple quadratures along first-order rays, second-order travel-time corrections can be computed. This approach is applicable to P and S waves. In case of S waves, it can include the computation of coupling effects. First-order ray tracing and dynamic ray tracing are used in this case for computing common rays, first-order travel times and geometrical spreading along them, using the Hamiltonian (48). The wavefield of S waves is obtained by solving second-order coupling equations along the common rays. The procedure yields standard ray-theory results for S waves propagating in isotropic

media, and approximate results in anisotropic media when the S waves are coupled or even decoupled.

Ray perturbation method for weakly dissipative media

In viscoelastic media, the density-normalized stiffness tensor a_{ijkl} is complex valued:

$$a_{ijkl}(x_n) = a^R_{ijkl}(x_n) - \mathrm{i}a^I_{ijkl}(x_n). \tag{60}$$

If a^I_{ijkl} is small, the viscoelastic medium can be considered as a perturbation of a perfectly elastic medium (Červený, 2001, sect. 5.5.3) which has a form of the imaginary-valued term $-\mathrm{i}a^I_{ijkl}$. Reference ray in the reference perfectly elastic medium and corresponding real-valued travel time T along the reference ray between points S and R can be obtained by standard ray tracing in perfectly elastic media. The imaginary travel time T^I (travel-time perturbation due to $-\mathrm{i}a^I_{ijkl}$) can be then obtained by quadratures along the reference ray:

$$T^I = \frac{1}{2}\int_S^R Q^{-1}(\tau)\mathrm{d}\tau. \tag{61}$$

The quantity Q in Equation 61 is a direction-dependent *quality factor* for anisotropic media, corresponding to the Hamiltonian (21):

$$Q^{-1} = a^I_{ijkl}p_j p_l g_i g_k. \tag{62}$$

For general Hamiltonians, the quality factor Q is given by the relation $Q^{-1} = -\mathrm{Im}\mathcal{H}(x_i, p_j)$.

The imaginary travel time T^I in Equation 61 is responsible for the exponential amplitude decay along the reference ray. For causal dissipation, the stiffness tensor (60) is frequency dependent. The above described perturbation approach is then equivalent to the perturbation scheme, in which $a^I_{ijkl}(x_n, \omega)$ is considered to be of the order of ω^{-1} for $\omega \to \infty$ (Kravtsov and Orlov, 1990; Gajewski and Pšenčík, 1992).

In an inhomogeneous isotropic, weakly dissipative medium, the expression (62) reduces to the well-known formula

$$Q^{-1} = -\mathrm{Im}V^2/\mathrm{Re}V^2 \doteq -2\mathrm{Im}V/\mathrm{Re}V, \tag{63}$$

in which V is the complex-valued velocity, $V = \alpha$ for P waves and $V = \beta$ for S waves. Complex-valued quantities α and β are generalizations (to the complex space) of real-valued α and β from Equation 14.

Concluding remarks. Applications, modifications, and extensions of the ray method

Seismic ray method has found broad applications both in global seismology and in seismic exploration. The advantages of the seismic ray method consist in its numerical efficiency, universality, conceptual clarity, and in its ability to investigate various seismic body waves independently of other waves. Although its accuracy is only limited, the seismic ray method is the only method which is able to give an approximate answer to many problems of high-frequency seismic body wave propagation in laterally varying, isotropic or anisotropic, perfectly elastic or dissipative, layered and block structures.

In classical global seismology, the seismic ray method has been traditionally used to study the internal structure of the whole Earth, assuming that the Earth is radially symmetric. The standard Earth's model, obtained in this way, is expressed in terms of distribution of elastic velocities as a function of depth.

At present, the applications of the seismic ray method are considerably broader. It is broadly used to study the 3-D local lateral inhomogeneities in the structure, the form and physical properties of structural interfaces, the local anisotropy, attenuation, etc. In addition to forward modeling, the ray perturbation methods are also broadly used for inversions based on measured travel times or whole waveforms. In lithospheric structural studies, particularly in crustal seismology, the ray-synthetic seismograms have been also often used for ultimate comparison with observed seismograms. The computation of ray-synthetic seismograms requires determination of not only travel times, but also ray-theory amplitudes and polarization of individual waves. Seismic ray method has also found broad applications in other branches of seismology. Very important examples are the localization of seismic sources and the simultaneous localization with structural inversion.

In most applications of the ray method in seismic exploration for oil, the use of local 3-D structures with structural curved interfaces is a necessity. Sophisticated algorithms have been developed and used to image the structures under consideration. At present, the most important role is played by migration algorithms. Seismic ray theory and its extensions have found important applications in these algorithms.

The ray method is not valid universally. We have briefly described three serious limitations of the ray method: (a) The ray method can be used only for high-frequency signals. (b) In models, in which heterogeneity of the medium exceeds certain degree, the ray field has chaotic character, particularly at large distances from the source. (c) The standard ray method cannot be used for computing S waves propagating in inhomogeneous, weakly anisotropic media. It must be replaced by the coupling ray theory. The coupling ray theory must be used even in moderately or strongly anisotropic media, in the vicinity of shear-wave singular directions.

The ray method fails, however, even in other singular situations. In smooth isotropic media, the most important type of singularity are *caustics*. Caustics may attain various forms. Various extensions of the ray method can be used to compute wavefields in caustic regions. These extensions are frequency dependent. See a detailed treatment of wavefields in caustic regions in Kravtsov and Orlov (1999), and also in Stamnes (1986). In models with

smooth structural interfaces, other singularities often appear. For edge and vertex points, see Ayzenberg et al. (2007). For critical singular regions, at which head waves separate from reflected waves, see Červený and Ravindra (1971). For the waves, whose rays are tangential to interfaces, see Thomson (1989).

Specific methods, designed for different types of singularities may be used for computing wavefields in singular regions. Disadvantage of these methods is that they are different for different singularities. Morever, singular regions often overlap, and the wavefield in the overlaping region requires again different treatment. It is desirable to have available a more general extension of the ray method, applicable uniformly in any of the mentioned singular regions, or, at least, in most of them. Such an extension would simplify ray computations considerably and could even lead to more accurate results.

Several such extensions of the ray method have been proposed. We do not describe them here in detail. Instead, we merely present references, in which more details and further references can be found. Let us mention the Maslov asymptotic ray theory introduced to seismology by Chapman and Drummond (1982), see also Thomson and Chapman (1985), Chapman (2004). Another extension of the ray method is based on the summation of Gaussian beams (Popov, 1982; Červený et al., 1982). For the relation of this method with the Maslov method see Klimeš (1984). The Gaussian beam summation method has found applications both in the forward modeling of seismic wavefields and in migrations in seismic exploration. It is closely related to the method of summation of Gaussian packets (Červený et al., 2007). Ray theory can be also used in the Born scattering theory (Chapman and Coates, 1994; Chapman, 2004). For waves reflected from a smooth structural interface separating two heterogeneous, isotropic or anisotropic media, the Kirchhoff surface integral method can be used. For details and many references see Chapman (2004, sect. 10.4). Another useful extension of the ray method is the one-way wave equation approach (Thomson, 1999).

Acknowledgments

The authors are very grateful to Luděk Klimeš and Ravi Kumar for valuable comments and recommendations. The research was supported by the consortium project Seismic Waves in Complex 3-D Structures, by research projects 205/07/0032 and 205/08/0332 of the Grant Agency of the Czech Republic; and by research project MSM0021620860 of the Ministry of Education of the Czech Republic.

Bibliography

Aki, K., and Richards, P., 1980. *Quantitative Seismology*. San Francisco: WH Freeman.

Ayzenberg, M. A., Aizenberg, A. M., Helle, H. B., Klem-Musatov, K. D., Pajchel, J., and Ursin, B., 2007. 3D diffraction modeling of singly scattered acoustic wavefields based on the combination of surface integral propagators and transmission operators. *Geophysics*, **72**, SM19–SM34.

Babich, V. M., 1956. Ray method of the computation of the intensity of wave fronts (in Russian). *Doklady Akademii Nauk SSSR*, **110**, 355–357.

Babich, V. M., 1961. Ray method of the computation of the intensity of wave fronts in elastic inhomogeneous anisotropic medium. In Petrashen, G. I. (ed.), *Problems of the Dynamic Theory of Propagation of Seismic Waves* 77 (in Russian). Leningrad: Leningrad University Press, Vol. 5, pp. 36–46. Translation to English: Geophysical Journal International., 118: 379–383, 1994.

Bakker, P. M., 1998. Phase shift at caustics along rays in anisotropic media. *Geophysical Journal International*, **134**, 515–518.

Bakker, P. M., 2002. Coupled anisotropic shear-wave ray tracing in situations where associated slowness sheets are almost tangent. *Pure and Applied Geophysics*, **159**, 1403–1417.

Bulant, P., and Klimeš, L., 2002. Numerical algorithm of the coupling ray theory in weakly anisotropic media. *Pure and Applied Geophysics*, **159**, 1419–1435.

Červený, V., 1972. Seismic rays and ray intensities in inhomogeneous anisotropic media. *Geophysical Journal of Royal Astronomical Society*, **29**, 1–13.

Červený, V., 2001. *Seismic Ray Theory*. Cambridge: Cambridge University Press.

Červený, V., and Jech, J., 1982. Linearized solutions of kinematic problems of seismic body waves in inhomogeneous slightly anisotropic media. *Journal of Geophysics*, **51**, 96–104.

Červený, V., and Ravindra, R., 1971. *Theory of Seismic Head Waves*. Toronto: Toronto University Press.

Červený, V., Molotkov, I. A., and Pšenčík, I., 1977. *Ray Method in Seismology*. Praha: Univerzita Karlova.

Červený, V., Popov, M. M., and Pšenčík, I., 1982. Computation of wave fields in inhomogeneous media. Gaussian beam approach. *Geophysical Journal of Royal Astronomical Society*, **70**, 109–128.

Červený, V., Klimeš, L., and Pšenčík, I., 1988. Complete seismic ray tracing in three-dimensional structures. In Doornbos, D. J. (ed.), *Seismological Algorithms*. New York: Academic, pp. 89–168.

Červený, V., Klimeš, L., and Pšenčík, I., 2007. Seismic ray method: recent developments. *Advances in Geophysics*, **48**, 1–126. http://www.sciencedirect.com/science/bookseries/00652687.

Chapman, C. H., 2002. Seismic ray theory and finite frequency extensions. In Lee, W. H. K., Kanamori, H., and Jennings, P. C. (eds.), *International Handbook of Earthquake and Engineering Seismology, Part A*. New York: Academic, pp. 103–123.

Chapman, C. H., 2004. *Fundamentals of Seismic Wave Propagation*. Cambridge: Cambridge University Press.

Chapman, C. H., and Coates, R. T., 1994. Generalized Born scattering in anisotropic media. *Wave Motion*, **19**, 309–341.

Chapman, C. H., and Drummond, R., 1982. Body-wave seismograms in inhomogeneous media using Maslov asymptotic theory. *Bulletin of the Seismological Society of America*, **72**, S277–S317.

Coates, R. T., and Chapman, C. H., 1990. Quasi–shear wave coupling in weakly anisotropic 3-D media. *Geophysical Journal International*, **103**, 301–320.

Farra, V., and Pšenčík, I., 2010. Coupled S waves in inhomogeneous weakly anisotropic media using first-order ray tracing. *Geophysical Journal International*, **180**, 405–417.

Fedorov, F. I., 1968. *Theory of Elastic Waves in Crystals*. New York: Plenum.

Gajewski, D., and Pšenčík, I., 1992. Vector wavefield for weakly attenuating anisotropic media by the ray method. *Geophysics*, **57**, 27–38.
Gjøystdal, H., Iversen, E., Laurain, R., Lecomte, I., Vinje, V., and Åstebol, K., 2002. Review of ray theory applications in modelling and imaging of seismic data. *Studia Geophsica et Geodaetica*, **46**, 113–164.
Hanyga, A., 1982. The kinematic inverse problem for weakly laterally inhomogeneous anisotropic media. *Tectonophysics*, **90**, 253–262.
Jech, J., and Pšenčík, I., 1989. First-order perturbation method for anisotropic media. *Geophysical Journal International*, **99**, 369–376.
Karal, F. C., and Keller, J. B., 1959. Elastic wave propagation in homogeneous and inhomogeneous media. *The Journal of the Acoustical Society of America*, **31**, 694–705.
Klimeš, L., 1984. The relation between Gaussian beams and Maslov asymptotic theory. *Studia Geophysica et Geodaetica*, **28**, 237–247.
Klimeš, L., 2002a. Lyapunov exponents for 2-D ray tracing without interfaces. *Pure and Applied Geophysics*, **159**, 1465–1485.
Klimeš, L., 2002b. Second-order and higher-order perturbations of travel time in isotropic and anisotropic media. *Studia Geophysica et Geodaetica*, **46**, 213–248.
Klimeš, L., 2006. Common-ray tracing and dynamic ray tracing for S waves in a smooth elastic anisotropic medium. *Studia Geophysica et Geodaetica*, **50**, 449–461.
Klimeš, L., 2010. Phase shift of the Green tensor due to caustics in anisotropic media. *Studia Geophysica et Geodaetica*, **54**, 269–289.
Kravtsov, Yu. A., 1968. "Quasiisotropic" approximation to geometrical optics. *Doklady Akademii Nauk SSSR*, **183**(1), 74–77 (in Russian).
Kravtsov, Yu. A., and Orlov, Yu. I., 1990. *Geometrical Optics of Inhomogeneous Media*. Heidelberg: Springer Verlag.
Kravtsov, Yu. A., and Orlov, Yu. I., 1999. *Caustics, Catastrophes and Wave Fields*. Heidelberg: Springer Verlag.
Popov, M. M., 1982. A new method of computation of wave fields using Gaussian beams. *Wave Motion*, **4**, 85–97.
Pšenčík, I., 1998. Green's functions for inhomogeneous weakly anisotropic media. *Geophysical Journal International*, **135**, 279–288.
Pšenčík, I., and Farra, V., 2007. First-order P-wave ray synthetic seismograms in inhomogeneous weakly anisotropic media. *Geophysical Journal International*, **170**, 1243–1252.
Stamnes, J. J., 1986. *Waves in Focal Regions*. Bristol: Adam Hilger.
Thomson, C. J., 1989. Corrections for grazing rays to 2-D seismic modelling. *Geophysical Journal International*, **96**, 415–446.
Thomson, C. J., 1999. The gap between seismic ray theory and full wavefield extrapolation. *Geophysical Journal International*, **137**, 364–380.
Thomson, C. J., and Chapman, C. H., 1985. An introduction to Maslov's asymptotic method. *Geophysical Journal of Royal Astronomical Society*, **83**, 143–168.
Virieux, J., 1996. Seismic ray tracing. In Boschi, E., Ekström, G., and Morelli, A. (eds.), *Seismic Modelling of Earth Structures*. Bologna: Editrice Compositori, pp. 223–304.

Cross-references

Energy Partitioning of Seismic Waves
Seismic Anisotropy
Seismic Diffraction
Seismic Waves, Scattering
Seismic, Migration
Seismic, Viscoelastic Attenuation
Seismic, Waveform Modeling and Tomography
Traveltime Tomography Using Controlled-Source Seismic Data

SEISMIC, RECEIVER FUNCTION TECHNIQUE

Rainer Kind[1], Xiaohui Yuan[2]
[1]Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences, Section 2.4, Seismology, Potsdam, Germany
[2]Section Seismology, GFZ German Research Centre for Geosciences, Potsdam, Germany

Synonyms

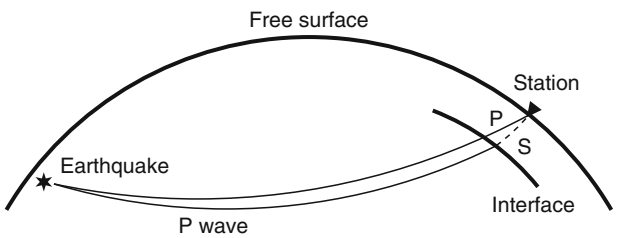
Receiver functions; Scattered teleseismic waves

Definition

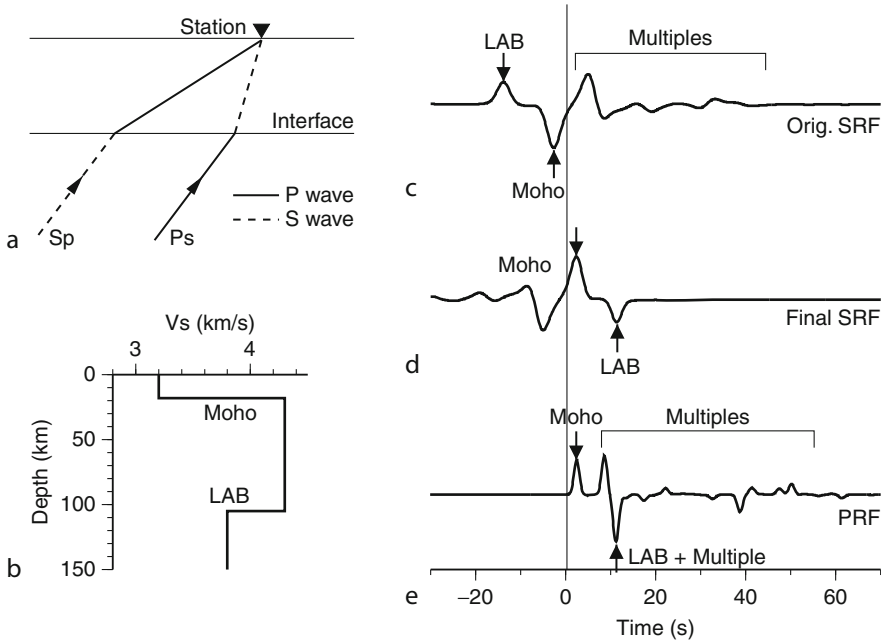
Receiver function. Response of the Earth's structure below seismic stations to incident teleseismic waves. Seismic discontinuities in the Earth's crust or upper mantle cause waves from distant earthquakes to be converted from P to S waves or vice versa, or to be multiply reflected between discontinuities before arriving the stations. Such scattered waves carry information about the seismic structure of the lithosphere and upper mantle. Scattered waves are weak signals and summation of many records is usually required to visualize the useful information.

Receiver function technique

If a seismic wave hits a discontinuity between two solid materials with different physical parameters, a part of the wave will be reflected and another part will be transmitted. There is also mode conversion between compressional (P) waves and shear (S) waves (see *Energy Partitioning of Seismic Waves*). Therefore, a P wave crossing a discontinuity will generate an S wave (called Ps), which will follow the P wave with a slower speed; and an S wave will produce a P wave (called Sp), which will run ahead of the S wave with a faster speed. If these wave types are recorded at the surface and their incidence angles and the material velocities are known, the depth of the discontinuity, where they are generated can be determined from the differential times of the two seismic phases, mother phase



Seismic, Receiver Function Technique, Figure 1 P receiver function geometry. The target zone below the three-component receiver is illuminated from below. The source is an earthquake at a large epicentral distance. When a plane P wave encounters interfaces below the receiver, Ps waves are generated. The incidence angle of the P wave is always bigger than that of the Ps waves. The travel path of S receiver functions (SRFs) is very similar, except that the incidence angle of the S wave is smaller than that of the Sp wave.

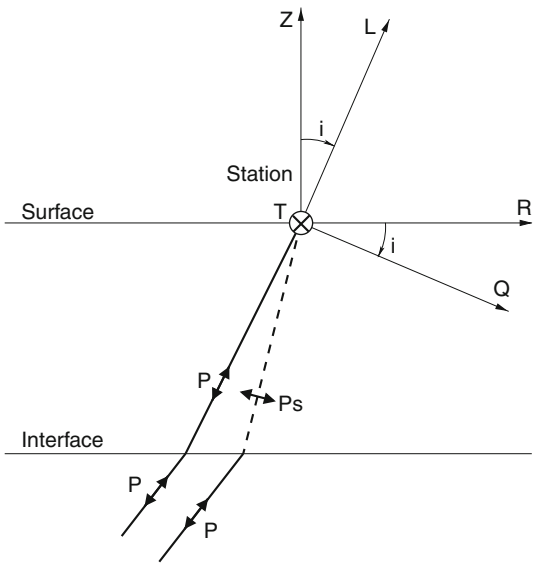


Seismic, Receiver Function Technique, Figure 2 (**a**) Sketch showing ray paths of teleseismic Ps and Sp converted waves below a station. Note that the piercing point of Sp is offset farther away from the station than that of Ps. (**b**) Velocity model for the calculation of P and S receiver functions. (**c**) Theoretical S receiver function (SRF), primary conversions arrive prior to S (which arrives at zero time and is omitted here), crustal multiples arrive after S. (**d**) S receiver function (trace **c**) after reversal of timescale and amplitude for better comparison with P receiver functions (PRFs). (**e**) P receiver functions (PRFs) of model in (**b**). The main difference of P and S receiver functions is that S receiver functions are free of crustal multiples at the time of the lithosphere–asthenosphere boundary (LAB) arrivals. Note that in the case of a homogeneous halfspace, no signal would be contained in the above traces (Yuan et al., 2007).

and daughter phase. The waveforms of both phases carry information about the structure of the discontinuity, which can be isolated by the receiver function technique. The ray path of such waves is shown in Figure 1. Figure 2 compares incident P and S phases for a simple model with two discontinuities (Moho and lithosphere–asthenosphere boundary, LAB, see *Earth's Structure, Continental Crust*; *Lithosphere, Continental*; *Lithosphere, Oceanic*), which cause converted and multiply reflected phases that lead finally to the receiver functions. There are a number of processing steps required in the receiver function technique, which will be considered in the following.

Separation of P and S waves

Since the converted daughter phase is a different wave type with different particle motion, it can be separated from the mother phase. The usual vertical, north–south and east–west (ZNE) components of a seismic record should be rotated, for better identification of signals, into a coordinate system where P, SV, and SH wave types are on different components. Due to heterogeneities close to the station, such a coordinate system (frequently called LQT or P-SV-SH system, see Figure 3) depends on the local angle of incidence and back azimuth. In many cases, only the horizontal components N and E are rotated into the radial (R) and transverse (T) components. Frequently

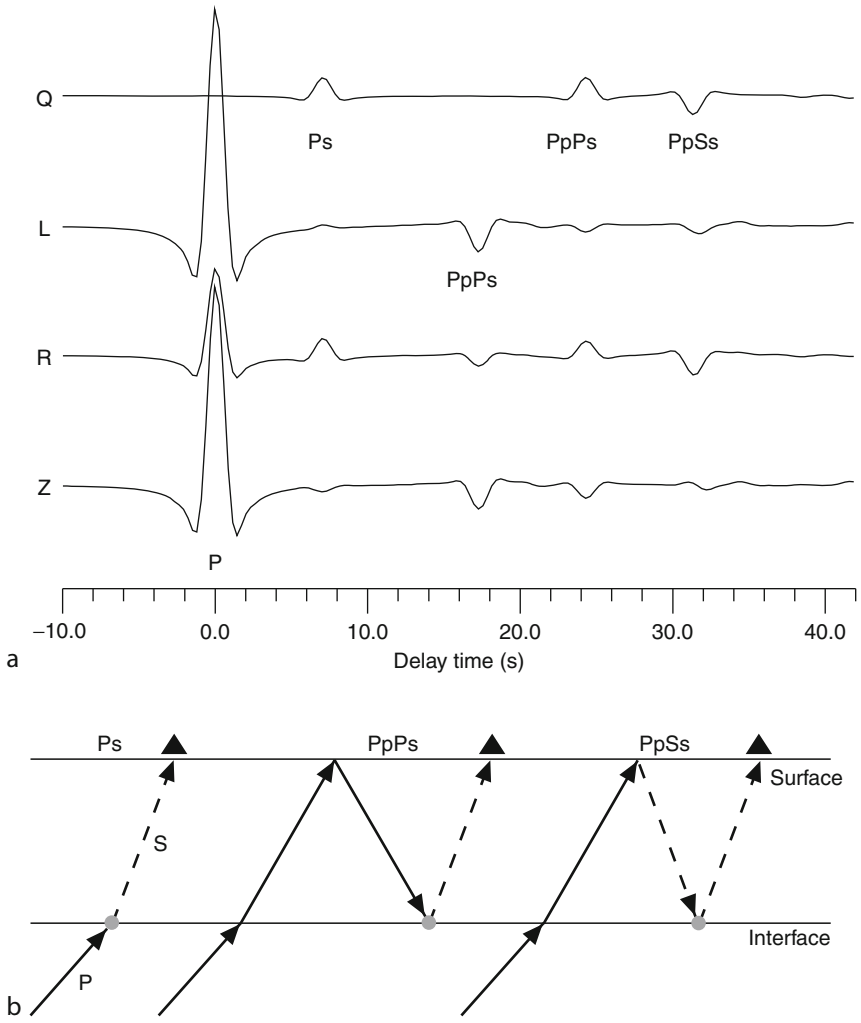


Seismic, Receiver Function Technique, Figure 3 Rotation of the radial (R) and vertical (Z) components around the angle of incidence into the L and Q system (Vinnik, 1977). The P wave is in this case only on the L component and the SV wave is only on the Q component. The SH wave remains on the T component.

the rotation is done according to the theoretical back azimuth between receiver and source and the theoretical angle of incidence according to epicentral distance. A local ray coordinate system may be determined from the diagonalization of the covariance matrix of the ZNE recordings within a certain time window (Montalbetti and Kanasewich, 1970; Kind and Vinnik, 1988). The length of the time window used plays an important role for determination of rotation angles with this technique. The influence of the free surface of the Earth is small and in most cases not considered. In Figure 4, theoretical seismograms in the ZR and LQ coordinate system are shown, along with ray diagrams for direct Ps phase and multiple phases PpPs and PpSs.

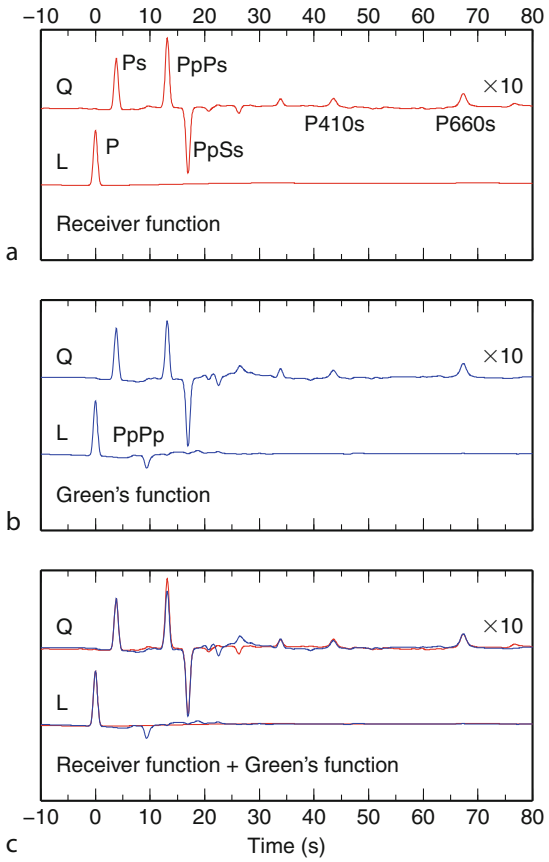
Deconvolution

After separation of P and SV wave types on separate components, it is much easier to identify and interpret the small converted signals. In earlier times when only analog data records were available, only a few records could be searched for converted phases. With the large amount of digital data now available (see *Seismological Networks*), weak signals can be enlarged relatively easily by summation of many records. A number of problems need to be considered before seismic records can be summed. Different waveforms and amplitudes generated by different earthquakes are another problem, which needs to be taken into account before summation of many traces can be applied. After rotation, records of one source and many distributed stations may be summed because all records are caused by a single event with the same source-time function. This is the well-known delay and sum technique. However the results of such a summation will be the response function at the source region (including, e.g., depth phases) and not the "receiver function." If we are interested in the structure beneath a seismic receiver,



Seismic, Receiver Function Technique, Figure 4 (**a**) Effects of rotation of theoretical Z and R coordinates into the LQ system. Phases arriving as S wave (Ps, PpPs, PpSs) at the station remain on the Q component only. Phases that arrive as P waves remain on the L component (P, PpPp). (**b**) Ray paths of crustal multiples.

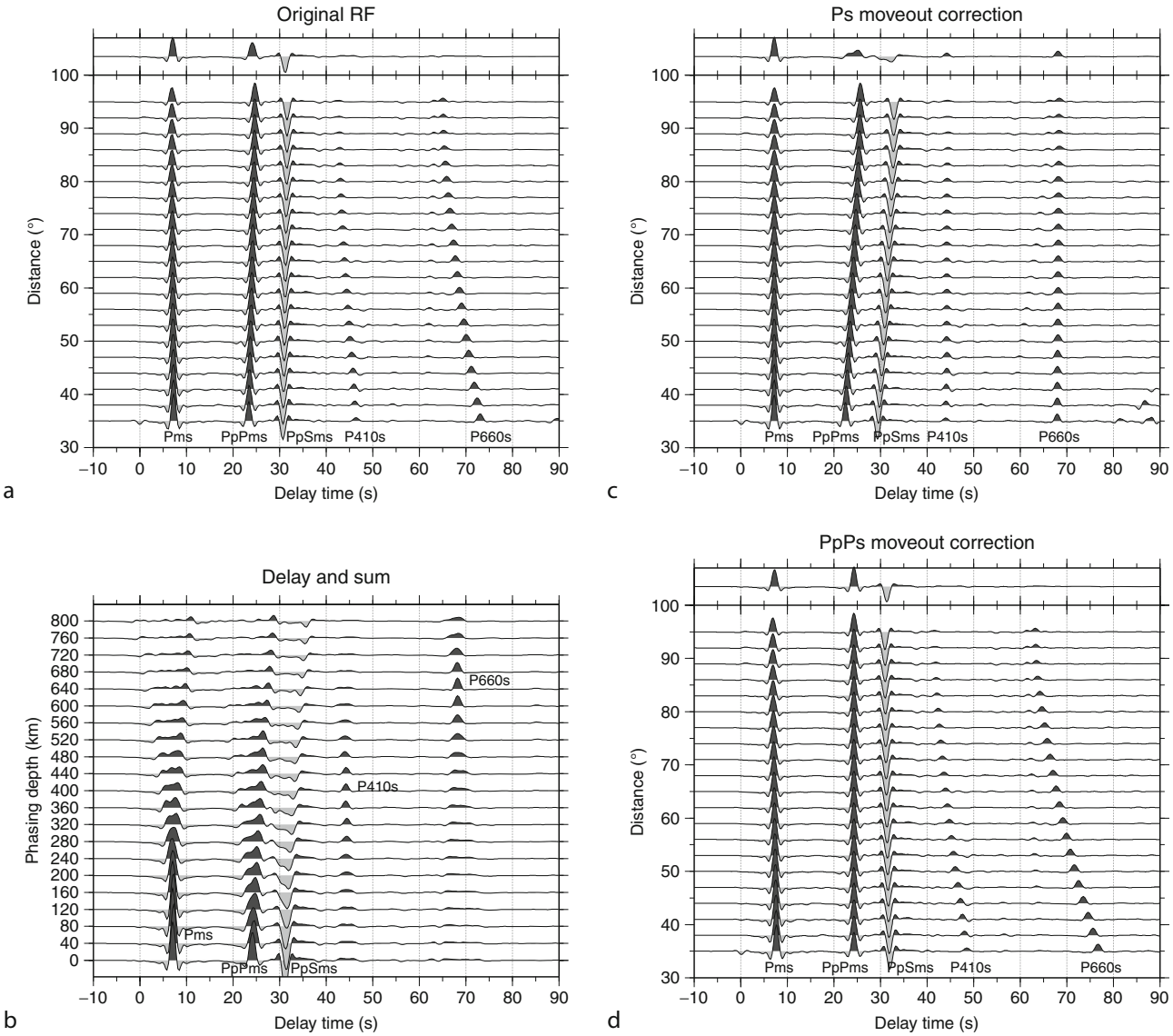
records from many earthquakes at this receiver with very different source-time functions and magnitudes must be summed. Only in this case we will obtain as the result of the summation the response of the structure beneath a receiver. Influences of the receiver structure are common to all records and will be enhanced by summation. In contrast, influences of the source or the source structure (depth phases, conversions near the source) will be reduced. Obtaining the structure beneath a receiver is especially interesting because in contrast to sources we can move the receivers where we need them and record teleseismic events to study the structure at nearly any location (with many problems at ocean bottom stations). The procedure used for source equalization is usually deconvolution. In the frequency domain, deconvolution of P receiver functions (PRFs) means division of the SV component by the P component and in S receiver functions (SRFs) it means division of the P component by the SV component (Langston, 1979; Gurrola et al., 1995). In time domain, a Wiener filtering approach (Kind et al., 1995) can be applied. For P receiver functions, an inverse filter can be generated by minimizing the least-square difference between the observed P waveform on the vertical component and the desired delta-like spike function. The inverse filter is then convolved with other components to obtain receiver functions. Ligorria and Ammon (1999) proposed an alternative approach to iteratively remove the source and propagation effects. The question which type of deconvolution is carried out does not seem to be very significant. After deconvolution, amplitude normalization is applied, resulting in amplitude values of the converted phase in percent of the incident phase. Deconvolution, however, has one principal disadvantage. One entire component is considered as the source-time function. For example, the entire P component over duration of perhaps 100 s is deconvolved from the SV component. This means all signals within this window on the P component are considered source. For example, P multiples within the crust at the receiver site will also be eliminated by this kind of deconvolution. In this sense, deconvolution prevents computation of the complete Green's function (impulse response) at the receiver site (Kumar et al., 2010; Baig et al., 2005; Bostock, 2004; Langston and Hammer, 2001). The modification of seismic records caused by deconvolution is demonstrated in Figure 5. Kumar et al. (2010) have shown that plain summation of many records of a seismic station results in SV component traces, which are nearly identical with deconvolved traces and P component traces are preserved. They only applied amplitude and sign equalization and summed all traces aligned along the maximum of the P signal (see Kumar et al., 2010). The only disadvantage of the new technique is that more traces are needed to obtain the same signal-to-noise ratio.



Seismic, Receiver Function Technique, Figure 5
(**a**) Deconvolved, (**b**) original theoretical seismograms (Kind, 1985) and (**c**) the overlay of both for a simple crust-mantle model. All phases except P are eliminated on the deconvolved P component. The deconvolved SV (Q) component differs significantly from the theoretical SV component. The P_MP phase has disappeared on the deconvolved SV component, as expected. However, the PpPs Moho multiple reflection has a significantly distorted amplitude (Kumar et al., 2010).

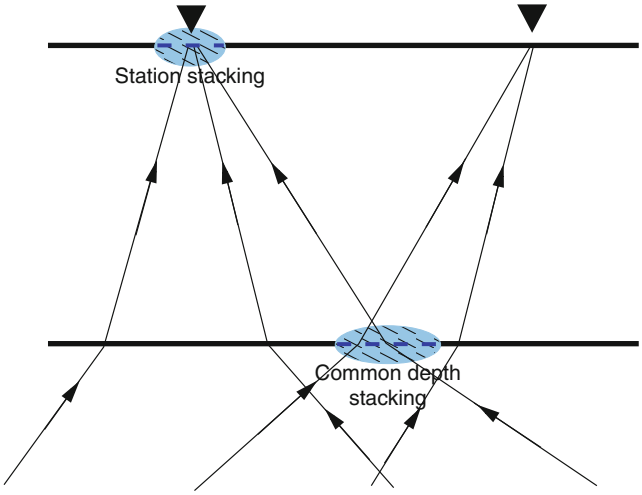
Moveout correction and summation

An important problem is that seismic phases travel with different velocities. This means that the differential times of seismic phases in one record depend on epicentral distances. Therefore, only records from similar epicentral distances can be summed. Traces from different distances may also be summed if the conversions arrive very closely in time to their mother phase (perhaps still possible for crustal conversions). For converted phases arriving later from greater depths, it is necessary to consider effects of their different slowness. Vinnik (1977) has solved this problem by using a delay and sum technique known from array processing and applying it to the upper mantle discontinuities. Results are displayed in a slowness-time plot (or conversion-depth-time plot, see Figure 6). Another solution is that the distance moveout correction procedure could be applied before summation, which is very well known in applied seismics (see *Deep Seismic Reflection and Refraction Profiling*). In moveout correction, the timescale is stretched or compressed in order to parallelize the same type of traveltime curves. A fixed reference



Seismic, Receiver Function Technique, Figure 6 Theoretical P receiver functions in original form and moveout corrected. A velocity model is needed to apply the moveout correction. (*Bottom left*) Delay and sum technique developed by Vinnik (1977). Delays are computed for a number of assumed depths of conversion.

slowness of 6.4 s/degree (or 67° epicentral distance) is frequently used at which the timescale remains unchanged. At smaller distances the timescale is compressed, and expanded at larger distances, in order to transform the traveltime curve of the converted phase under investigation into one parallel to the mother signal. After moveout correction, a number of traces may be lined up and summed into a single trace. A group of traces used for summation could be all records from either one station or records with piercing points in a certain geographical region (box) at a certain depth (common conversion point technique, Dueker and Sheehan, 1997, see Figure 7). The advantage of the moveout correction and summation technique is that single traces representing groups of traces can be easily compared. After moveout correction for direct conversions, the summation traces contain only such phases. Surface multiples are destructively superimposed because of their different slowness. Moveout corrections cannot only be applied to Ps or Sp conversions, but also to other phases like multiples. In this case, the summation trace contains only a certain type of multiples (see Figure 6). Moveout correction and summation also solves a special problem in S receiver functions. Not every precursor of S on the P component is an S-to-P



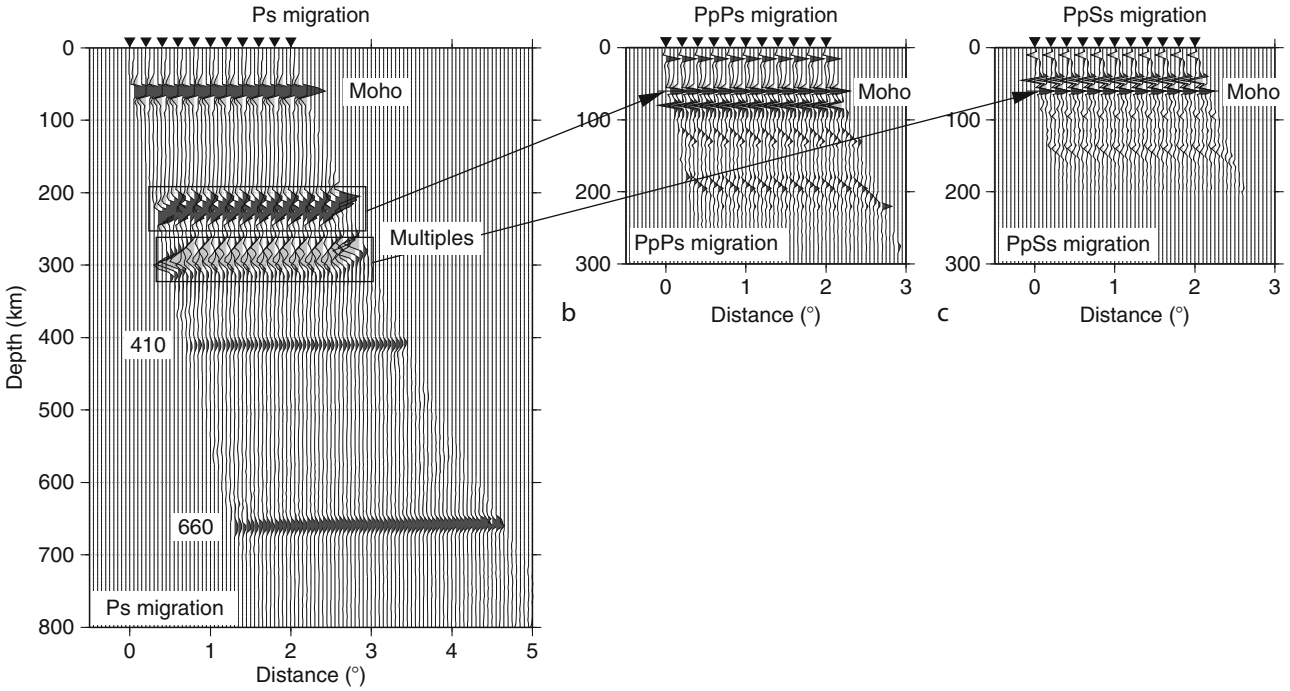
Seismic, Receiver Function Technique, Figure 7 Cartoon showing receiver function rays in a simple Earth model. Single station summation is useful for shallower studies or for averages over larger regions in greater depths. In case of dense deployments of several stations, common depth summation, (CCP) may be used.

conversion. Surface multiples of the P phase can also arrive on the vertical component as precursors of S (Bock, 1994). These phases have, fortunately, a different slowness and are suppressed in the delay and sum technique or the moveout correction and sum technique.

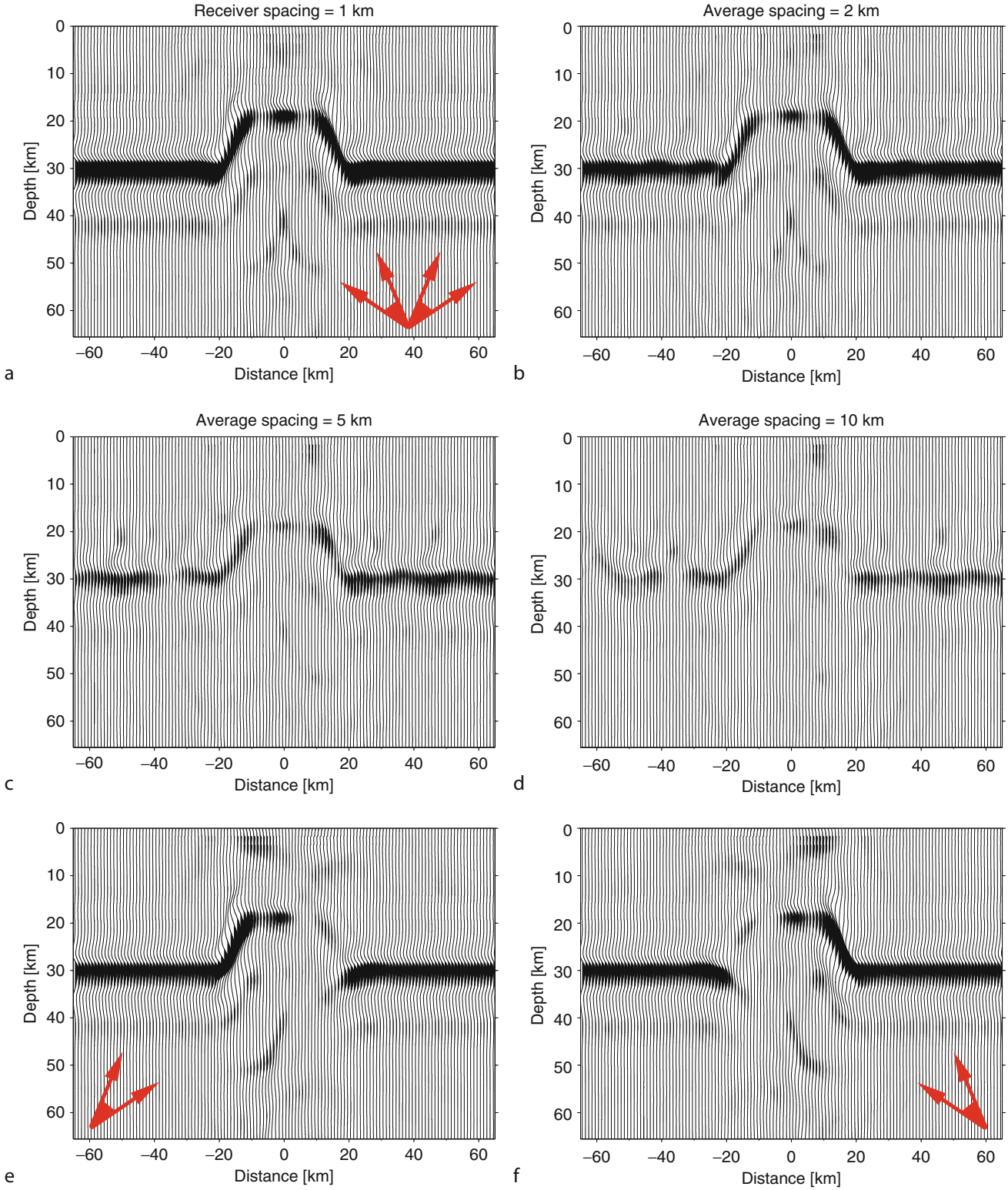
CCP stack and migration

Stacking by single stations is meaningful if piercing points from different stations are not overlapping in the depth range studied. If several stations are closely spaced and piercing points are overlapping at the study depth, summation by common regions of piercing points at that depth may be more meaningful (common conversion point stacking, CCP). In this case, piercing point locations for conversions at a defined depth are computed and all traces with piercing points inside a defined area are summed (see Figure 7).

Stacked time domain receiver functions are not yet the goal of the processing, even if clear converted phases can be recognized. We need to know where in the three-dimensional Earth the conversion occurs. Therefore, the final goal is the migration from the time series into



Seismic, Receiver Function Technique, Figure 8 Principle of depth migration of seismic records (theoretical seismograms in this case). Each ray from each station is back projected into an assumed model of the Earth. The amplitudes of the seismogram are distributed along this ray path. In the left figure it is assumed that all energy in the seismogram originates from direct conversions; in the other figures it is assumed that all energy is caused by multiples. Migration can, like moveout correction, only be done for one wave type at a time. The energy of the other phase type is migrated to the wrong depth and not focused.



Seismic, Receiver Function Technique, Figure 9 (**a**–**d**) Migrated theoretical 1-Hz receiver functions of a dome structure with different receiver spacing. A spacing of 5 or less kilometers is required to obtain a good image of the structure. (**e**–**f**) Effects of illuminating the structure from limited back azimuths (see *arrows*, Ryberg and Weber, 2000). Station spacing is 1 km.

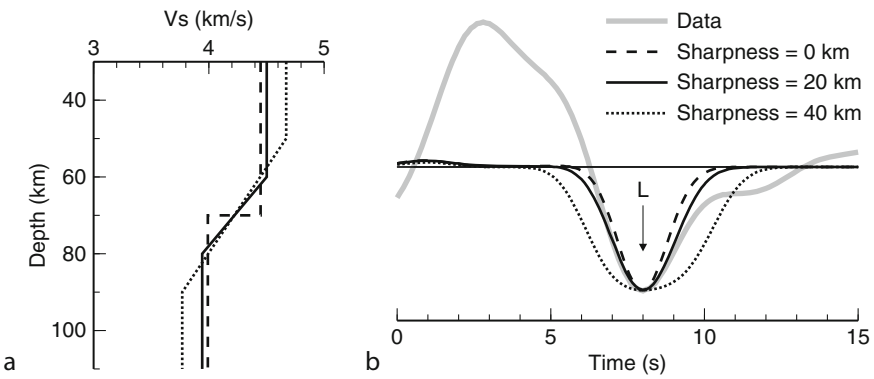
a depth model (see *Seismic, Migration*). Depth migration with a known velocity depth model is necessary for this purpose. Migration is a very well-known technique in controlled source seismics. In earthquake seismology with relatively sparse station distribution more simple versions of migration with one-dimensional models are often used. The amplitudes of the receiver function traces are distributed along the ray path of a known velocity depth model. The Earth model is divided in boxes and amplitudes of all rays in one box are summed. The result is an approximate two-dimensional (or three-dimensional) distribution of seismic energy in space where hopefully seismic discontinuities can be identified (see Figure 8). In the case of P receiver functions, such images are blurred with multiples from the surface, which produce apparent discontinuities. One of the first examples of receiver function depth migration is given by Kosarev et al. (1999). A more sophisticated development is given, e.g., by Bostock (2002).

The results of the processing steps described so far indicate the great similarity with steep angle–controlled source techniques. The topography of seismic discontinuities in the interior of the Earth is displayed either in the time or space domain. Distance moveout corrections and depth migration are the most important common steps. The main differences are the frequency content (which determines the resolution) and depth penetration. Steep angle seismics uses frequencies of many Hertz, whereas the teleseismic signals have periods from one to many seconds. Steep angle seismics does not always reach the Moho, whereas receiver functions have practically unlimited depth penetration since the signals are incident from below. Ryberg and Weber (2000) conclude from computations of theoretical seismograms of two-dimensional models that a station spacing of a few kilometers is required if structures with a few kilometers scale length need to be resolved (see Figure 9). A station density of a few kilometers has so far been achieved only in very few deployments (e.g., Hansen and Dueker, 2009).

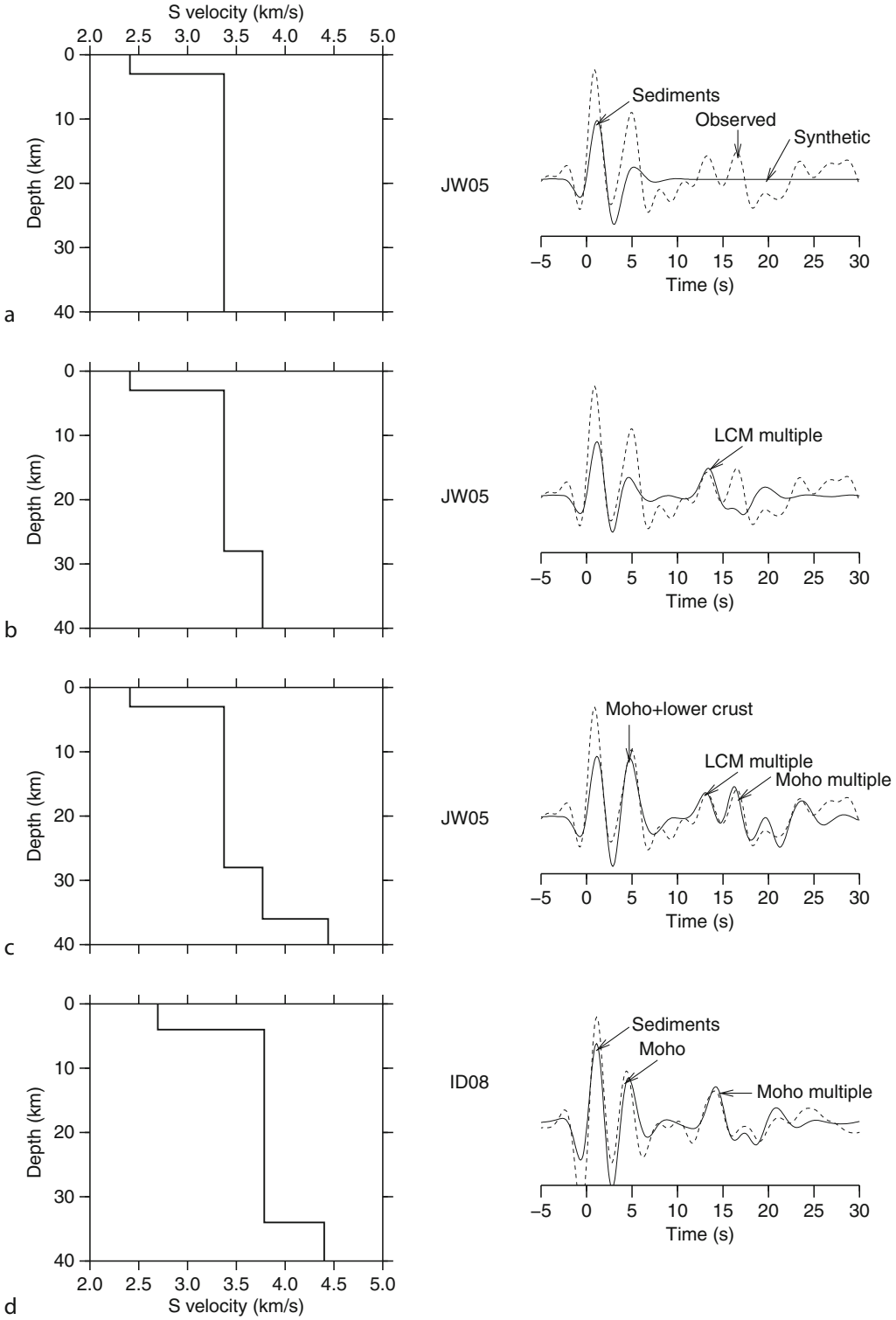
Waveform modeling

The above described processing steps for receiver functions lead to the detection and location of discontinuities in the interior of the Earth and their topography, ranging from the base of the sedimentary layers to the upper mantle transition zone between 410- and 660-km depth. A frequently applied additional step in receiver functions is the inversion of the waveforms over the entire length of the receiver function at a station into a one-dimensional velocity depth model (e.g., Ammon et al., 1990; Kind et al., 1995). Such an inversion technique has the capability of fitting the data nearly perfectly. However, problems involved are the nonuniqueness of the inversion and the assumption that seismic phases, which have been generated by a three-dimensional Earth, can be transformed into a one-dimensional model (Ammon et al., 1990). Although the nonuniqueness problem has been improved recently by using long-period signals (Jacobsen and Svenningsen, 2008), even a perfect fit of the observed waveforms with theoretical waveforms encounters the problem of fitting three-dimensional effects and possible noise into a one-dimensional model.

Somewhat closer to reality seems to be the inversion of isolated waveforms of conversions from individual discontinuities in P or S receiver functions (e.g., Moho or LAB). An example is given by Li et al. (2007), see Figure10. They summed about 8,000 S-to-P converted signals from the LAB (S receiver functions) in the western US and inverted the width of the summed signal in terms of thickness of the LAB gradient. They obtained a transition zone of not more than 20 km. However, determinations of the size of the velocity jump and of the thickness of the LAB from summation traces may depend on



Seismic, Receiver Function Technique, Figure 10 Modeling of the LAB summation phase in the western US (Li et al., 2007, Figure 8). (**a**) Models with different sharpness and velocity contrast of the LAB. Dashed line: LAB as a first-order discontinuity; solid line: LAB as a 20-km thick gradient transition zone and dotted line: LAB as a 40-km thick gradient transition zone. (**b**) Waveforms of data and synthetic S receiver functions. The thick gray line is the summation of all the S receiver function stacks for each station after an alignment of the LAB phase. A 3–20 s bandpass filter is applied prior to alignment and summation for a better phase correlation. Synthetic S receiver functions correspond to the models shown in (**a**). The large positive swing in the data at less than 5 s is the averaged Moho signal, which is not computed in the synthetic seismograms.



Seismic, Receiver Function Technique, Figure 11 Forward waveform modeling of P receiver functions with relatively simple models (Mohsen et al., 2005).

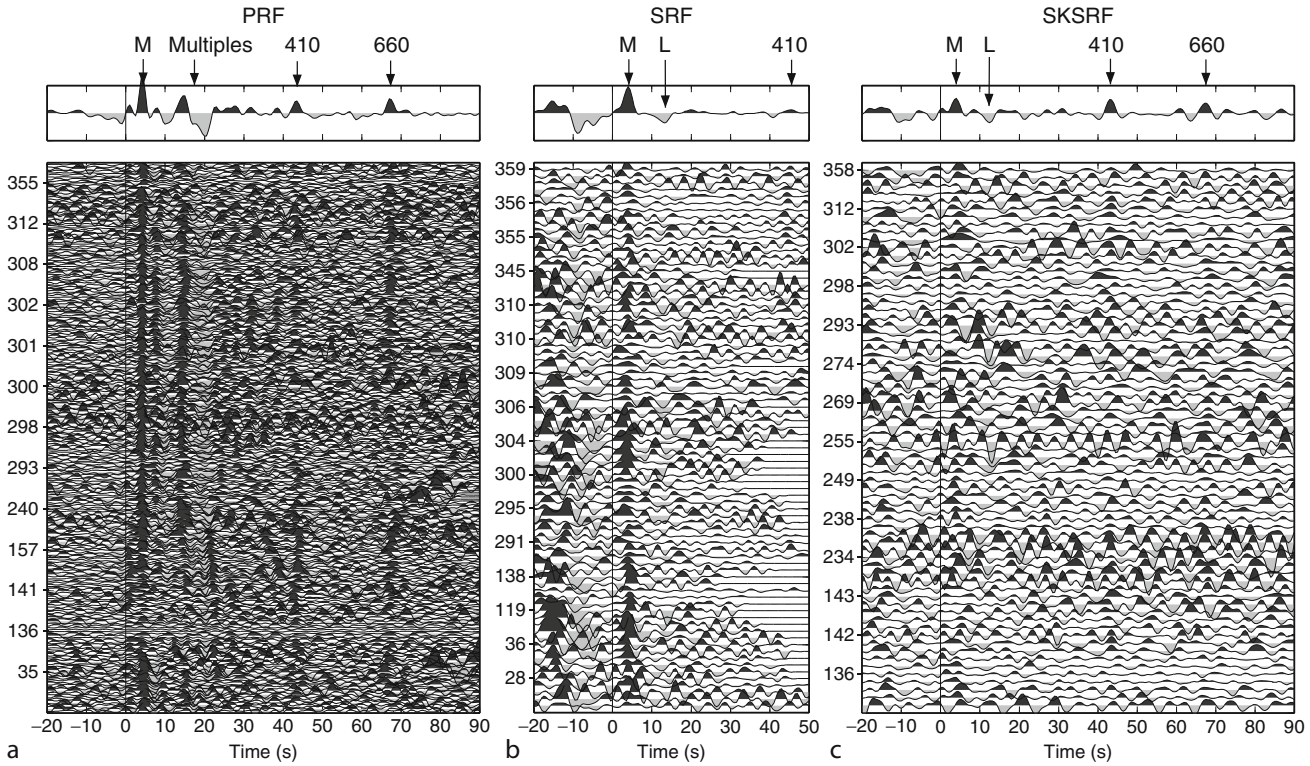
focusing and defocusing effects of the topography within the region of piercing points and erroneous transform topography into the vertical velocity structure. Therefore, this 20-km gradient must be considered as an upper limit in this case. Another example is given by Mohsen et al. (2005) in Figure 11. They show how each one of the observed seismic phases may be modeled with good accuracy with a relatively simple model.

Two data examples

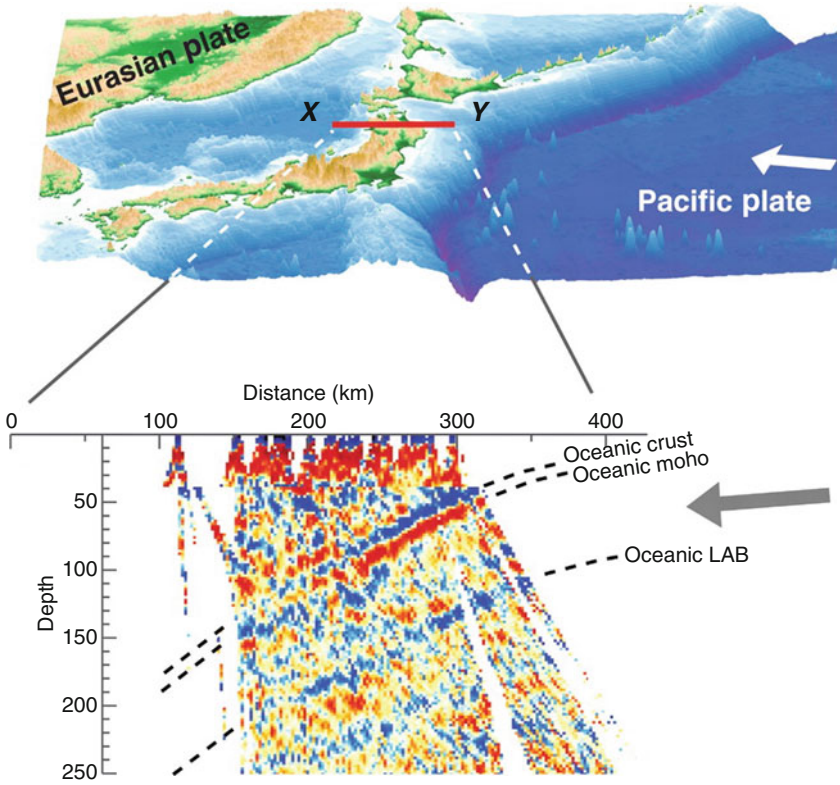
In Figure 12 are shown P, S, and SKS receiver functions in time domain from a seismically quiet station in northern Canada. The interesting point in these data is that in S and SKS receiver function (SKSRF), a clear negative phase is observed at 12–13 s, which corresponds to a depth of about 110 km. A negative signal indicates a low velocity zone. Such a zone in that depth may be caused by the LAB. However, from surface wave studies a much larger depth of the LAB is expected in this region. These contradicting results pose very interesting questions about the structure of the upper mantle beneath old cratons. In Figure 13 is shown an excellent example of depth migrated P receiver functions beneath a subduction zone (Kawakatsu et al., 2009). The subducting oceanic lithosphere is completely visible, including upper and lower boundary of the oceanic crust and the lower boundary of the oceanic lithosphere (see *Lithosphere, Oceanic*).

Summary

Receiver functions or converted waves belong to a class of small secondary scattered waves generated by a relatively strong main phase like P or S. Similar scattered phases are precursors of other main phases like PP, SS, or P′P′, which have the great advantage to cover areas that are difficult to access, like oceans (see *Body Waves*; *Seismic Phase Names: IASPEI Standard*). These techniques are since about 30 years the main tools to study discontinuities of material properties in the Earth. The other main seismic tool is tomography, which is sensitive to smooth changes of material properties. The success of the scattered wave techniques was made possible by the great extension of high quality seismic networks in the last decades. Permanent and mobile networks supplement each other. Especially studies of the lithospheric plates and the mantle transition zone have gained much from the new techniques. Many more detailed results can be expected when much denser networks will be used in future.



Seismic, Receiver Function Technique, Figure 12 PRFs, SRFs, and SKSRFs of station YKW3 (Yellow Knife, Canada). The receiver functions are corrected for moveout with a reference slowness of 6.4 s/degree and a reference model of the IASP91 (Kennett and Engdahl, 1991). The number of the PRFs is more than 200, the numbers of the SRFs and SKSRFs are ~80 each. A three-pole low-pass filter with corner frequency of 3 S is applied to all the data. Back azimuth values for some sample traces are labeled to the left of each trace. The summation trace is plotted above each receiver function section. Main conversion phases are labeled on the summation traces.



Seismic, Receiver Function Technique, Figure 13 P receiver functions for ocean bottom station WP2 and receiver function image of subducting Pacific plate beneath northeast Japan. In the topographic map, the inverted red triangle indicates the location of WP2, which is deployed at a depth of 460 m below the seafloor with a water column of 5,566-m thick. (**a**) P receiver function image along the profile XY using dense land seismic data of Hi-net from Japan. *Red* and *blue* colors indicate velocity increase (from shallow to deep) and decrease at the point, respectively. The top surface of the slab and the oceanic Moho are clearly imaged, as well as the bottom surface of the slab (i.e., subducting LAB). (**b**) P receiver function and S receiver function for WP2. Negative phases associated with a shear wave velocity drop marked as LAB appear to correspond to the strong negative (*blue*) signature of the subducting slab below Japan. (After Kawakatsu et al., 2009.)

A recent review of the receiver function technique has been given by Rondenay (2009). A number of additional processing steps are the use of multiples to determine average crustal properties (e.g., Zhu and Kanamori, 2000), the role of anisotropy (e.g., Levin and Park, 1997; Schulte-Pelkum et al., 2005), or three-dimensional inversion (e.g., Bostock, 2002).

Bibliography

Ammon, C. J., Randall, G. E., and Zandt, G., 1990. On the nonuniqueness of receiver function Inversions. *Journal of Geophysical Research*, **95**, 15303–15318.

Baig, A. M., Bostock, M. G., and Mercier, J.-P., 2005. Spectral reconstruction of teleseismic Green's functions. *Journal of Geophysical Research*, **110**, B08306.

Bock, G., 1994. Multiples as precursors of S, SKS and ScS. *Geophysical Journal International*, **119**(2), 421–427.

Bostock, M. G., 2002. Kirchhoff-approximate inversion of teleseismic wave fields. *Geophysical Journal International*, **149**, 787–795.

Bostock, M. G., 2004. Green's functions, source signatures, and the normalization of teleseismic wave fields. *Journal of Geophysical Research*, 109.

Dueker, K. G., and Sheehan, A. F., 1997. Mantle discontinuity structure from midpoint stacks of converted P to S waves across the Yellowstone hotspot track. *Journal of Geophysical Research*, **102**, 8313–8327.

Gurrola, H., Baker, G. E., and Minster, J. B., 1995. Simultaneous time-domain deconvolution with application to the computation of receiver functions. *Geophysical Journal International*, **120**, 537–543.

Hansen, S., and Dueker, K., 2009. P- and S-Wave Receiver Function Images of Crustal Imbrication beneath the Cheyenne Belt in Southeast Wyoming. *Bulletin. Seismological Society of America*, **99**, 1953–1961, doi:10.1785/0120080168.

Jacobsen, B. H., and Svenningsen, L., 2008. Enhanced uniqueness and linearity of receiver function inversion. *Bulletin. Seismological Society of America*, **98**, 1756–1767.

Kawakatsu, H., Kumar, P., Takei, Y., Shinohara, M., Kanazawa, T., Araki, E., and Suyehiro, K., 2009. Seismic evidence for sharp lithosphere-asthenosphere boundaries of oceanic plates. *Science*, **324**, 499–502.

Kennett, B. L. N., and Engdahl, E. R., 1991. Travel times for global earthquake location and phase identification. *Geophysical Journal International*, **105**, 429–465.

Kind, R., 1985. The reflectivity method for different source and receiver structures and comparison with GRF data. *Journal of Geophysics*, **58**, 146–152.

Kind, R., and Vinnik, L. P., 1988. The upper-mantle discontinuities underneath the GRF array from P-to-S converted phases. *Journal of Geophysics*, **62**, 138–147.
Kind, R., Kosarev, G. L., and Petersen, N. V., 1995. Receiver functions at the stations of the German Regional Seismic Network (GRSN). *Geophysical Journal International*, **121**, 191–202.
Kosarev, G., Kind, R., Sobolev, S. V., Yuan, X., Hanka, W., and Oreshin, S., 1999. Seismic evidence for a detached Indian Lithospheric mantle beneath Tibet. *Science*, **283**, 1306–1309.
Kumar, P., Kind, R., and Yuan, X. H., 2010. Receiver function summation without deconvolution. *Geophysical Journal International*, **180**, 1223–1230.
Langston, C. A., 1979. Structure under Mount Rainer, Washington, inferred from teleseismic body waves. *Journal of Geophysical Research*, **84**, 4749–4762.
Langston, C. H. A., and Hammer, J. K., 2001. The vertical component P-wave receiver function. *Bulletin. Seismological Society of America*, **91**(6), 1805–1819.
Levin, V., and Park, J., 1997. P-SH conversions in a flat-layered medium with anisotropy of arbitrary orientation. *Geophysical Journal International*, **131**, 253–266.
Li, X. Q., Yuan, X. H., and Kind, R., 2007. The lithosphere-asthenosphere boundary beneath the western United States. *Geophysical Journal International*, **170**(2), 700–710.
Ligorria, J. P., and Ammon, C. J., 1999. Iterative deconvolution and receiver function estimation. *Bulletin. Seismological Society of America*, **89**, 1395–1400.
Mohsen, A., Hofstetter, R., Bock, G., Kind, R., Weber, M., Wylegalla, K., Rumpker, G., and DESERT Group, 2005. A receiver function study across the Dead Sea Transform. *Geophysical Journal International*, **160**, 948–960.
Montalbetti, J. F., and Kanasewich, E. R., 1970. Enhancement of teleseismic body phases with a polarization filter. *Geophysical Journal International*, **21**(2), 119–129.
Rondenay, S., 2009. Upper mantle imaging with array recordings of converted and scattered teleseismic waves. *Surveys in Geophysics*, **30**, 377–405.
Ryberg, T., and Weber, M., 2000. Receiver function array: a reflection seismic approach. *Geophysical Journal International*, **41**, 1–11.
Schulte-Pelkum, V., Monsalve, G., Sheehan, A., Pandey, M. R., Sapkota, S., Bilham, R., and Wu, F., 2005. Imaging the Indian subcontinent beneath the Himalaya. *Nature*, **435**, 1222–1225, doi:10.1038/nature03678.
Vinnik, L. P., 1977. Detection of waves converted from P to SV in the mantle. *Physics of the Earth and Planetary Interiors*, **15**, 39–45.
Yuan, X., Kind, R., Li, X., and Wang, R., 2006. S receiver functions: synthetics and data example. *Geophysical Journal International*, **175**(2), 555–564.
Yuan, X., Li, X., Wölbern, I., and Kind, R., 2007. Tracing the Hawaiian mantle plume by converted seismic waves. In Ritter, J. R. R., and Christensen, U. R. (eds.), *Mantle Plumes – A Multidisciplinary Approach*. Berlin: Springer, pp. 49–69.
Zhu, L. P., and Kanamori, H., 2000. Moho depth variation in southern California from teleseismic receiver functions. *Journal of Geophysical Research*, **150**, 2969–2980.

Cross-references

Body Waves
Deep Seismic Reflection and Refraction Profiling
Earth's Structure, Continental Crust
Energy Partitioning of Seismic Waves
Lithosphere, Continental
Lithosphere, Oceanic
Seismic Phase Names: IASPEI Standard
Seismic, Migration
Seismological Networks

SEISMIC, REFLECTIVITY METHOD

Mrinal K. Sen
Department of Geological Sciences, UT Institute for Geophysics, Jackson School of Geosciences,
The University of Texas at Austin, Austin, TX, USA

Synonyms

Reflection matrix method

Definition

Reflectivity method is a semi-analytic method for computing synthetic seismograms in layered elastic media. The method has now been extended to anisotropic and laterally varying media.

The reflectivity method

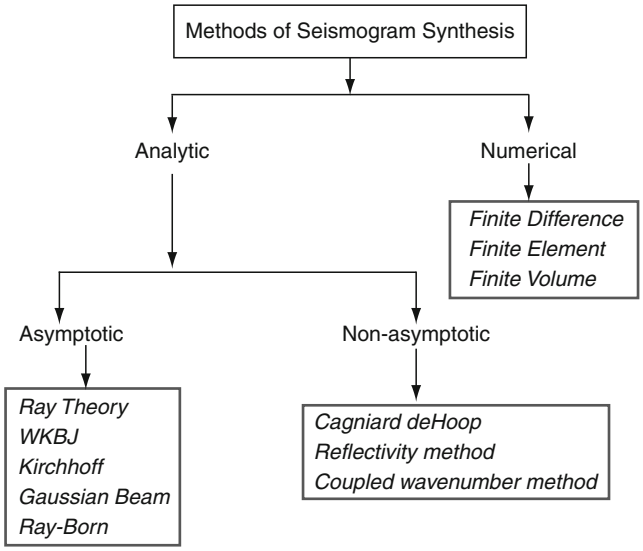
Introduction

Synthetic seismograms are now used routinely for interpretation of field seismic data and evaluating the performance of seismic processing algorithms. Generation of synthetic seismograms essentially involves solution of the partial differential equation (PDE) for seismic wave propagation. Different forms of the PDE are available for different earth models (e.g., acoustic, isotropic, elastic, anisotropic, homogeneous, and laterally inhomogeneous). Brute force numerical methods such as finite differences and finite elements can, in principle, handle models of any complexity. However, they are limited by the speed of computation and computer memory. Even with our latest computer technology, generation of realistic 3D synthetic seismograms is not possible within reasonable clock time.

Much effort has been spent over the years in developing methods that are fast and accurate. We are left with three choices:

- Approximate the earth model and generate analytic or semi-analytic solutions.
- Derive approximate (asymptotic) solution for realistic heterogeneous models.
- Derive pure numerical solution for general heterogeneous models.

Figure 1 provides a fairly complete list of the methods currently available for seismic modeling. Pure numerical methods based on finite-difference (Virieux, 1984) and finite-element (Marfurt, 1984; Komatitsch and Tromp, 1999; DeBasabe and Sen, 2007) approaches generate complete solutions but these methods become prohibitively expensive at high frequencies. The ray-based methods generate asymptotic solutions at infinite frequency and can be applied to models of general complexity as long as rays can be traced through the medium (Cerveny, 2001). Although the original WKBJ method is valid for layered media, its extension to laterally varying media is the Maslov method (Chapman and Drummond, 1982).



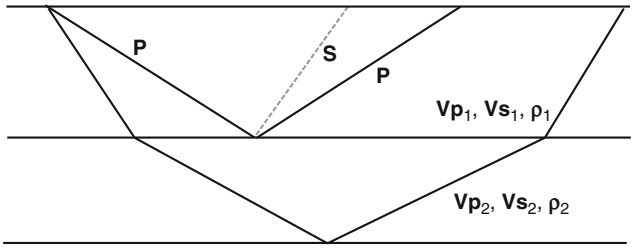
Seismic, Reflectivity Method, Figure 1 Different methods for computing synthetic seismograms: the reflectivity method is a semi-analytic method that is generally valid for layered earth models.

The methods such as the reflectivity (Fuchs and Muller, 1971) or Cagniard-deHoop method (e.g., Aki and Richards, 2000) are included in the category of non-asymptotic analytic method. They were originally proposed for layered (1D) isotropic media and are still commonly used in such media although some extensions of these methods to anisotropy and lateral heterogeneity have indeed been reported. Although I included these methods in the category of "analytic method," their final evaluation is done numerically. Thus strictly speaking, they should be included in the category of "semi-analytic" methods. The speed of computation of these methods can be improved by using asymptotic approximation at the final stage of evaluation of an oscillatory integral.

The primary goal of this entry is to provide an overview of the reflectivity method – perhaps the most popular method of seismogram synthesis. In the following sections, I will summarize the theory and numerical methods pertaining to reflectivity and provide some numerical examples. Finally, I will discuss an extension of the reflectivity method to laterally inhomogeneous media.

Plane waves

The plane waves are fundamental to understanding wave propagation. It has been demonstrated that a point source response can be generated by a weighted sum of individual plane wave responses (e.g., Aki and Richards, 2000). In a homogeneous medium, plane waves propagate without spreading. In a stratified medium, plane waves remain planar during their propagation. However, as they propagate through a stack of flat layers (Figure 2), they undergo changes in amplitude and phase.



Seismic, Reflectivity Method, Figure 2 A stack of layers in which each layer is characterized by its elastic parameters and layer thickness: as an incident plane wave propagates through the stack of layer, it undergoes mode conversion and changes in amplitude due to reflection and transmission.

- The amplitude changes can be accounted for by imposing boundary conditions of continuity of displacement and stress resulting in expressions for the so-called plane wave reflection and transmission coefficients.
- Travel time changes can be computed using individual layer velocities for locally linear ray-paths.
- Changes in propagation angle can be computed by imposing Snell's law of preservation of horizontal slowness.
- In a stratified medium, an incident plane wave generates converted waves and internal multiples with appropriate amplitudes given by reflection and transmission coefficients.

Figure 2 displays a stack of layers in which each layer is characterized by its elastic parameters (densities and, P- and S-wave velocities) and thicknesses in an elastic isotropic medium. An incident plane P-wave characterized by a ray (normal to the plane) given by the angle it makes with the vertical axis is shown in the top layer. At the first layer boundary, it generates a reflected P-wave and a reflected SV-wave (converted). It also generates several reverberations with the top free surface (not shown in the figure). As the plane wave enters the second layer, it changes its propagation direction and similarly generates several converted waves and multiples. The primary task in generating synthetic seismograms is to keep track of all of these paths and account for all these amplitude and phase changes. The reflectivity method does exactly that. However, before we look into the details of the reflectivity method, let us first examine some simple cases with a few rays.

Simple plane wave synthetics

It is fairly straightforward to generate plane wave synthetic seismograms for a layered medium for a few rays or plane waves. Two fundamental parameters involving plane wave seismograms are the ray parameter or horizontal slowness denoted by p and vertical delay time denoted by τ. They are defined as follows

$$p = \frac{\sin\theta}{v}, \quad q = \frac{\cos\theta}{v}, \tag{1}$$

where, θ is the vertical angle of the ray, q is called the vertical slowness and

$$\tau = 2hq, \tag{2}$$

where, h is the layer thickness.

Now referring to Figure 2, the *PP* reflection delay time for layer 1 is

$$\tau_1^{PP} = 2h_1 q_1^P, \tag{3}$$

and the *PS* reflection delay time is

$$\tau_1^{PS} = h_1\left(q_1^P + q_1^s\right). \tag{4}$$

In the above equations, the subscript represents the layer and the superscript represents the wavemode under consideration. The plane wave response of the primary *PP* mode from the interface between layer 1 and layer 2 can be written down in the frequency-slowness domain as

$$R(\omega, p) = R_1^{PP} \exp\left(i\omega 2h_1 q_1^{PP}\right), \tag{5}$$

where, R_1^{PP} is the *PP* plane wave reflection coefficient for interface 1. Similar expression can be written down for the *PS* mode primary reflection.

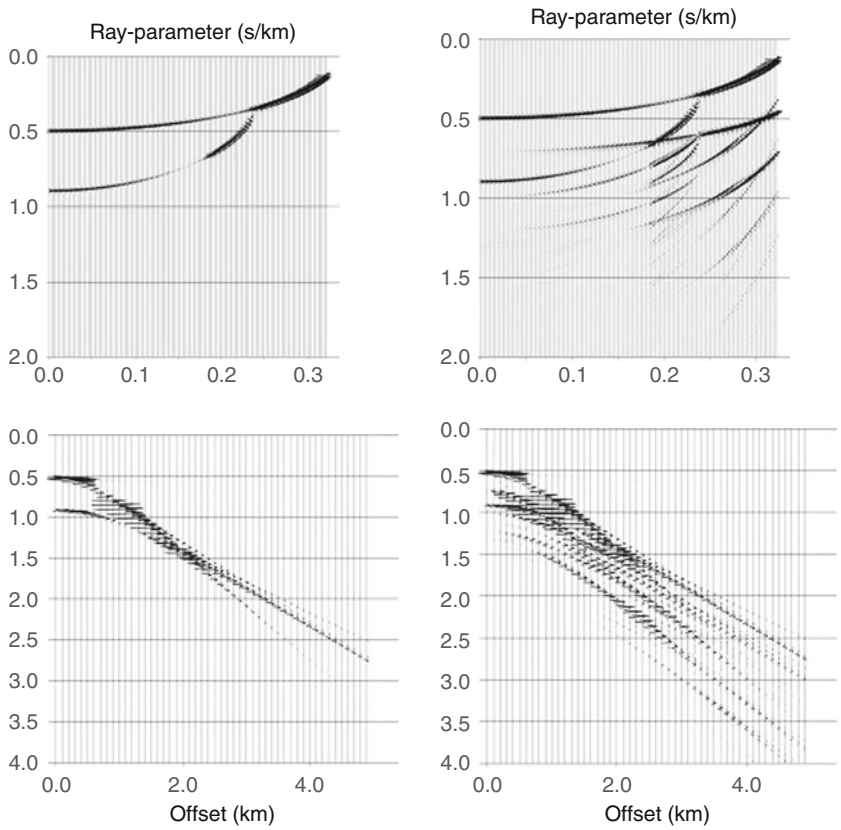
$$R(\omega, p) = R_1^{PS} \exp\left(i\omega h_1\left(q_1^P + q_1^S\right)\right). \tag{6}$$

To include the reflection from the second layer, we simply need to add a term appropriate for the second layer (similar to Equation 3) resulting in the following expression

$$R(\omega, p) = T_{1D}^{PP} R_2^{PP} T_{1U}^{PP} \exp\left(i\omega\left(2h_1 q_1^{PP} + 2h_2 q_2^{PP}\right)\right), \tag{7}$$

where we include the effects of transmission using downgoing (with subscript D) and upgoing (with subscript U) transmission coefficients. Expressions similar to Equation 7 can also be written down for mode-converted waves.

Examples of plane wave and point source synthetic seismograms are shown in Figure 3; these are generated for a vertical point force and only the traces of vertical component of displacement are shown. The upper panel



Seismic, Reflectivity Method, Figure 3 Examples of plane wave and point source synthetic seismograms for a model comprising two elastic layers over a half space: these were generated for a vertical point force and only the traces of vertical component of displacement are shown. The upper panel shows the plane wave seismograms in τ-p domain while the lower panel shows the corresponding x-t domain synthetic seismograms. The left panel shows seismograms for primary *PP* reflections only. The right panel shows the seismograms that include primary *PP*, *PS* and internal multiples from the second layer.

shows the plane wave seismograms in τ-*p* domain while the lower panel shows the corresponding *x*-*t* domain synthetic seismograms. The left panel shows seismograms for primary *PP* reflections only. Note two distinct reflections from the two layer boundaries. The right panel shows seismograms that include primary *PP*, *PS,* and internal multiples from the second layer.

It is fairly straightforward to compute seismograms for a few selected phases for a few layers. In reality, we have many layers and an infinite number of ray-paths – these are not easy to count. The reflectivity method offers an elegant algorithm for computing full response without having to count rays. Of course, we may choose to compute a few selected phases under reflectivity formulation as well, if desired.

Theory

Detailed developments of the theory of the reflectivity method for isotropic media can be found in Fuchs and Muller (1971) and Kennett (1983). Derivations for an anisotropic layered medium can be found in Booth and Crampin (1983) and Fryer and Frazer (1984). We start with the two fundamental equations in seismology, namely, the constitutive relation and the equation of motion given by,

$$\tau = \mathbf{C} : \nabla \mathbf{u} \tag{8}$$

$$\rho\omega^2 \mathbf{u} = \nabla \cdot \tau + \mathbf{f}, \tag{9}$$

where, τ is a second rank stress tensor, **C** is the fourth rank elastic stiffness tensor, **u** is the displacement vector, ω is the angular frequency, and **f** is a body force term. Note that the above equations are in frequency domain.

The fundamental assumption in reflectivity is that the medium is layered and therefore, we can apply transformations over *x* and *y* coordinates to horizontal wave numbers K_x and K_y or horizontal slowness p_x and p_y. For isotropic and layered transversely isotropic media, we can make use of cylindrical symmetry and transform to radial ray parameter *p* and azimuth ϕ. For a general anisotropic medium, we simply apply Fourier transforms over *x* and *y* coordinates to transform Equations 8 and 9 into horizontal wave numbers or ray-parameter domain. Once that is achieved, all the analytic developments are done in the frequency ray-parameter domain.

The equation of motion and the constitutive relation are thus transformed into the following system of ordinary differential equations in depth *z*, by applying a Fourier transform in *x* and *y*

$$\mathbf{b} = i\omega\, \mathbf{A}\mathbf{b} + \mathbf{f}, \tag{10}$$

where, $\mathbf{b} = \left[u_x\, u_z\, u_y\, \tau_{xz}\, \tau_{zz}\, \tau_{yz}\right]^T = \mathbf{b}(\omega, \mathbf{p})$is the stress-displacement vector which is a function of frequency ω and horizontal slowness *p*, **A**(ω, **p**) is the system matrix that is a function of elastic coefficients, and **f** is a body force term. For isotropic and transversely isotropic media, the Equation 10 decouples into two systems, namely a P-SV (4 × 4) system and an SH (2 × 2) system. The solution of the ODE (Equation 10) can be carried out by a propagator matrix method (Gilbert and Backus, 1966). It is well known that the propagator matrix is generally unstable due to growing exponentials (e.g., Jensen et al., 1993). Stable solutions can be obtained by one of the three methods:

- A global matrix approach (Schmidt and Tango, 1986).
- Compound matrix approach where we define a new system of ODE in which the elements of the new system matrix are the minors of the original system matrix. The original 4 × 4 P-SV system maps into a 6 × 6 system (Dunkin, 1965; Phinney et al., 1987).
- An invariant imbedding or a reflection matrix approach (Kennett, 1983).

Of the three methods listed above, the compound matrix and the reflectivity methods have been widely used in seismology. Table 1 summarizes the two methods. The reflectivity method has been very popular because of its ray-interpretation and easy generalization to azimuthally anisotropic media. In the unconditionally stable reflection matrix approach (Kennett, 1983), the propagation uses the eigenvalues and eigenvectors of the system matrix **A**; the eigenvalues are the vertical phase functions. These

Seismic, Reflectivity Method, Table 1 Reflection matrices or compound matrices?

Reflectivity approach	Compound matrix approach
Unconditionally stable	**Unconditionally stable**
For the P-SV case, the algorithm evaluates four 2 × 2 complex reflection/transmission matrices and iteration equations for each frequency and ray-parameter.	For the P-SV case, the algorithm has fewer floating-point operations than the reflectivity method. The method is computationally faster than the reflectivity method.
Derivatives or differential seismograms can be computed analytically for isotropic and vertically transversely isotropic (VTI) media	Derivatives or differential seismograms can be computed analytically for isotropic and VTI media
Derivatives can be computed by semi-analytic approach for general anisotropic media	–
Requires VERY CAREFUL coding	**Coding is very EASY**
Most efficient for general anisotropic media	• Very slow for general anisotropy
Ray-interpretation is possible. Selected rays and selected number of multiples can be included. The algorithm is VERY FLEXIBLE	Complete solution is derived

eigenvalues and eigenvectors are used to define four upgoing and downgoing reflection and transmission coefficient matrices, $\mathbf{R}_D$, $\mathbf{T}_D$, $\mathbf{R}_U$, and $\mathbf{T}_U$. These are propagated through the stack of layers to obtain a composite reflection matrix that includes the effects of reflection, transmission, mode conversion, and internal multiples. Kennett (1983) derived the following iteration equation (Kennett, 1983; p.127), which can be used to compute the R/T matrices of a zone AC when those of zones AB and BC are known:

$$\begin{aligned}
\mathbf{R}_D^{AC} &= \mathbf{R}_D^{AB} + \mathbf{T}_U^{AB}\mathbf{R}_D^{BC}\left[\mathbf{I} - \mathbf{R}_U^{AB}\,\mathbf{R}_D^{BC}\right]^{-1}\mathbf{T}_D^{AB} \\
\mathbf{T}_D^{AC} &= \mathbf{T}_D^{BC}\left[\mathbf{I} - \mathbf{R}_U^{AB}\,\mathbf{R}_D^{BC}\right]^{-1}\mathbf{T}_D^{AB} \\
\mathbf{R}_U^{AC} &= \mathbf{R}_U^{BC} + \mathbf{T}_D^{BC}\mathbf{R}_U^{AB}\left[\mathbf{I} - \mathbf{R}_D^{BC}\,\mathbf{R}_U^{AB}\right]^{-1}\mathbf{T}_U^{BC} \\
\mathbf{T}_U^{AC} &= \mathbf{T}_U^{AB}\left[\mathbf{I} - \mathbf{R}_D^{BC}\,\mathbf{R}_U^{AB}\right]^{-1}\mathbf{T}_U^{BC}.
\end{aligned} \tag{11}$$

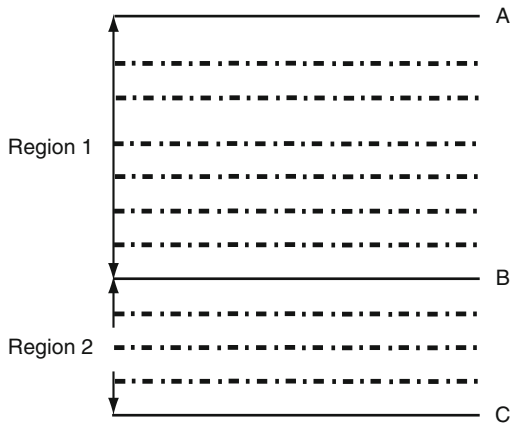
Equation 11 is the most fundamental development of the reflectivity formulation. Given the R/T matrices (which include layer propagation and interaction terms) of two consecutive regions AB and AC (Figure 4), we can compute the response of region AC using Equation 11. Thus, we can propagate through the stack of layers resulting in an unconditionally stable algorithm.

Let us examine the downward reflection matrix in Equation 11 in slightly more detail. Using the matrix identity

$$(\mathbf{I} - \mathbf{A})^{-1} = I + \mathbf{A}\mathbf{A} + \mathbf{A}\mathbf{A}\mathbf{A} + \ldots, \tag{12}$$

we can write

$$\begin{aligned}
\mathbf{R}_D^{AC} &= \mathbf{R}_D^{AB} + \mathbf{T}_U^{AB}\mathbf{R}_D^{BC}\left[\mathbf{I} - \mathbf{R}_U^{AB}\,\mathbf{R}_D^{BC}\right]^{-1}\mathbf{T}_D^{AB} \\
&= \mathbf{R}_D^{AB} + \mathbf{T}_U^{AB}\mathbf{R}_D^{BC}\mathbf{T}_D^{AB} + \mathbf{T}_U^{AB}\mathbf{R}_D^{BC}\mathbf{R}_U^{AB}\,\mathbf{R}_D^{BC}\mathbf{T}_D^{AB} + \ldots
\end{aligned} \tag{13}$$



Seismic, Reflectivity Method, Figure 4 Zones AB and BC from a stack of layers; if the upgoing and downgoing reflection/transmission matrices for these two zones are known, those for the entire stack *AC* can be generated using the iteration equations.

Figure 5 clearly shows that Equation 13 includes all the internal multiples and mode-converted waves. We may also choose to retain a few terms in the expansion and thus compute exact response of a few rays.

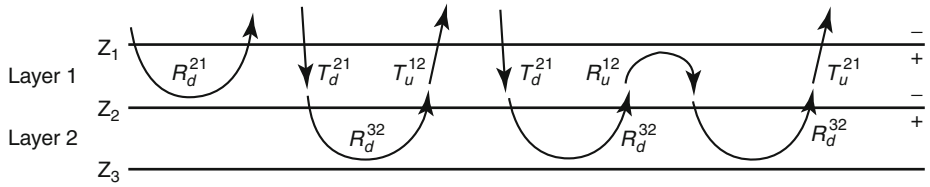
While computing the up/downgoing reflection/transmission matrices through the stack of layers, we need to save the appropriate matrices through the source, receiver layers, and the free surface. Finally, one can compute the entire stress/displacement vector comprising the components of displacement and traction at any receiver location.

Note that we obtain our solution in the frequency ray-parameter domain. An inverse temporal Fourier transforms results in (τ-p) seismograms. Synthetics in the offset-time domain can be obtained by plane wave synthesis of the (ω, p) or (τ-p) seismograms. A general flow chart for computing synthetic seismograms by the reflectivity method is shown in Figure 6. Note that the ray-parameter, layer and frequency loops can be interchanged depending on the application. As stated earlier, intermediate results in (τ-p) domain are very useful for understanding wave propagation and amplitude effects since they are devoid of spherical spreading loss (Fryer, 1980).

Computational issues

It has been demonstrated that the reflectivity algorithm reviewed in this entry is unconditionally stable (Kennett, 1983). However, one must incorporate some practical strategies in computing noise-free synthetic seismograms (e.g., Mallick and Frazer, 1987).

- *Spatial aliasing*: Generation of offset synthetics from plane wave seismograms requires evaluation of an oscillatory integral. The integrand becomes more and more oscillatory with increase in offset and frequency. Thus a practical strategy would be to use a frequency dependent sampling interval in ray-parameters. Alternatively, one can use the minimum ray-parameter increment appropriate for the highest frequency and the largest offset of interest. We also need to use smooth tapers at the high and low ends of the ray-parameter window to avoid truncation phases.
- *Temporal aliasing*: It can be avoided either by using a very large time window or adding a small imaginary part in the frequency following a method outlined in Phinney (1965).
- *Parallelization*: For high frequency and large offset calculation, the reflectivity calculation may be computationally intensive especially for azimuthally anisotropic media, where reflectivity calculation needs to be performed for a large number of ray parameters. Note that (Figure 6) almost the entire reflectivity computation can be done in parallel resulting in an algorithm that may be termed "embarrassingly parallel." One of the simplest ways to parallelize is to distribute the computation of ray-parameter traces equally to the available nodes of a computer (e.g., Roy et al., 2005). The master node collects all the ray parameter traces (note that there is no communication in between the



Seismic, Reflectivity Method, Figure 5 A demonstration of how the iteration equation includes all the reflection and transmission effects.

General Flow chart

- Loop over ray-parameter
  - Loop overe layer
    - Compute R/T (up going and downgoing) for all the interfaces
    - loop over frequency
      - evaluate iteration equation
    - end loop over frequency
  - end loop over layer
  - inverse FFT for (tau-p) seismograms
- End loop over ray-parameter
- Sum the plane wave responses to compute point source synthetics

Seismic, Reflectivity Method, Figure 6 A general flow chart for developing a reflectivity code.

nodes required for the calculation of ray parameter trace). This results in an almost linear speedup as a function of the number of nodes. MPI (message passing interface) can be used for parallelization; the reflectivity code can also be parallelized using simple compiler directives, if coded carefully.

- *Attenuation*: In order to include attenuation, we need to specify Q values for P- and S- waves for each layer. For frequency dependent attenuation, the velocities are complex and dependent on frequency. In such cases, the layer loop (Figure 6) becomes the deepest loop.

Applications

Exploration scale modeling: Numerous applications of reflectivity synthetic seismograms for exploration can be found in literature. This algorithm is often used to generate synthetic seismograms from well logs and then used in calibrating a seismic gather at the well location and well tie (e.g., Sen, 2006). Figure 7a shows α_0, β_0, density, and VTI anisotropy parameters ε and δ derived from a well log as a function of two way time. The plane wave synthetics for an isotropic and VTI model and their differences are shown in Figure 7b. Note that for isotropic calculations, ε and δ values were set equal to zero; for VTI calculations, P- and S-wave velocities from well logs were treated as vertical P and S velocities respectively. Offset synthetics for isotropic and VTI models and their differences are shown in Figure 7c. As expected, the effect of anisotropy is more pronounced at large offsets or ray-parameters.

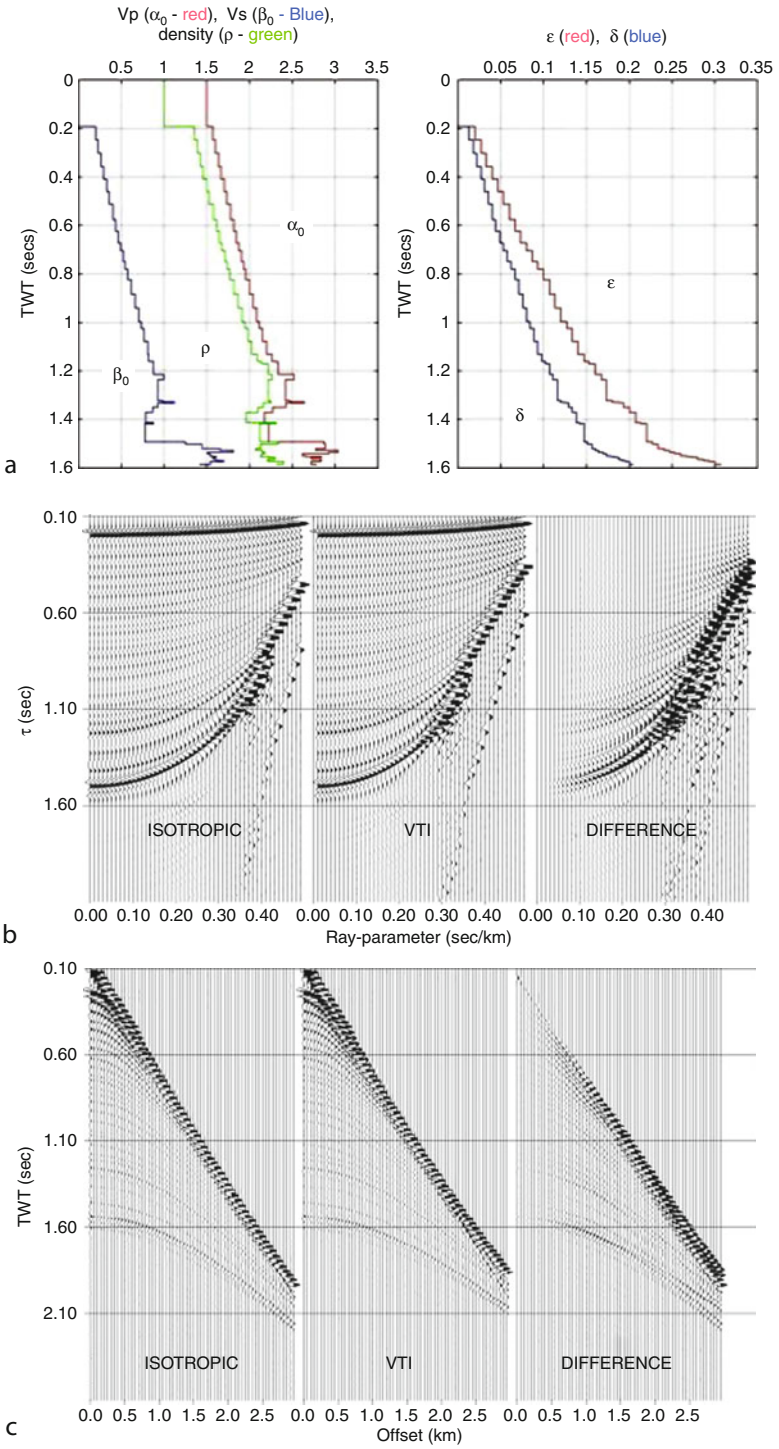
Modeling anisotropy in the core–mantle transition zone (D″): For modeling at regional and teleseismic distances, we need to apply earth-flattening transformation to the models prior to computing synthetic seismograms using a reflectivity code. An example of such modeling is given in Pulliam and Sen (1998) in which attempts were made to model shear wave splitting observations of S-wave phases that propagate nearly horizontally through the core–mantle transition zone. Figure 8 shows the data and the corresponding synthetic seismograms for a station at Hockley from earthquakes from Tonga-Fiji. The model used for reflectivity calculation includes transversely isotropic layers with a horizontal axis of symmetry within the D″ zone.

Modeling in azimuthally anisotropic media (comparison with FD): An example of comparison of azimuthally anisotropic reflectivity synthetic seismograms with those computed by a 3D azimuthally anisotropic finite-difference code is shown in Figure 9 (Bansal and Sen, 2008). The synthetics computed by the two methods are in excellent agreement.

Extension to laterally heterogeneous media

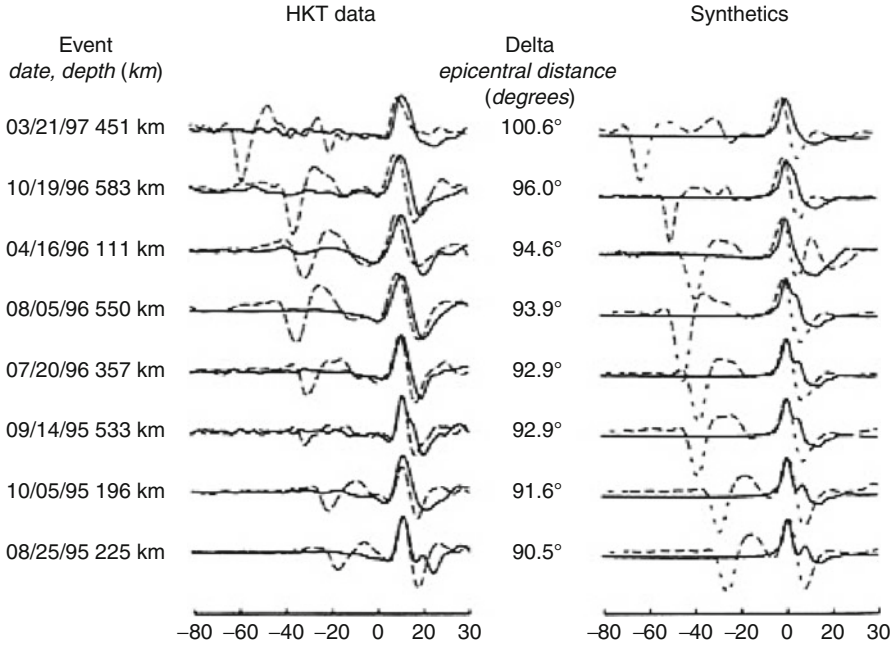
Extension of the reflectivity method to two dimensions, where the velocities and densities are allowed to vary laterally as well, is nontrivial. Koketsu et al. (1991) and Sen and Pal (2009) developed extended reflectivity methods for the case of homogeneous layers separated by irregular or curved interfaces which do not cross each other. Lateral media variations within a layer cannot be easily taken into account in a reflectivity-type formalism because the concept of an interface is inherent in it except using the concepts of pseudo-differential operators and Fourier Integral operators (McCoy et al., 1986).

Across a flat surface, the horizontal slowness p is conserved according to Snell's law. However, across an irregular interface, an incident ray with a particular p is scattered into different plane waves with slownesses p'. This is mathematically equivalent to Fourier transforming the dependence of z on x to the scattered set of p' or k' (wave number). Now the coefficients $\mathbf{R_D}(p)$, $\mathbf{R_U}(p)$, $\mathbf{T_D}(p)$, and $\mathbf{T_U}(p)$ become functions of both p and p'. The 2D extension of the method as proposed by Koketsu et al. (1991) involves explicit evaluation of boundary conditions that results in a matrix formulation involving several matrix inversions in the coupled ray-parameter domain; numerous numerical artifacts are caused by such matrix operations. An alternate asymptotic approach was

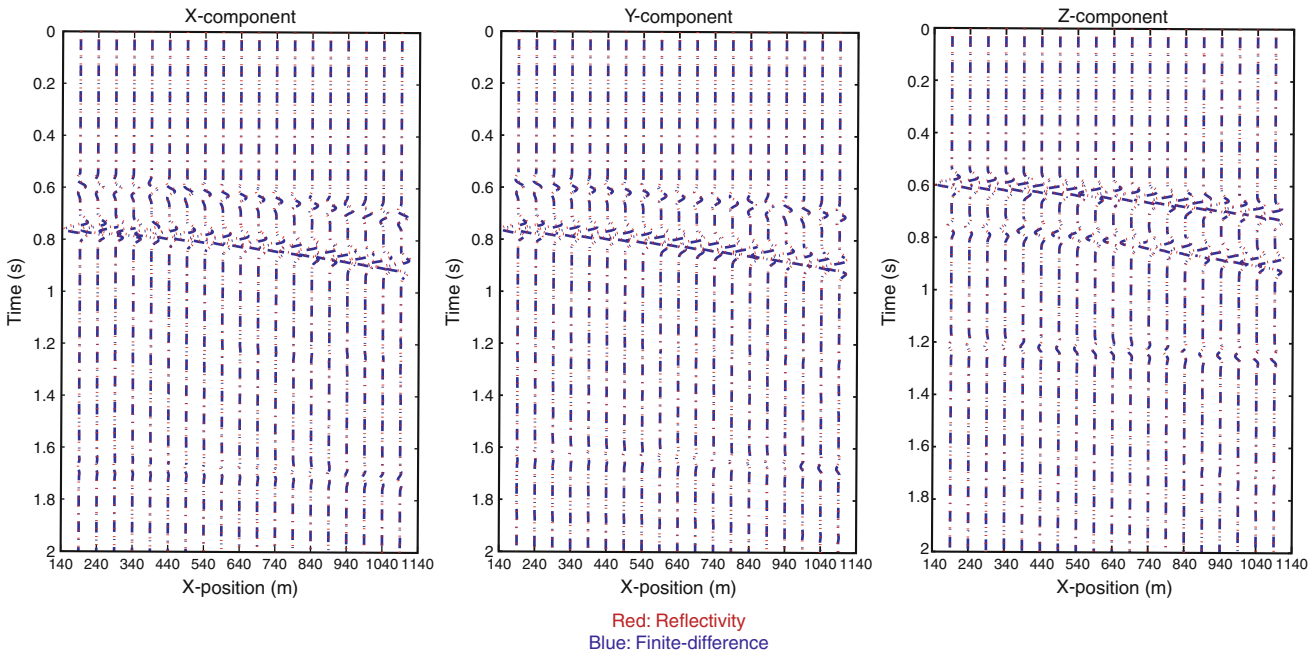


Seismic, Reflectivity Method, Figure 7 (**a**) *Vp*, *Vs*, density, *e* and *d* for a well log model; (**b**) plane wave synthetic seismograms for an isotropic model (*left*), VTI model (*middle*), and the difference (*right*); (**c**) offset synthetic seismograms for an isotropic model (*left*), VTI model (*middle*), and the difference (*right*).

proposed by Sen and Pal (2009) by examining a tangent-plane or Kirchhoff formulation in the plane wave domain to replace the exact boundary condition evaluation (Sen and Frazer, 1991). All other reflectivity operations such as the invariant imbedding or iterative computation of reflection, transmission, multiple, and mode conversions can be readily applied even under this approximation. This new algorithm computes noise-free seismograms even



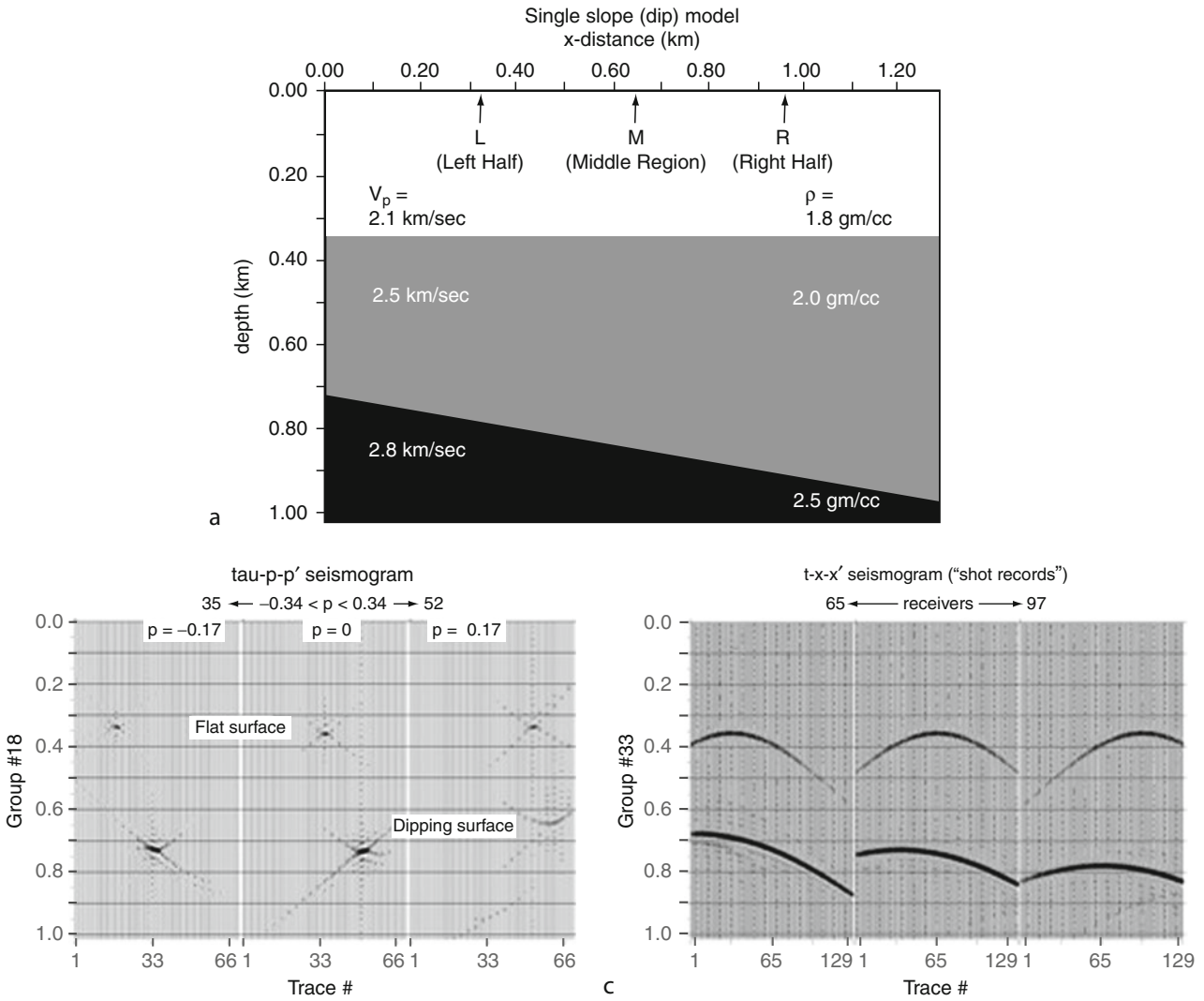
Seismic, Reflectivity Method, Figure 8 Comparison of S-wave arrivals from Tonga-Fiji events recorded at Hockley station. Synthetic seismograms (*right panel*) were computed using a reflectivity code for a model containing transversely isotropic layers with a horizontal axis of symmetry within the core–mantle transition zone. Solid lines represent transverse and dashed lines represent radial component seismograms. Reproduced with permission from Geophysical Journal International.



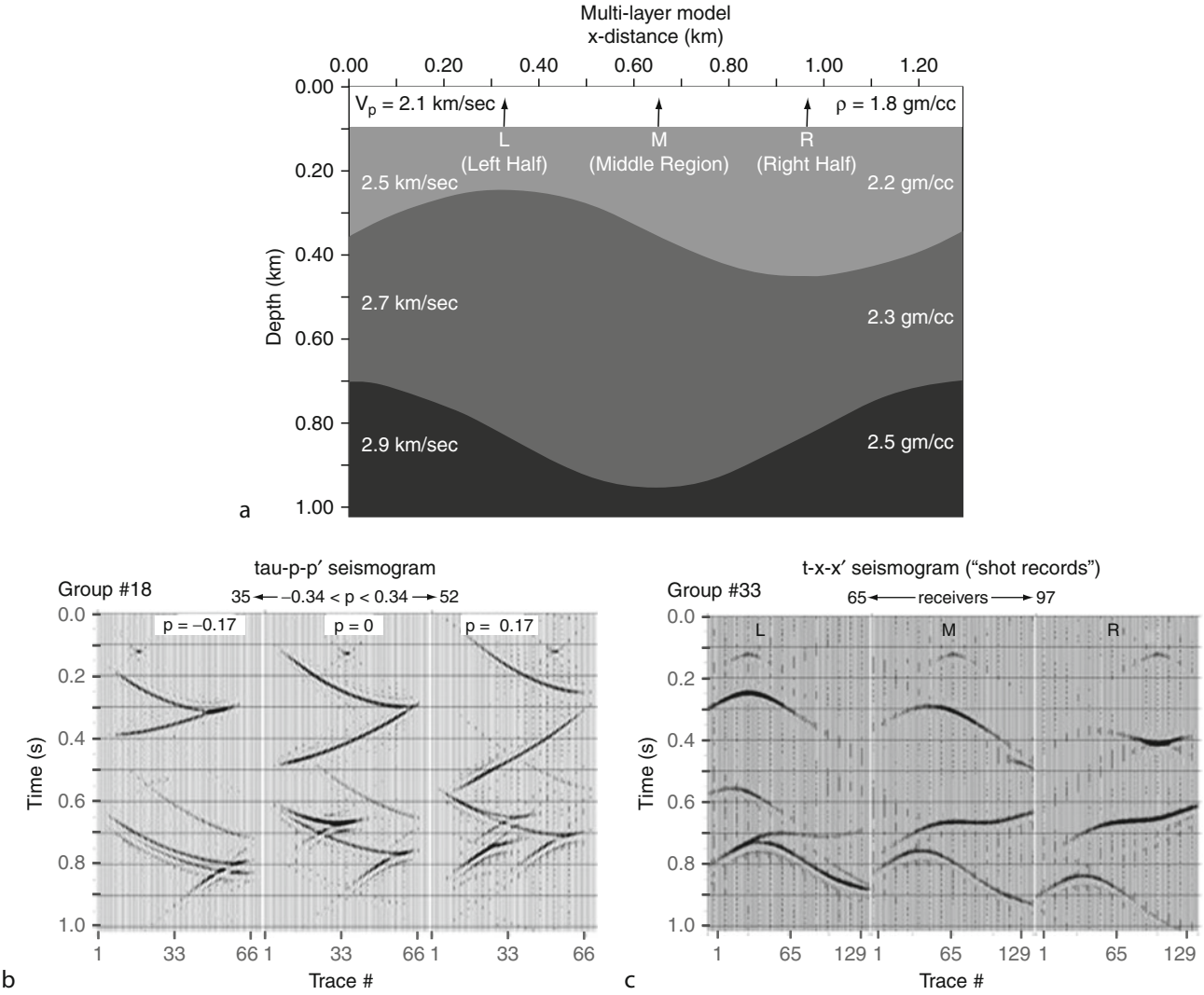
Seismic, Reflectivity Method, Figure 9 A comparison between reflectivity and 3D finite-difference synthetic seismograms for a two-layer azimuthally anisotropic earth model. Red: reflectivity; Blue: finite difference. Reproduced with permission from Geophysical Prospecting.

with coarse sampling of the interfaces and ray-parameters. Approximate calculation of reflection/transmission coefficients, however, does not include multiple interaction of a plane wave with an interface.

Sen and pal (2009) performed computations in the coupled slowness domain and thus, multiple shot-receiver data can be synthesized rapidly. Intermediate results in the coupled slowness space provide important insight into understanding wave propagation in heterogeneous media. Figure 10 shows synthetic seismograms for a model with two homogeneous layers over a half space (Figure 10a) such that the top interface is flat and the second interface is dipping with a constant dip. Figure 10b shows synthetic seismograms in (τ-p-p') domain; note that for the flat layer, the scattered p is the same as the incident ray-parameter and the same for the dipping reflector is different from the incident ray-parameter. Synthetic shot gathers for three shot locations shown in Figure 10a are displayed in Figure 10c. These were all computed using a 2D reflectivity code in which internal multiples were not included. Similarly, Figure 11 shows the synthetic seismograms for a 2 layer over half space model where both the interfaces are curved. Note the triplicated arrivals in the seismograms in both τ-p-p' domain (Figure 11b) and shot gathers (Figure 11c). Note the triplicated arrivals in the seismograms due to curved interfaces.



Seismic, Reflectivity Method, Figure 10 (**a**) A model with two homogeneous layers over a half space; (**b**) τ-p-p' synthetic seismograms for the model shown in (**a**): note the shift in p value for the reflection from the dipping interface; (**c**) shot gathers for the mode shown in (**a**) – asymmetry in the reflection pattern from the second interface is modeled very well. Reproduced with permission from Geophysical Journal International.



Seismic, Reflectivity Method, Figure 11 (**a**) A model with two homogeneous layers over a half space; (**b**) τ-*p*-*p*′ synthetic seismograms for the model shown in (**a**): note the shift in *p* values and triplications for the reflection from the curved interfaces; (**c**) shot gathers for the mode shown in (**a**) – triplications in the reflections modeled very well. Reproduced with permission from Geophysical Journal International.

Summary

Seismic modeling is crucial to understanding seismic wave propagation and interpreting field seismic records. Several methods for seismic modeling are available; they vary in accuracy, model assumptions, and computational speed. Of these, the reflectivity method is perhaps the most popular. The original reflectivity formulation is valid for 1D elastic media. It is capable of computing a complete response of a layered earth model. If needed, this can also be designed to compute responses of a few selected phases. The cost of computation grows with the number of layers, increase in frequency, and offset. Most isotropic models can now be computed fairly rapidly on standard desktop workstations or personal computers. Azimuthally anisotropic models require computations in two wave numbers and therefore can be computationally demanding. Parallel algorithms can be easily developed for rapid computation of reflectivity synthetic seismograms for large models. For 1D full waveform inversion, the reflectivity forward modeling is invoked a large number of times where a parallel reflectivity algorithm is also essential. The reflectivity algorithm has been used in a wide variety of applications including seismic modeling in exploration and whole earth scales. Full waveform inversion methods based on reflectivity forward modeling are becoming increasingly popular (e.g., Sen and Roy, 2003). The method has also been extended to modeling in laterally heterogeneous media and to electromagnetic modeling in layered media (Sena et al., 2008).

Bibliography

Aki, K., and Richards, P. G., 2000. *Quantitative Seismology*. San Francisco: W. H. Freeman.

Bansal, R., and Sen, M. K., 2008. Finite difference of S-wave splitting in anisotropic media. *Geophysical Prospecting*, **56**, 293–312.

Booth, D. C., and Crampin, S., 1983. The anisotropic reflectivity technique: theory. *Geophysical Journal of the Royal Astronomical Society*, **72**(3), 755–766.

Cerveny, V., 2001. *Seismic Ray Theory*. Cambridge: Cambridge University Press.

Chapman, C. H., and Drummond, R., 1982. Body-wave seismograms in inhomogeneous media using Maslov asymptotic theory. *Bulletin. Seismological Society of America*, **72**, S277–S317.

DeBasabe, J., and Sen, M. K., 2007. Grid dispersion and stability criteria of some finite element methods for seismic wave propagation modeling. *Geophysics*, **72**(6), T81–T95.

Dunkin, J. W., 1965. Computation of modal solutions in layered elastic media at high frequencies. *Bulletin. Seismological Society of America*, **55**, 335–338.

Fryer, G. J., 1980. A slowness approach to the reflectivity method of seismogram synthesis. *Geophysical Journal of the Royal Astronomical Society*, **63**(3), 747–758.

Fryer, G. J., and Frazer, L. N., 1984. Seismic waves in stratified anisotropic media. *Geophysical Journal International*, **78**(3), 697–710.

Fuchs, K., and Mueller, G., 1971. Computation of synthetic seismograms with the reflectivity method and comparison with observations. *Geophysical Journal of the Royal Astronomical Society*, **23**, 417–433.

Gilbert, F., and Backus, G. E., 1966. Propagator matrices in elastic wave and vibration problems:. *Geophysics*, **31**, 326–332.

Jensen, F. B., Kuperman, W. A., Porter, M. B., and Schmidt, H., 1993. *Computational Ocean Acoustics*. New York: AIP.

Kennett, B. L. N., 1983. *Seismic Wave Propagation in Stratified Media*. Cambridge: Cambridge University Press.

Koketsu, K., Kennett, B. L. N., and Takenaka, H., 1991. 2-D reflectivity method and synthetic seismograms for irregularly layered structures – II. Invariant embedding approach. *Geophysical Journal International*, **105**, 119–130.

Komatitsch, D., and Tromp, J., 1999. Introduction to the spectral element for three dimensional seismic wave propagation. *Geophysical Journal International*, **139**, 806–822.

Mallick, S., and Frazer, L. N., 1987. Practical aspects of reflectivity modeling. *Geophysics*, **52**, 1355–1364.

Marfurt, K. J., 1984. Accuracy of finite-difference and finite-element modeling of the scalar and elastic wave equations. *Geophysics*, **49**, 533–549.

McCoy, J. J., Fishman, L., and Frazer, L. N., 1986. Reflection and transmission at an interface separating transversely inhomogeneous acoustic half spaces. *Geophysical Journal of the Royal Astronomical Society*, **85**, 543–562.

Phinney, R. A., 1965. Theoretical calculation of spectra of first arrivals in layered elastic mediums. *Journal of Geophysical Research*, **70**, 5107–5123.

Phinney, R. A., Odom, R. I., and Fryer, G. J., 1987. Rapid generation of synthetic seismograms in layered media by vectorization of the algorithm. *Bulletin. Seismological Society of America*, **77**, 2218–2226.

Pulliam, J., and Sen, M. K., 1998. Anisotropy in the core-mantle transition zone may indicate chemical heterogeneity. *Geophysical Journal International*, **135**, 113–128.

Roy, I. G., Sen, M. K., and Torres-verdin, C., 2005. Full waveform inversion using a distributed system of computers. *Concurrency and Computation: Practice and Experience*, **17**, 1635–1685.

Scmidt, H., and Tango, G., 1986. Efficient global matrix approach to the computation synthetic seismograms. *Geophysical Journal of the Royal Astronomical Society*, **84**, 331–359.

Sen, M. K., 2006. *Seismic Inversion*. USA: Society of Petroleum Engineers.

Sen, M. K., and Frazer, L. N., 1991. Multifold phase space path integral synthetic seismograms. *Geophysical Journal International*, **104**, 479–487.

Sen, M. K., and Pal, A., 2009. A reflectivity method for laterally varying media. *Geophysical Journal International*, **178**(2), 792–812.

Sen, M. K., and Roy, I. G., 2003. Computation of differential seismograms and iteration adaptive regularization in pre-stack seismic inversion. *Geophysics*, **68**(6), 2026–2039.

Sena, A., Sen, M. K., and Stoffa, P. L., 2008. Modeling of ground penetrating radar data in stratified media using the reflectivity technique. *Journal of Geophysics and Engineering*, **5**, 129–146.

Virieux, J., 1984. SH-wave propagation in heterogeneous media: velocity-stress finite-difference method. *Geophysics*, **49**, 1933–1957.

Cross-references

Numerical Methods, Finite Difference
Numerical Methods, Finite Element
Seismic Waves, Scattering
Seismic, Ray Theory
Seismic, Waveform Modeling and Tomography

SEISMIC, VISCOELASTIC ATTENUATION

Vernon F. Cormier
Physics Department, University of Connecticut, Storrs, CT, USA

Synonyms

Seismic intrinsic attenuation

Definition

Linear viscoelastic attenuation. The fractional loss of seismic energy in a material in which elastic deformation (strain) induced by one cycle of a seismic wave or mode lags in time the applied stress associated with the wave or mode.

Apparent seismic attenuation. The loss of energy in a propagating seismic wave or standing mode due to viscoelasticity combined with the loss of scattered energy redistributed in time and space by heterogeneity.

Introduction

The amplitude of seismic waves decreases with increasing distance from earthquake, explosion, and impact sources. How this amplitude decreases, how rapidly it occurs, and how it depends on frequency of the seismic waves is fundamentally important to the efforts to describe Earth structure and seismic sources. The decay of amplitude of seismic waves with increasing distance of propagation through earth is known as seismic wave attenuation. The attenuation occurring under high-temperature rheological

conditions in the earth's interior can be called seismic viscoelastic attenuation.

Seismic attenuation and its variation with location within the Earth are useful for determining the anelastic properties of the earth as a function of depth. Seismic attenuation also shows large lateral variations that can be related to lateral variations in geological and geophysical properties not as easily detected by measurement of seismic velocities. In addition to providing information on a physical property, research in seismic attenuation has also been strongly motivated by more practical problems. One problem has been the prediction of ground motion due to probable earthquakes in different regions. The frequency content and decay with distance of this strong ground motion is an important input to the design of earthquake resistant structures and to disaster forecasting (see *Earthquakes, Strong-Ground Motion*). Another problem has been to estimate the size and detectability of underground nuclear tests (see *Seismic Monitoring of Nuclear Explosions*).

How do seismic waves attenuate?

The attenuation of seismic waves is due to three effects: geometric spreading, intrinsic attenuation, and scattering.

Geometric spreading

Geometric spreading leads to an energy density decrease that occurs as an elastic wave front expands with increasing distance from its source. In a homogeneous Earth of constant velocity and density, the geometric spreading of a seismic body wave is proportional to the reciprocal of the distance between source and receiver. In the real Earth, velocity and density vary strongly with depth and less so laterally. Given a model of this variation, however, the geometric spreading of a body wave can be easily calculated (see *Seismic, Ray Theory*).

Intrinsic viscoelastic attenuation

Intrinsic (viscoelastic) attenuation occurs at high temperatures due to internal friction during the passage of an elastic wave. It is controlled by the thermal and defect properties of the medium in which the wave is propagating. It can result in a phase lag between strain and stress giving rise to strain energy dissipation and associated frequency dependence (dispersion) of the relevant modulus or speed of the propagating elastic wave. The microscopic mechanisms of intrinsic attenuation have been described in several different ways, including the resistive and viscous properties of oscillator models of the atoms in crystalline lattices, the movement of interstitial fluids between grain boundaries and cracks (O'Connell and Budiansky, 1977), and the frictional sliding of cracks. Jackson (1993, 2007) reviews laboratory experiments that investigate microscopic mechanisms of intrinsic attenuation. This article concentrates on the measurement of intrinsic attenuation from recordings of seismic waves at great distance.

Scattering attenuation

Scattering attenuation occurs when elastic energy is scattered and redistributed into directions away from the receiver or into waves arriving in later time windows at the receiver (see *Seismic Waves, Scattering*). Scattering takes place by reflection, refraction, and mode conversion of elastic energy by wavelength-scale irregularities in the medium. These irregularities are discontinuous or rapid variations in the velocity and/or density of the medium. In the crust and uppermost mantle, variations in velocity and density can be particularly strong in the lateral as well as the vertical direction.

Linear viscoelasticity

Rheology

A stress is a vector force per unit area applied to a solid. A strain is non-dimensional measure of the deformation of the solid due to the applied stress, such as the change in a length element divided by the original length. The equation that relates stress and strain is sometimes termed the *rheology* or the *constitutive* relation (see *Mantle Viscosity*). A linear viscoelastic rheology can be described by a linear differential equation:

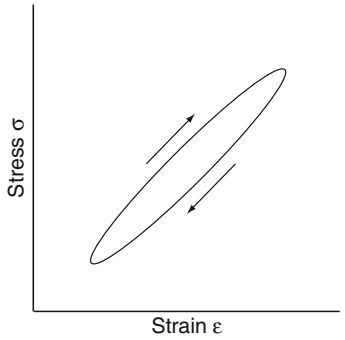
$$L_1\sigma(t) = L_2\varepsilon(t) \tag{1}$$

Where L_1 and L_2 are any linear combinations of operators of the time $\frac{d^n}{dt^n}$ or $\int dt^n$. This type of equation can describe both the elastic strain of a material over a short time interval of applied stress as well as its viscous behavior and flow over a longer time interval (Gross, 1953; Nowick and Berry, 1972; Jackson et al., 2005; Kohistedt, 2007).

Anelastic hysteresis

Seismic oscillations at distances beyond several fault lengths from an earthquake excite small strains less than 10^{-6}. These strains are recoverable during a cycle of seismic oscillation and lag the applied stress of the oscillation in time. Because of the time lag, a cycle of increasing and decreasing stress does not produce a perfectly proportional increase and decrease in strain. Instead a hysteresis loop occurs (Figure 1). The area enclosed by the hysteresis loop is a measure of the energy lost due to heat and internal friction. During the stress cycle associated with the passage of a seismic wave, the energy lost to this internal friction is not available to deform the adjacent regions of the solid ahead of the wave front and therefore the amplitude of the wave decreases.

From the hysteresis curve, one can see that the stress–strain relation cannot be described by a simple constant of proportionality in the time domain. A more complicated relation involving an integral over time is required to describe strain at any instant of time as a function of the prior time history of the applied stress. By Fourier transforming the rheologic equation, however, and keeping only terms describing the short-term anelastic behavior, the stress–strain relation can be simply expressed



Seismic, Viscoelastic Attenuation, Figure 1 Stress–strain hysteresis curve showing the behavior of strain during a cycle of applied stress induced by a propagating seismic wave.

by means of either a complex elastic modulus $\hat{G}(\omega)$ or by its reciprocal, the complex elastic compliance, $\hat{J}(\omega)$:

$$\hat{\sigma}(\omega) = \hat{G}(\omega)\varepsilon(\omega) \tag{2a}$$

$$\hat{\varepsilon}(\omega) = \hat{J}(\omega)\hat{\sigma}(\omega) \tag{2b}$$

The elastic modulus $\hat{G}$ and compliance $\hat{J}$ must be a complex numbers to describe the phase lag of strain. $\hat{G}$ and $\hat{J}$ must also be frequency dependent because the phase lag of strain depends on the time history of stress, the shape of the hysteresis curve changing with different load histories. All the usual measures of anelasticity, including the frequency-dependent quality factor $Q(\omega)$ and the dispersion of the complex phase velocity $\hat{v}(\omega)$ can be expressed in terms of the complex compliance $\hat{J}(\omega)$ (Jackson and Anderson, 1970).The trend of the frequency dependencies can be inferred from the time lag of strain from applied stress.

A feature of the complex modulus is that its real part will be smaller at zero or very low frequency and larger at infinite or very high frequency. That is, there will be an instantaneous response of strain to the applied stress, which is smaller than the eventual equilibrium response after longer time. The difference between the modulus at infinite frequency $G(\infty)$, representing the instantaneous or *unrelaxed* response, and the low-frequency limit of the modulus $G(0)$, for the equilibrium or *relaxed* response, is called the *modulus defec*t ΔG, with

$$\Delta G = G(\infty) - G(0) \tag{3}$$

The relaxed and unrelaxed moduli are pure real numbers that can be determined by observing a sequence of hysteresis curves for increasing frequencies of monochromatic loads. The frequency dependence of the real part of the modulus G at frequencies between 0 and ∞ implies that the propagation of a stress pulse will be dispersive, with higher frequencies traveling faster than lower frequencies.

Q and complex velocity

Since simple mechanical systems, composed of springs and dashpots and simple electric circuits also obey linear equations of the form of Equations 2a, b, there are analogies between the quantities describing these systems and quantities in the stress–strain relation. For example, strain behaves like voltage, stress like current, and the complex compliance $\hat{J}$ like the complex impedance of an electric circuit. Similar to the resonance phenomenon in circuits and mechanical systems, a Q can be defined by the average energy W per cycle divided by the energy lost or work done per cycle, ΔW:

$$Q = \frac{W}{\Delta W} \tag{4}$$

Large Q's imply small energy loss; small Q's imply large loss. Q is a measure of the area contained in the hysteresis loop of a stress–strain cycle. The inverse of (Equation 4), Q^{-1}, is sometimes simply termed the *attenuation or internal friction* (Knopoff, 1964*)*.

Plane waves of frequency ω and propagating in the + or − direction can be defined by the phasor $\exp(i\omega t - \hat{k}t)$ where $\hat{k}$ is a complex wave number $\frac{\omega}{\hat{c}}$ and $\hat{c}$ is a complex velocity defined from the local density ρ and complex modulus $\hat{G}$, with

$$\hat{c} = \sqrt{\frac{\hat{G}}{\rho}} \tag{5}$$

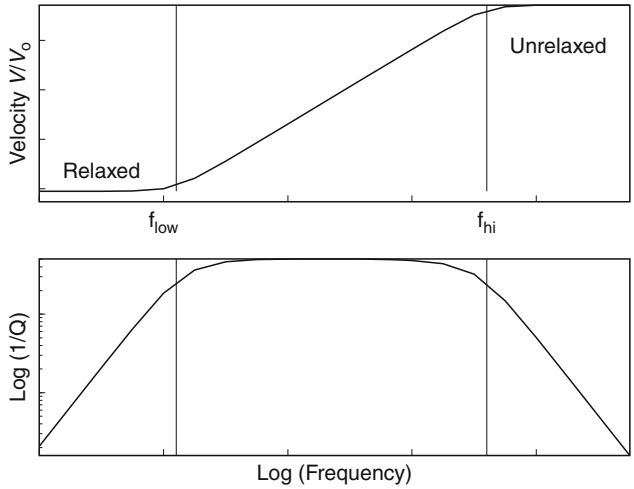
From the average energy density and loss per cycle of a complex plane wave, it can be shown that $Q = \frac{\mathrm{Re}(\hat{G})}{\mathrm{Im}(\hat{G})}$. It is often less confusing to report the reciprocal parameter Q^{-1}, which represents the usually small perturbations to perfect elasticity.

The $Q^{-1}(\omega)$ relaxation spectrum

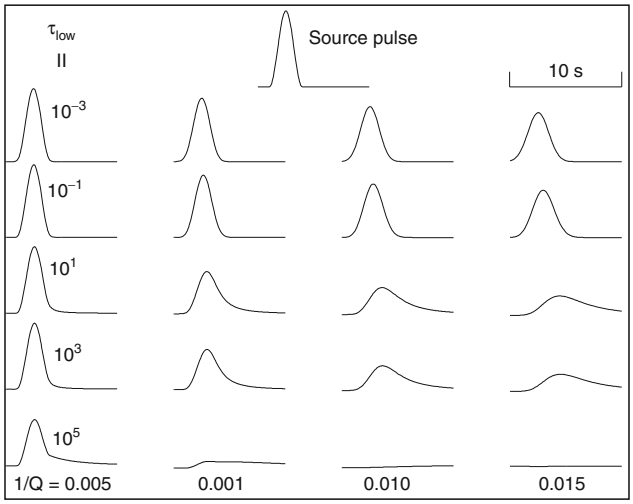
Since $\hat{G}$ depends on frequency, Q also depends on frequency. Zener (1960) described the frequency-dependent effects on an elastic modulus of a solid having a single characteristic time τ for the relaxation of stress. A distribution of relaxation times can be constructed to give a Q^{-1} having a general dependence on frequency. The function $Q^{-1}(\omega)$ is called the *relaxation spectrum*. In the Earth and in many solid materials, the relaxation spectrum is observed to be slowly varying and nearly constant over a broadband of frequencies. A theoretical requirement is that the attenuation Q^{-1} cannot increase faster than ω^1 or decrease faster than ω^{-1}. Figure 2 shows how a continuous distribution of relaxations can produce a Q^{-1} that is nearly constant with a frequency over a broadband. Once the limits of an absorption band are specified, however, it is not possible to have an arbitrarily high Q^{-1} (low viscoelastic Q) over an arbitrarily broad-frequency band without making an unrealistically large modulus defect ΔG. Measured modulus defects in shear are typically less than 25%.

Velocity dispersion

Although the dispersion in elastic moduli had long been known and predicted from the theories of viscoelasticity,



Seismic, Viscoelastic Attenuation, Figure 2 Viscoelastic dispersion of seismic velocity (*top*) and attenuation (*bottom*) showing a relaxation spectrum constant with frequency between two corner frequencies.



Seismic, Viscoelastic Attenuation, Figure 3 Pulse distortion showing the effects of viscoelastic dispersion for variable low-frequency corner and peak attenuation.

it was only widely recognized in seismology when velocity models determined in the low-frequency band from the normal modes of the earth (0.0001–0.01 Hz) were compared with velocity models determined in a high-frequency band (0.1–10 Hz) of body waves (Dziewonski and Anderson, 1981). The models were found to differ and the difference was found to agree with the amount of dispersion predicted from average Q models of the Earth. For example, since the preliminary reference Earth model (PREM), was derived from observations of both the travel times of body waves as well as the eigen frequencies of free oscillations, it reports velocities referenced at both 0.001 Hz and at 1 Hz.

Another more subtle effect of this velocity dispersion can be seen in the propagation of pulses as body waves. A stress disturbance that propagates from its point of initiation as a symmetric narrow Gaussian or triangle-shaped function in time gradually evolves into an asymmetric pulse (Figure 3). High frequencies traveling faster than low frequencies are preferentially loaded into the front of the pulse (Futterman, 1962; Carpenter, 1967). Common theories for the physical mechanism of earthquakes as either frictional slip on a plane or a propagating crack triggered by tectonic stress often predict a far-field displacement pulse that has either a different or opposite form of asymmetry than that predicted for the effect of viscoelastic attenuation. These differences can assist in separating the effects of the source-time history from the effects of viscoelastic attenuation.

Effects of scattering

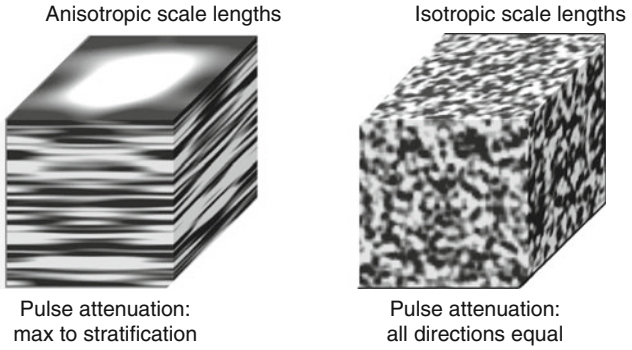
Equivalent medium

At frequencies that are so low that wavelengths are much larger than the characteristic scales of heterogeneity, the attenuative effects of scattering can usually be neglected. At sufficiently low frequency, little energy is lost to scattering, and the medium behaves like an equivalent medium, having properties that are an average of small-scale heterogeneities.

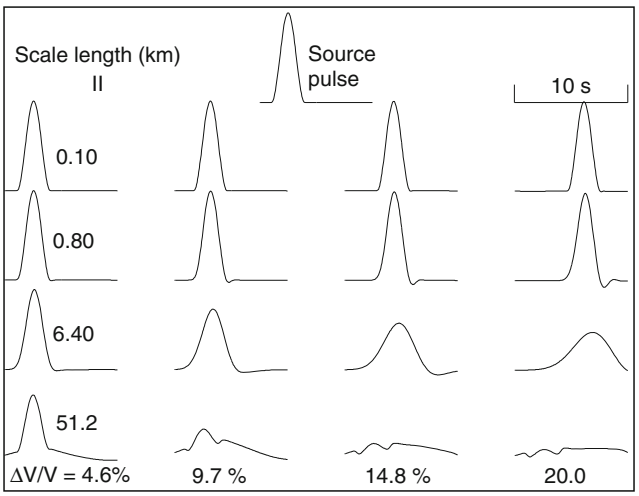
Stochastic dispersion

The most complicated domain in which to perform calculations is where the wavelength is of the order of the scale length of the heterogeneity (Figure 4). In this domain, the presence of heterogeneities can profoundly alter the propagation of the wavefield, both the initial cycle of a body wave pulse as well as the motion immediately following the initial cycle or *coda*. The effects of scattering can be calculated in a one-dimensional medium consisting of thin planar layers in which the velocity in each layer is assigned randomly (O'Doherty and Anstey, 1971; Richards and Menke, 1983). A prediction of such experiments is that body waves will exhibit a stochastic dispersion in which high-frequency energy is transferred into the coda following the first several cycles. This stochastic dispersion may have some biasing effects on measures of intrinsic attenuation. In measures of the spectrum taken over a narrow time window, different results can be obtained, depending on the length of window analyzed, with less attenuation of higher frequencies estimated from longer time windows.

Pulse measurements such as width and rise time may also be biased because higher-frequency energy has been transferred out of the pulse into the later coda. This behavior is opposite to the effects of intrinsic attenuation on a propagating pulse, in which higher frequencies arrive at the beginning of the pulse. A symmetrically shaped displacement source pulse loses less of its symmetry as it propagates through the heterogeneous medium (Figure 5). Anisotropy of the scale lengths of heterogeneity can also be important factor (Hong and Wu, 2005), attenuation being strongest for paths for which the



Seismic, Viscoelastic Attenuation, Figure 4 Example heterogeneity in the Earth and the directional dependence of attenuation of a body wave pulse for wavelengths that are approximately equal to either the dominant scale length (isotropic heterogeneity) or the dominant scale length in the direction of propagation (anisotropic heterogeneity).



Seismic, Viscoelastic Attenuation, Figure 5 Pulse distortion showing the effects of scattering attenuation for variable scale lengths and velocity perturbation calculated by Cormier and Li (2002) using the Dynamic Composite Elastic Modulus (DYCEM) theory of Kaelin and Johnson (1998).

wavelength is of the order of the characteristic scale length in the medium in that direction.

Effects of anisotropy

The existence of general anisotropy in the real part of the elastic modulus has the potential to bias some estimates of anelastic attenuation from either shear wave pulses or surface waves. In a medium having general anisotropy, the decompositions of shear wave motion into SH and SV motion will each contain the interference of two orthogonal shear wave polarizations that are neither SH nor SV (see *Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting*). The broadening of the SH component due to the interference of two quasi-S waves arriving close in time can be mistaken for the broadening due to anelastic attenuation. The regions of the deep Earth characterized by the strongest elastic anisotropy are the upper 400 km of the mantle (Silver, 1996) and the lowermost 400 km of the mantle near the core-mantle boundary (Panning and Romanowiz, 2006). The effects of elastic anisotropy must be removed by combined analysis of SV and SH components of motion, resolving the polarizations of two quasi-S waves, before viscoelastic attenuation can be properly measured.

Measurement and modeling attenuation

Measurements of amplitude of seismic waves may be taken directly from seismograms or from their frequency spectra. To measure the attenuation, we must predict its effects from a model and vary the parameters of the model to fit the observed amplitude, amplitude ratio, or waveform. The effects of intrinsic attenuation in any modeling algorithm operating in the frequency domain can be simply obtained by allowing elastic moduli and/propagation velocities to become complex. Elastic boundary conditions, reflection and transmission at boundaries, travel times, and amplitudes are calculated exactly as in a non-attenuating solid but with elastic moduli and associated velocities analytically continued to complex values. This step of analytic continuation of real moduli to complex moduli is the same whether one wishes to predict the waveform of a body wave or surface wave or spectrum of free oscillations. The size of the imaginary part of the elastic moduli, parameterized by the value of Q^{-1}as a function of depth and frequency, is chosen to match an observed waveform, spectrum, amplitude ratio, or spectral ratio.

The attenuation operator for body waves

As an example of these procedures, consider an experiment with body waves. The effects on a body wave of source radiation, geometric spreading, reflection-transmission, and intrinsic attenuation are most conveniently expressed in the frequency domain by a product of complex functions. The complex $\hat{\Omega}(\vec{x},\omega)$ spectrum of a body wave propagation from a point $\vec{x}_o$ to a receiver at $\vec{x}$ is

$$\hat{\Omega}(\vec{x},\omega) = \hat{B}(\vec{x}_o,\vec{x},\omega)\,\hat{S}(\omega)\,\hat{A}(\omega) \tag{6}$$

The function $\hat{S}(\omega)$ is the Fourier transform of the source-time function. $\hat{B}(\vec{x}_o,\vec{x},\omega)$ incorporates a product of reflection-transmission coefficients, reverberations at source and receiver, geometric spreading, and source radiation pattern. $\hat{A}(\omega)$ is defined by

$$\hat{A}(\omega) = \exp[i\omega\,\hat{T}(\omega)] \tag{7}$$

where $\hat{T}(\omega)$ is the complex travel time obtained by integrating the reciprocal of complex velocity along a ray or normal to the wave front of the body wave:

$$\hat{T}(\omega) = \int_{\text{ray}} \hat{c}(\omega)\mathrm{d}s \tag{8}$$

For body waves, the dominant effect of attenuation on amplitude and phase is given by $\hat{A}(\omega)$. The effects of attenuation on reflection-transmission coefficients and geometric spreading, which have been lumped into $\hat{B}$ are much smaller and can be neglected unless the attenuation is very large (Q is very small). For $Q >> 1$, $\hat{A}(\omega)$ can be rewritten as

$$\hat{A}(\omega) = \exp\left[\frac{-\omega t*(\omega)}{2}\right] \exp\left\{i\omega\left[\mathrm{Re}\ \hat{T}(\infty) - \frac{H[t*(\omega)]}{2}\right]\right\} \tag{9}$$

where

$$t*(\omega) = \int_{\text{ray}} \frac{Q^{-1}}{\hat{c}(\omega)} ds \tag{10}$$

In Equation 9, the attenuation effect is contained in the factor $\exp\left[\frac{-\omega t*(\omega)}{2}\right]$, and the dispersive effect is in the factor $\exp\left\{i\omega\left[\mathrm{Re}\ \hat{T}(\infty) - \frac{H[t*(\omega)]}{2}\right]\right\}$. The operator H is a Hilbert transform. In a band of frequencies in which Q and t^* are nearly constant

$$H[t*(\omega)]/2 = \frac{\ln(\omega/\omega_0)}{\pi}\ t* \tag{11}$$

where ω_0 is a reference frequency contained in the band (Liu et al., 1976). The value of $T(\infty)$ need not be known and can be replaced by some reference time or predicted from an Earth model for the phase being analyzed. The Hilbert transform relation in Equation 11 for the dispersive phase of $\hat{A}(\omega)$ says that $\hat{A}(\omega)$ must be a minimum phase filter in the frequency domain. In general, the Fourier transform of the source-time function, $\hat{S}(\omega)$, is not a minimum phase filter, which can help in the separation and discrimination of the source spectrum from the effects of $\hat{A}(\omega)$ in the total expression for the far-field spectrum $\Omega(\vec{x}, \omega)$.

The phase given by Equation 11 will be accurate only between and far from the low- and high-frequency corners of the relaxation spectrum. Accurate representations of $\hat{A}(\omega)$ across a broad-frequency band can be obtained for general relaxation spectra by substituting expressions for complex velocity $\hat{c}(\omega)$ in Equation 8 obtained by superposing multiple Zener relaxations centered on single relaxation times whose strength is varied to achieve a desired shape for the relaxation spectrum. A useful expression for $\hat{c}(\omega)$ that is accurate for all frequencies across a relaxation spectrum, which is flat between two corner frequencies, can be derived from formulae for complex modulus given by Minster (1978), and is

$$\hat{c}(\omega) = c_{\text{ref}}(\omega_0)\frac{\sqrt{1 + 2\pi Q^{-1}\ln[\psi(\omega)]}}{\mathrm{Re}\sqrt{1 + 2\pi Q^{-1}\ln[\psi(\omega_0)]}} \tag{12a}$$

where

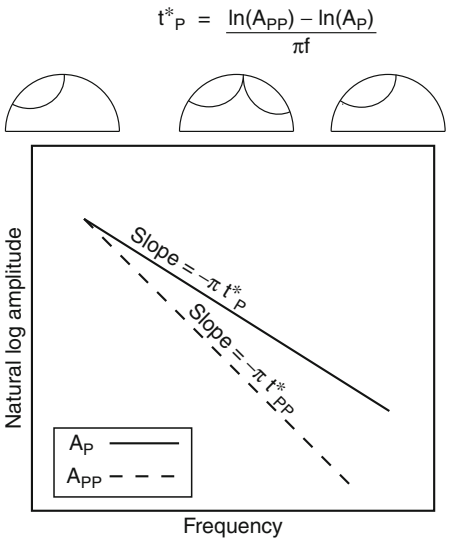
$$\psi(\omega) = \frac{i\omega + 1/\tau_1}{i\omega + 1/\tau_2} \tag{12b}$$

with τ_1 and τ_2 the relaxation times corresponding to the low and high frequency corners respectively. $c_{ref}(\omega_0)$ is a real velocity at the reference frequency ω_0.

Most measurements of attenuation attempt to measure only the amplitude effect of attenuation through the term $\exp[-\omega t*(\omega)/2]$ from the spectral shape of body waves. There are basically two types of experiments commonly reported: matching of (1) spectral decay rates and (2) spectral ratios. In experiment (1) a shape for the displacement source spectrum $S(\omega)$ is assumed usually to be a flat level followed by a decay of ω^{-2} above a corner frequency. The additional decay observed at high frequencies in data spectra is taken as a measure of $t*$ in $\exp[-\omega t*(\omega)/2]$. In experiment (2), a ratio of two different seismic phases from the same source is observed in which the source spectrum is assumed to approximately cancel and the factor related to ratios of geometric spreading and near source and receiver crustal reverberations can be assumed to contribute a simple constant scalar factor. If the phases analyzed are recorded at the same receiver and are incident at nearly the same angles, then crustal reverberations at the source and receiver will approximately cancel. Both types of experiments usually apply some type of smoothing to the spectra to remove biasing effects of spectral holes caused by interfering crustal multiples, source complexities, scattering, and multipathing that are not included in the simple propagation model. Figure 6 illustrates an attenuation experiment of this type.

Since t^* measures only the path-integrated effect of attenuation, many such experiments for different ray paths, bottoming at a range of different depths, are needed to construct a model of Q as a function of depth. Serious breakdowns in this approach, however, exist for cases in which the factorization of the observed spectrum into a product of a geometric spreading, source spectrum, and crustal effects is no longer accurate. One such case is when the body waves in question experience frequency-dependent effects of diffraction near caustics or grazing incidence to discontinuities. The spectral ratios of PKnKP waves, for example, are dominated by the effects of frequency-dependent reflection and transmission coefficients at grazing incidence to the core-mantle boundary. Instead of decreasing linearly with increasing frequency, an observed spectral ratio increases with frequency and exhibits a curvature in a log-log plot, which is consistent with a Q near infinity ($Q^{-1} = 0$) in the other core (Cormier and Richards, 1976).

It is becoming more common to model and invert for viscoelastic attenuation parameters in the time domain, including not only the magnitude of the viscoelastic attenuation parameter Q^{-1}, but also its frequency dependence. Examples of such a study are the inversions for Q^{-1} in the inner core assuming either a viscoelastic



Seismic, Viscoelastic Attenuation, Figure 6 The steps (*top*) to measure the path-integrated attenuation t^* of P waves in the mantle from a log-log plot (*bottom*) of stacked PP and P spectra (A_{PP} and A_P). The distances of observed P and PP spectra are chosen such that each turning ray path of PP is identical in shape and length to that of the single turning ray path of P in the mantle (Adapted from figures in Warren and Shearer, 2000).

(Li and Cormier, 2002) or a scattering origin of attenuation (Cormier and Li, 2002). In these studies, the combined effects of mantle attenuation and source-time function were first modeled by fitting P waves observed in the great circle range 30–90°. Attenuation in the liquid outer core was assumed to be zero. Parameters defining a viscoelastic relaxation spectrum in the inner core were then varied to match the observed PKIKP waveforms. Care must be taken to examine a broad range of attenuation parameters because waveform inversions of this type are very non-linear.

Free oscillations and surface waves

Measurements of attenuation in the low-frequency band of the free oscillations of the Earth are conducted in the frequency domain by observing the width of the individual resonance peaks associated with each mode. These measurements face special problems associated with the broadening produced by lateral heterogeneity of elastic Earth structure. This heterogeneity splits the degenerate modes of a radially symmetric Earth, making a set of modes that would have the same frequency have slightly different frequencies. The slightly different frequencies of the split modes may not be easily resolved in the data spectra and can be confused with the broadening of a single resonance peak of a mode caused by attenuation.

Lateral heterogeneity also complicates the measurement of viscoelastic attenuation of surface waves. Heterogeneity introduces focusing, defocusing, and multipathing, all of which must be accurately modeled to understand the separate attenuative effects of viscoelasticity.

The frequency band of free-oscillation and surface waves (0.001–0.1 Hz), however, offers the best hope of obtaining radially symmetric whole-Earth models of viscoelastic attenuation in this frequency band. This is because lateral variations in attenuation structure are averaged by the gravest modes of oscillation and surface waves that make multiple circuits around the Earth. Computational advances have made the division between free-oscillation and surface wave studies fuzzier, with common approaches now amounting to time-domain modeling of complete low-frequency (<0.1 Hz) seismograms for combined three-dimensional models of viscoelasticity and heterogeneity.

Numerical modeling

Fully numerical modeling of the seismic wavefield allows the combined effects of heterogeneity and viscoelasticity in three-dimensions to be predicted. If the numerical technique is formulated in the frequency domain, substituting a complex velocity for an assumed relaxation spectrum can incorporate viscoelastic attenuation.

If the technique is formulated in the time domain by a finite difference approach, it is neither simple nor efficient to incorporate attenuation by convolution of the wavefield calculated in a non-attenuating medium with an attenuation operator $A(t)$ for individual waves propagating in the attenuating medium, where $A(t)$ is the Fourier transform of $\widehat{A}(\omega)$ defined in Equation 9. Instead, time-domain *memory functions* can be defined to describe a viscoelastic relaxation (Robertsson et al., 1994; Blanch et al., 1995) that can be integrated over time simultaneously with the equations describing particle velocity or displacement and stress. In practice, only three-memory functions, distributed evenly over the logarithm of their characteristic times, are required to simulate a broad frequency band in which Q^{-1} varies slowly.

Interpretation of attenuation measurements in the earth

Shear versus bulk attenuation

In the most general theory of viscoleasticity, it is possible to have with energy loss to occur during both a cycle of volumetric strain as well as shear strain. Since the velocity of a P wave depends on both the bulk and shear moduli, the attenuation Q_P^{-1} of a P wave can be written as a linear combination of the attenuations Q_K^{-1} and Q_S^{-1} defined from complex shear and bulk moduli:

$$Q_P^{-1} = L\,Q_S^{-1} + (1 - L)\,Q_K^{-1} \tag{13}$$

where $(4/3)(V_S/V_P)^2$ and V_P and V_S are the compressional and shear velocities respectively (Anderson, 1989). Although plausible mechanisms for defects in bulk moduli have been found in both laboratory measurements and analytic models of specific attenuation mechanisms, measurements on real data find that bulk dissipation in

the earth is small and, in most cases, can be neglected. One exception may occur when the pressure and temperature state in a narrow depth regions of the earth are close to those near a phase transition, either solid–liquid (Stevenson, 1983) or solid–solid (Ricard and Chambat, 2009). Except for these regions, intrinsic attenuation occurs almost entirely in shear, associated with lateral movement of lattice defects, grain boundaries and/or fluids rather than with changes in material volume. Hence, for viscoelastic attenuation purely in shear in a Poisson solid, for which $V_P = \sqrt{3}\, V_S$,

$$Q_P^{-1} = \frac{4}{9}\, Q_s^{-1} \tag{14}$$

and the parameter for path-integrated attenuation of S waves or $t*_S$ is approximately $4\, t*_P$. Most experiments confirm these values. There is a suggestion, however, that the apparent Q_P^{-1} tends to approach Q_S^{-1} and $t*_S < 4\, t*_P$ at frequencies higher than 1 Hz. These observations are likely evidence of scattering rather than of bulk attenuation because the effects of scattering increase at higher frequencies. With scattering, the apparent Q_S^{-1} tends to approach the apparent Q_P^{-1}, especially when they are measured from pulse widths or spectra taken in a frequency band and medium for which wavelengths are of the order of richest heterogeneity scale lengths of the medium. Thus, the assumption of viscoelastic attenuation occurring mainly in shear can aid in separating the effects of scattering from intrinsic attenuation in body wave pulses.

Frequency dependence

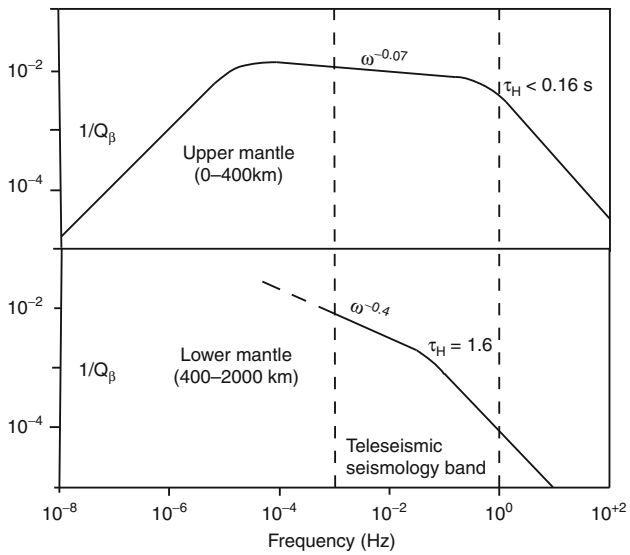
When the results of attenuation measurements determined from free-oscillations and body waves in the 0.0001–0.1 Hz band began to be compared with observations of body wave spectra in the 1–10 Hz band, it became apparent that even under the assumption of a white source spectrum that an increase in Q with frequency was necessary to explain the amplitude of spectra in the 1–10 Hz band.

Thermal activation

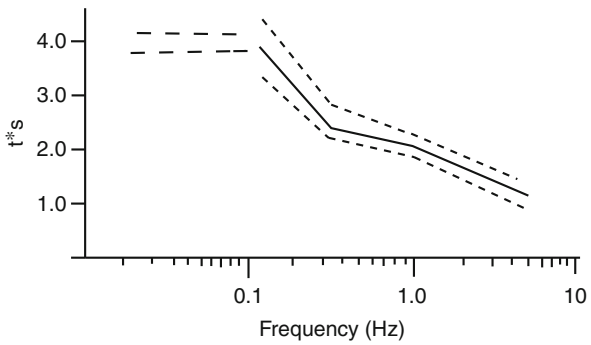
Frequency dependence of viscoelastic attenuation has been interpreted in terms of physical mechanisms of attenuation that are thermally activated. In these mechanisms, the low-frequency corner f_L is tied to a relaxation time τ_L, where $f_L = 1/(2\,\pi\,\tau_L)$. The time τ_L depends on temperature T and pressure P as follows:

$$\tau_L = \tau_0 \exp\left(\frac{E* + PV*}{RT}\right) \tag{15}$$

where $E*$ and $V*$ are the activation energy and volume, respectively. Both the low- and high-frequency corners (f_L, f_H) of an absorption band are assumed to be similarly affected, temperature and pressure acting to slide the absorption band through a band of frequencies. A typical width to expect for the relaxation spectrum of the mantle is about 5 orders of magnitude in frequency, $\tau_L/\tau_H = 10^5$ (Minster and Anderson, 1981; Anderson and Given, 1982). A simplified model of an absorption band with depth in the earth's mantle is shown in Figure 7. The movement of the absorption band toward lower frequencies longer periods in the mantle below 400 km depth is consistent with the type of behavior shown in Figure 8 for the $t*_S$ measured from shear waves of an earthquake. The difference in the location of the absorption band with respect to the band of seismic frequencies is consistent with models of the temperature and pressure profiles of



Seismic, Viscoelastic Attenuation, Figure 7 A frequency and depth-dependent model of shear attenuation in the Earth's mantle derived from modeling broadband shear waves. This model has been used by the National Earthquake Information Center (NEIC) to correct for viscoelastic attenuation in the reported radiated elastic energy from earthquakes (Boatwright and Choy, 1986; Choy and Boatwright, 1995).



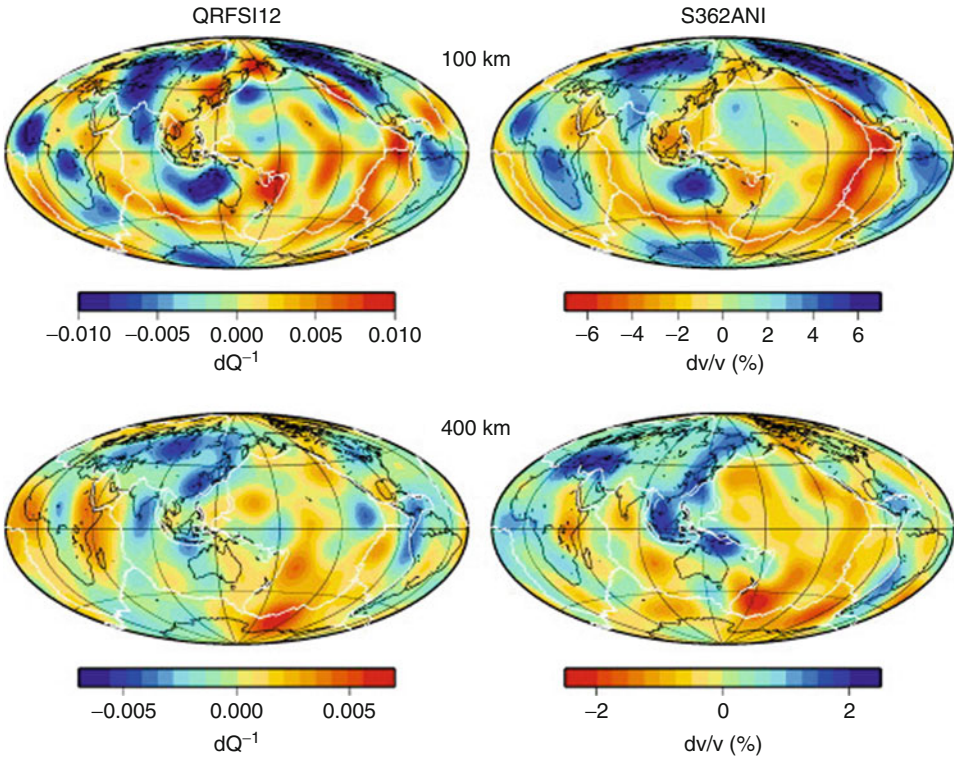
Seismic, Viscoelastic Attenuation, Figure 8 Path-integrated attenuation $t*$ of S waves in the mantle as a function of frequency determined from modeling broadband shear waves predicted from the frequency and depth-dependent attenuation model shown in Figure 7 (Choy and Cormier, 1986).

the Earth's mantle for specific values of E^* and V^*. Relaxation times are also affected by the grain size of minerals, which may increase from millimeter to centimeter in the upper 400 km of the mantle (Faul and Jackson, 2005). A rapid increase in temperature with depth can rapidly change the location of the absorption band with respect to the seismic band. Given a specific temperature profile, estimated values of the activation energy and volume, and grain size, a combined shear velocity and Q_S profile can be predicted and modified to fit an observed shear velocity profile.

Regional variations

Romanowicz and Mitchell (2007) review and interpret both global and many regional variations in intrinsic attenuation, including correlations with velocity perturbations. Tomographic images of perturbations to seismic velocities and attenuations in the mantle can qualitatively be interpreted as images of lateral temperature variations, leaving open the possibility of additional contributions to the observed heterogeneity from chemical variations. In the upper mantle, tectonically active regions overlain by radiogenically younger crust are more attenuating than the mantle underlying inactive regions such as continental shields (Figure 9). The shape of the frequency dependence across the seismic band seems to remain similar in different regions, although the Q at a given frequency is lower for a tectonically young region than for an older shield region. Although motion of interstitial water and partial melt can produce high attenuation, grain boundary and defect mechanisms in a dry mantle have been shown to be equally effective in explaining regions of high attenuation in the upper mantle, except possibly behind island arcs and directly beneath active spreading ridges (Karato and Jung, 1998; Faul and Jackson, 2005).

Generally, perturbations in attenuation Q^{-1} inversely correlate with those in shear wave velocity (Roth et al., 2000). The correlations between shear velocity and shear attenuation appear to be consistent with thermal activation, in which the dispersive effect of attenuation acts jointly with variations in the high-frequency corner of the mantle relaxation spectrum to produce the observed variations in travel time and frequency content. Deep chemical differences between the upper mantle beneath shields and that beneath young continents and oceans as well as in the deep mantle, however, have been suggested by comparing anomalies in shear velocity versus bulk velocity V_K, where $V_K = \sqrt{V_P^2 - (4/3)V_S^2}$. Milder lateral temperature differences in the mid- and lower mantle tend to make the relaxation spectrum more laterally stable in height, width, and location within a frequency band, reducing the observed lateral heterogeneity in velocity and attenuation in these regions.



Seismic, Viscoelastic Attenuation, Figure 9 Inverted upper mantle shear velocity perturbations and shear attenuation. (From Dalton et al., 2009).

Global models of attenuation (e.g., Gung and Romanowicz, 2004) often do not have the resolving power to detect spatially concentrated regions of high attenuation and sharp spatial gradients found in regional studies beneath and near island arcs, mid-ocean spreading ridges, and hot mantle plumes. The dense path coverage required of higher-frequency (0.1–2 Hz) body waves to resolve smaller spatial scales usually is lacking, except in regions containing dense seismic networks. Even larger scale, long-established lateral variations, such as high attenuation west of the Rocky Mountains in North America and low-attenuation east (Der et al., 1982), are not very apparent in some global studies (Warren and Shearer, 2002).

Strain dependence

Laboratory measurements of Q in rocks find dependence in strain beginning at strains of about 10^{-6}. The strain dependence decreases with increasing confining pressure. The Q is also strongly dependent on moisture and interstitial fluids between cracks in rocks and grain boundaries and soils. These observations are consistent with a physical mechanism of frictional sliding across cracks. Unlike viscoelastic relaxations, which are representative of all linear mechanisms, frictional sliding is an inherently non-linear mechanism, depending on strain amplitude.

Estimates, when non-linear effects occur, may be made by calculating the strain associated with the seismic wave being analyzed. A rough estimate can be obtained by assuming that the wave front is a plane wave and dividing the particle velocity by the propagation velocity. For example, the particle velocities of body waves observed in strong ground motion recordings from 0 km to 10 km from the hypocenter of a magnitude 6 earthquake are typically 0.01 m/s. If the body wave propagates at 3 km/s, the strain observed at the strong ground motion site is roughly 0.01 m/s divided by 3×10^3 m/s or strain = 3.3×10^{-5}. This value is likely to be in the non-linear regime of surficial rocks having open cracks or pores. In this strain regime, it becomes important to solve the elastic equation of motion with non-linear terms in its rheology (Bonilla et al., 2005), including terms proportional to the square of strain.

Summary

The intrinsic attenuation of seismic waves in the earth has been found to be consistent with loss mechanisms that are thermally activated. The observed regional and frequency dependences of seismic Q agree with the expected lateral variations in a geotherm having a rapid temperature increase in the upper 400 km of the mantle, followed by a slower vertical and lateral variation in the mid- and lower mantle. High velocities correlate with regions of low attenuation; low seismic velocities correlate with regions of high attenuation. Measurements are consistent with losses primarily in shear rather than bulk deformations.

The existence of lateral heterogeneity in the elastic properties of the Earth complicates the measurement of viscoelastic properties. The longer scale lengths of heterogeneity can split modes of free oscillation and focus and defocus body waves and surface waves. Shorter scale lengths scatter seismic energy, broaden the waveforms of body waves, and redistribute energy into different time and angular windows. Observations that are useful for discriminating between the effects of scattering attenuation versus viscoelastic attenuation include the ratio of apparent P wave attenuation to apparent S wave attenuation, the rate of velocity dispersion within a frequency band, and the apparent viscoelastic modulus defect. The intensity of heterogeneity in percent fluctuation of velocities and densities is higher at shorter scale lengths at shallower depths in the Earth's crust and upper mantle. There is still a need for experiments that determine finer details of how the distribution of heterogeneity changes with depth and lateral location in the Earth and its anisotropy of scale lengths. Many, if not most experiments, have not completely removed the effects of heterogeneity on the apparent attenuation, making their results an upper bound on the viscoelastic Q^{-1}.

Laboratory experiments find a transition from linear to non-linear rheology at strains of the order of 10^{-6}. The observed strain dependence of Q and its dependence on pressure in the shallow crust agree with a mechanism of frictional sliding of cracks. It is still unknown how and at what strain levels linear superposition begins to break down close to a seismic source.

Although a consensus has been reached on the major features and thermal activation of intrinsic attenuation in most of the Earth's upper mantle, this is less true of other deep regions of the Earth. Definitive experiments are still needed for the determination of Q^{-1} in the lowermost 400 km of the mantle, where increased lateral heterogeneity exists across a broad spatial spectrum, complicating the separation of its effects from those of viscoelasticity. A concept unifying lateral variations in velocity, elastic anisotropy, scattering, and apparent attenuation in the uppermost inner core is needed (e.g., Calvet and Margerin, 2008).

Bibliography

Anderson, D. L., 1989. *Theory of the Earth*. Boston: Blackwell Scientific Publications.

Anderson, D. L., and Given, J. W., 1982. The absorption band Q model for the Earth. *Journal of Geophysical Research*, **87**, 3893–3904.

Blanch, J. O., Robertsson, J. O. A., and Symes, W. W., 1995. Optimally efficient constant Q modeling. *Geophysics*, **60**, 176–184.

Boatwright, J., and Choy, G., 1986. Teleseismic estimates of the energy radiated by shallow earthquakes. *Journal of Geophysical Research*, **91**, 2095–2112.

Bonilla, L. F., Archuleta, R. J., and Lavallée, D., 2005. Hysteretic and dilatant behavior of cohesionless soils and their effects on nonlinear site response: field data, observations and modeling. *Bulletin. Seismological Society of America*, **95**, 2373–2395.

Calvet, M., and Margerin, L., 2008. Constraints on grain size and stable iron phases in the uppermost inner core from multiple scattering modeling of seismic velocity and attenuation. *Earth and Planetary Science Letters*, **267**, 200–212.

Carpenter, E. W., 1967. Teleseismic signal calculated for underground, underwater, and atmospheric explosions. *Geophysics*, **32**, 17–32.
Choy, G. L., and Boatwright, J. L., 1995. Global patterns of radiated seismic energy and apparent stress. *Journal of Geophysical Research*, **100**, 18205–18228.
Choy, G. L., and Cormier, V. F., 1986. Direct measurement of the mantle attenuation operator from broadband P and S waves. *Journal of Geophysical Research*, **91**, 7326–7342.
Cormier, V. F., and Li, X., 2002. Frequency dependent attenuation in the inner core: Part II. A scattering and fabric interpretation. *Journal of Geophysical Research,* **107**(B12), doi:10.1029/2002JB1796.
Cormier, V. F., and Richards, P. G., 1976. Comments on "The Damping of Core Waves" by Anthony Qamar and Alfredo Eisenberg. *Journal of Geophysical Research*, **81**, 3066–3068.
Dalton, C. A., Ekstrom, G., and Dziewonski, A. M., 2009. Global seismological shear velocity and attenuation: A comparison with experimental observations. *Earth and Planetary Science Letters*, **284**, 65–75.
Der, Z. A., McElfresh, T. W., and O'Dannell, 1982. An investigation of regional variations and frequency dependence of anelastic attenuation in the United States in the 0.5–4 Hz. band. *Geophysical Journal of the Royal Astronomical Society*, **69**, 67–100.
Dziewonski, A. M., and Anderson, D. L., 1981. Preliminary reference Earth model. *Physics of the Earth and Planetary Interiors*, **24**, 297–356.
Faul, U. H., and Jackson, I., 2005. The seismological signature of temperature and grain size variations in the upper mantle. *Earth and Planetary Science Letters*, **234**, 119–134.
Futterman, W. I., 1962. Dispersive body waves. *Journal of Geophysical Research*, **67**, 5279–5291.
Gross, B., 1953. *Mathematical Structure of the Theories of Viscoelasticity*. Paris: Hermann.
Gung, Y., and Romanowiz, B. A., 2004. Q tomography of the upper mantle using three component long period waveforms. *Geophysical Journal International*, **147**, 831–830.
Hong, T.-K., and Wu, R.-S., 2005. Scattering of elastic waves in geometrically anisotropic random media and its implication to sounding of heterogeneity in the Earth's deep interior. *Geophysical Journal International*, **163**, 324–338.
Jackson, I., 1993. Progress in the experimental study of seismic attenuation. *Annual Review of Earth and Planetary Sciences*, **21**, 375–406.
Jackson, I., 2007. Properties of rocks and minerals – physical origin of anelasticity and attenuation in rocks. In: Schubert, G. (ed.), *Treatise on Geophysics*, 2, Amsterdam: Elsevier, pp. 493–525.
Jackson, D. D., and Anderson, D. L., 1970. Physical mechanisms of seismic wave attenuation. *Reviews of Geophysics and Space Physics*, **8**, 1–63.
Jackson, I., Webb, S., Weston, L., and Boness, D., 2005. Frequency dependence of elastic wave speeds at high temperature: a direct experimental demonstration. *Physics of the Earth and Planetary Interiors*, **148**, 85–96.
Kaelin, B., and Johnson, L. R., 1998. Dynamic composite elastic medium theory. Part II. Three-dimensional media. *Journal of Applied Physics*, **84**, 5458–5468.
Karato, S.-I., and Jung, H., 1998. Water partial melting and the origin of the seismic low velocity zone in the upper mantle. *Earth and Planetary Science Letters*, **157**, 193–207.
Knopoff, L., 1964. Q. *Review of Geophysics*, **2**(4), 625–660.
Kohistedt, D. L., 2007. Properties of rocks and minerals – constitutive equations, rheological behavior, and viscosity of rocks. In: Schubert, G. (ed.), *Treatise on Geophysics*, 2, Amsterdam: Elsevier, 390–417.
Li, X., and Cormier, V. F., 2002. Frequency dependent attenuation in the inner core: Part I. A viscoelastic interpretation. *Journal of Geophysical Research*, **107**(B12), doi:10.1029/2002JB001795.
Liu, H-P., Anderson, D. L., and Kanamori, H., 1976. Velocity dispersion due to anelasticity: implications for seismology and mantle composition. *Geophysical Journal of the Royal Astronomical Society*, **47**, 41–58.
Minster, J. B., 1978. Transient and impulse responses of a one-dimensional linearly attenuating medium—I. Analytical results. *Geophysical Journal of the Royal Astronomical Society*, **52**, 479–501.
Minster, B., and Anderson, D. L., 1981. A model of dislocation-controlled rheology for the mantle. *Philosophical Transactions of the Royal Society of London*, **299**, 319–356.
Nowick, A. S., and Berry, B. S., 1972. *Anelastic Relaxation in Crystalline Solids*. New York: Academic, p. 677.
O'Connell, R. J., and Budiansky, B., 1977. Viscoelastic properties of fluid-saturated cracked solids. *Journal of Geophysical Research*, **82**, 5719–5735.
O'Doherty, R. F., and Anstey, N. A., 1971. Reflections on amplitudes. *Geophysical Prospecting*, **19**, 430–458.
Panning, M. P., and Romanowicz, B. A., 2006. A three dimensional radially anisotropic model of shear velocity in the whole mantle. *Geophysical Journal International*, **167**, 361–379.
Ricard, Y., and Chambat, F., 2009. Seismic attenuation in a phase change coexistence loop. *Physics of the Earth and Planetary Interiors*, **176**, 124–131.
Richards, P. G., and Menke, W., 1983. The apparent attenuation of a scattering medium. *Bulletin. Seismological Society of America*, **73**, 1005–1021.
Robertsson, J. O. A., Blanch, J. O., and Symes, W. W., 1994. Viscoelastic finite-difference modeling. *Geophysics*, **59**, 1444–1456.
Romanowicz, B., and Mitchell, B., 2007. Deep earth structure: Q of the earth from crust to core. In: Schubert, G. (ed.), *Treatise on Geophysics*, 1, Amsterdam: Elsevier, 731–774.
Roth, E. G., Wiens, D. A., and Zhao, D., 2000. An empirical relationship betweeen seismic attenuation and velocity anomalies in the upper mantle. *Geophysical Research Letters*, **27**, 601–604.
Silver, P. G., 1996. Seismic anisotropy beneath the continents: Probing the depths of geology. *Annual Review of Earth and Planetary Sciences*, **24**, 385–432.
Stevenson, D. J., 1983. Anomalous bulk viscosity of two-phase fluids and implications for planetary interiors. *Journal of Geophysical Research*, **88**, 2445–2455.
Warren, L. M., and Shearer, P. M., 2000. Investigating the frequency dependence of mantle Q by stacking P and PP spectra. *Journal of Geophysical Research*, **105**(B11), 25391–25402.
Warren, L. M., and Shearer, P. M., 2002. Mapping lateral variations in upper mantle attenuation by stacking P and PP spectra. *Journal of Geophysical Research*, **107**(B12), 2342.
Zener, C., 1960. *Elasticity and Anelasticity of Metals*. Chicago: The University of Chicago Press.

Cross-references

Body Waves
Earth's Structure, Core
Earth's Structure, Lower Mantle
Earth's Structure, Upper Mantle
Earthquakes, Early and Strong Motion Warning
Earthquakes, Strong-Ground Motion
Energy Partitioning of Seismic Waves
Free Oscillations of the Earth
Lithosphere, Mechanical Properties
Mantle D″ Layer
Mantle Viscosity
Numerical Methods, Finite Difference

Propagation of Elastic Waves: Fundamentals
Seismic Anisotropy
Seismic Diffraction
Seismic Monitoring of Nuclear Explosions
Seismic Phase Names: IASPEI Standard
Seismic Properties of Rocks
Seismic Tomography
Seismic Velocity-Temperature Relationships
Seismic Wave Propagation in Real Media: Numerical Modeling Approaches
Seismic Waves, Scattering
Seismic, Ray Theory
Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting
Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat

SEISMIC, WAVEFORM MODELING AND TOMOGRAPHY

Yanghua Wang
Department of Earth Science and Engineering, Centre for Reservoir Geophysics, Imperial College London, London, UK

Synonyms

Waveform inversion

Definition

Tomography. An inversion method to produce slicing image of the internal structures of an object, by recording wavefield propagating through and scattered/refracted/reflected back from the object, and observing the difference in the effects on the wave energy impinging on those structures.

Introduction

Waveform tomography is an imaging technology using seismic data to reconstruct the Earth subsurface physical properties, represented by seismic velocity, attenuation coefficient, etc. It is usually formulated as an inverse problem. Many seismic tomography methods use the arrival time information extracted from the waveform recordings as the input data to an inversion process. These methods are generically referred to as travel time tomography. Comparing to travel time tomography, waveform tomography has at least two major features. First, the input data consist of the seismic waveforms themselves, as opposed to travel times, amplitudes, or some other secondary attributes of the recorded data. This makes waveform tomography have a better resolution than travel time tomography. Secondly, the underlying numerical method is based on the full wave equation, as opposed to a ray approximation or a Born approximation. This makes waveform tomography more accurate than the travel time counterpart. These two features also make the inverse problem of waveform tomography more difficult to solve.

However, as field seismic data usually lack low-frequency information, travel time tomography may provide a reliable starting model for the iterative waveform inversion (Pratt et al., 2002). If waveform tomography is implemented in the frequency domain, the inversion process proceeds from low to high-frequency components. In the time domain, a band-pass filter can be applied to the seismic data, and inversion uses band-pass filtered data with low frequencies first and proceeds to higher frequencies. Even for the frequency-domain implementation, in order to suppress the low signal-to-noise ratio effect of frequency data and to make the inversion procedure robust, waveform tomography usually uses a group of frequencies simultaneously in an iterative inversion, and proceeds from low to high-frequency groups, generating a high-resolution image of the subsurface model.

Waveform tomography has been used successfully on transmission data, such as crosshole seismic data, and to wide-angle reflection/refraction seismic data with a certain degree of success. It has also been used for regional-scale studies on the crustal and upper mantle velocity structure, using scattering waves, surface waves and SH-waves from either exploration seismics or broad-band teleseismograms. In addition, there is a good progress on its application to reflection-seismic data with limited source-receiver offsets (Wang and Rao, 2009). The seismic reflection method is a routine practice in the hydrocarbon exploration, and the data dominated by the pre-critical reflection energy, reflected back from subsurface contrasts in physical parameters, are well suited for seismic migration for the structural image. Applying the waveform tomography technique to these reflection data can quantitatively extract the geophysical parameters, for identifying different lithologies and different fracture characteristics and even for indicating the hydrocarbon distribution directly.

Waveform modeling

Seismic wave propagation satisfies both Newton's second law and Hook's law, and can be expressed by the following five simultaneous first-order differential equations in two-dimensional case:

$$
\begin{aligned}
\rho \frac{\partial u}{\partial t} &= \frac{\partial \tau_{xx}}{\partial x} + \frac{\partial \tau_{xz}}{\partial z}, \\
\rho \frac{\partial w}{\partial t} &= \frac{\partial \tau_{xz}}{\partial x} + \frac{\partial \tau_{zz}}{\partial z}, \\
\frac{\partial \tau_{xx}}{\partial t} &= c_{11} \frac{\partial u}{\partial x} + c_{13} \frac{\partial w}{\partial z} + c_{15} \left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x} \right), \\
\frac{\partial \tau_{zz}}{\partial t} &= c_{31} \frac{\partial u}{\partial x} + c_{33} \frac{\partial w}{\partial z} + c_{35} \left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x} \right), \\
\frac{\partial \tau_{xz}}{\partial t} &= c_{51} \frac{\partial u}{\partial x} + c_{53} \frac{\partial w}{\partial z} + c_{55} \left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x} \right),
\end{aligned}
\tag{1}
$$

where u and w are the particle-velocity components in the horizontal and vertical directions, respectively, τ_{xx}, τ_{xz} and τ_{zz} are stress components, ρ is density, and

$$\begin{bmatrix} c_{11} & c_{13} & c_{15} \\ c_{31} & c_{33} & c_{35} \\ c_{51} & c_{53} & c_{55} \end{bmatrix} \tag{2}$$

are the elastic constants in Hook's law relating stress to strain in the two-dimensional case. The first two equations are derived from Newton's second law, and the last three from Hooke's law for an elastic medium. Eliminating stress components, these five simultaneous equations can be coupled into two equations,

$$\begin{aligned} &-\rho\frac{\partial^2 u}{\partial t^2}+\frac{\partial}{\partial x}\left(c_{11}\frac{\partial u}{\partial x}+c_{15}\frac{\partial w}{\partial x}\right)+\frac{\partial}{\partial x}\left(c_{15}\frac{\partial u}{\partial z}+c_{13}\frac{\partial w}{\partial z}\right)\\ &+\frac{\partial}{\partial z}\left(c_{51}\frac{\partial u}{\partial x}+c_{55}\frac{\partial w}{\partial x}\right)+\frac{\partial}{\partial z}\left(c_{55}\frac{\partial u}{\partial z}+c_{53}\frac{\partial w}{\partial z}\right)=0,\\ &-\rho\frac{\partial^2 w}{\partial t^2}+\frac{\partial}{\partial x}\left(c_{51}\frac{\partial u}{\partial x}+c_{55}\frac{\partial w}{\partial x}\right)+\frac{\partial}{\partial x}\left(c_{55}\frac{\partial u}{\partial z}+c_{53}\frac{\partial w}{\partial z}\right)\\ &+\frac{\partial}{\partial z}\left(c_{31}\frac{\partial u}{\partial x}+c_{35}\frac{\partial w}{\partial x}\right)+\frac{\partial}{\partial z}\left(c_{35}\frac{\partial u}{\partial z}+c_{33}\frac{\partial w}{\partial z}\right)=0, \end{aligned} \tag{3}$$

defined in terms of horizontal and vertical wave components u and w. These equations are valid for the Earth media with arbitrary anisotropy and heterogeneity.

In the forward modeling of seismic wave propagation based upon either five or two simultaneous equations, there has been steady development in the finite-difference implementations. As the time derivatives are approximated by $\partial u/\partial t \approx (u_{i+1}-u_i)/\Delta t$ and $\partial^2 u/\partial t^2 \approx (u_{i+1}-2u_i+u_{i-1})/\Delta t^2$, where i is the time index, $t_i = i\Delta t$, and Δt is the step size in time, the wavefield at the current time can be calculated based on the previous wavefield: $u_{i+1} = f(u_i)$ or $u_{i+1} = f(u_i, u_{i-1})$; it is therefore an explicit finite-difference scheme (Alford et al., 1974; Kelly et al., 1976). For the spatial derivatives, a second-order staggered grid scheme for numerical stability was proposed by Virieux (1986), and was extended to a fourth-order scheme for greater accuracy by Levander (1988).

In the frequency domain, the two simultaneous equations become

$$\begin{aligned} &\omega^2\rho u+\frac{\partial}{\partial x}\left(c_{11}\frac{\partial u}{\partial x}+c_{15}\frac{\partial w}{\partial x}\right)+\frac{\partial}{\partial x}\left(c_{15}\frac{\partial u}{\partial z}+c_{13}\frac{\partial w}{\partial z}\right)\\ &+\frac{\partial}{\partial z}\left(c_{51}\frac{\partial u}{\partial x}+c_{55}\frac{\partial w}{\partial x}\right)+\frac{\partial}{\partial z}\left(c_{55}\frac{\partial u}{\partial z}+c_{53}\frac{\partial w}{\partial z}\right)=0,\\ &\omega^2\rho w+\frac{\partial}{\partial x}\left(c_{51}\frac{\partial u}{\partial x}+c_{55}\frac{\partial w}{\partial x}\right)+\frac{\partial}{\partial x}\left(c_{55}\frac{\partial u}{\partial z}+c_{53}\frac{\partial w}{\partial z}\right)\\ &+\frac{\partial}{\partial z}\left(c_{31}\frac{\partial u}{\partial x}+c_{35}\frac{\partial w}{\partial x}\right)+\frac{\partial}{\partial z}\left(c_{35}\frac{\partial u}{\partial z}+c_{33}\frac{\partial w}{\partial z}\right)=0. \end{aligned} \tag{4}$$

where ω is the angular frequency. This system of wave equations can be solved in parallel for individual frequencies, using also a finite-difference method. Once all frequency components of the wavefield at any spatial position are obtained, performing inverse Fourier transform with respect to the frequency variable will produce the time-domain seismic trace at this location.

However, the frequency-domain implementation is much more time consuming, compared to its counterpart in the time domain. This is because the frequency-domain calculation involves the solution of linear algebraic equations as the follows. With a finite-differencing scheme, Equation 4 may be presented in a matrix form as

$$\mathbf{Mu} = \mathbf{s}, \tag{5}$$

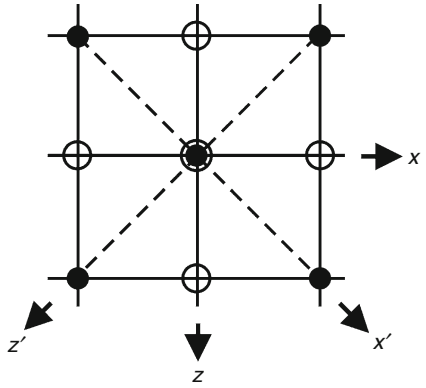
where $\mathbf{M}$ is a matrix approximating the partial differential operators, $\mathbf{u}$ is a vector representing the two components of the wavefield at all grids, and $\mathbf{s}$ is a vector representing the source term, which is zero everywhere except at the location of the source (added to the right-hand side of the wave equations). To obtain wavefield $\mathbf{u}$, one needs to solve the linear algebraic equation system 5, that is, to solve the inverse matrix, $\mathbf{M}^{-1}$. It is therefore an implicit finite-difference scheme.

The matrix $\mathbf{M}$ represents a significant storage requirement, which is largely determined by the numerical bandwidth of $\mathbf{M}$ and by the manner in which the structural sparsity of the matrix is maintained in any solution method. The bandwidth of the differencing matrix is determined by the number of nodes needed for a spatial derivative. For example, for a second-order finite-differencing, a first derivative needs 3×3 nodes, and for a fourth-order 5×5 nodes. A minimal, rotated computational star is a scheme that can minimize the bandwidth of the matrix and meanwhile improve the accuracy for finite-difference modeling. It does not require any new grid points as used in a second-order finite-differencing but can produce an equivalent of fourth-order accuracy in finite-differencing (Štekl and Pratt, 1998). As shown in Figure 1, the ○ symbol represents the five nodes needed in second-order finite-differencing of a first derivative, and the • symbol represents the five nodes in a 45° rotated finite-differencing star. Coupling these two second-order finite-differencing stars does not use additional points outside the 3×3 nodes, and there will be no increase in the numerical bandwidth of the differencing matrix. Therefore, the increment in computational cost and storage requirement over the ordinary second-order scheme is negligible.

In a rotated coordinate system, the elastic constants (2) become

$$\begin{bmatrix} c_{11}' & c_{13}' & c_{15}' \\ c_{31}' & c_{33}' & c_{35}' \\ c_{51}' & c_{53}' & c_{55}' \end{bmatrix} = \mathbf{B}\begin{bmatrix} c_{11} & c_{13} & c_{15} \\ c_{31} & c_{33} & c_{35} \\ c_{51} & c_{53} & c_{55} \end{bmatrix}\mathbf{B}^T, \tag{6}$$

where the rotation operator is



Seismic, Waveform Modeling and Tomography, Figure 1 The minimal, rotated computational star for finite-difference modeling. The symbol ○ represents the only five nodes required in the ordinary, second-order finite-differencing star. The symbol • represents the five nodes in a rotated computational star. The combination of these two stars indicates the coupling of the central node to the nearest neighbors on the grid. The final star does not use additional points outside the 3×3 nodes, but coupling includes those nodes not presented in the original star.

$$\mathbf{B} = \begin{bmatrix} \cos^2\theta & \sin^2\theta & 2\sin\theta\cos\theta \\ \sin^2\theta & \cos^2\theta & 2\sin\theta\cos\theta \\ -\sin\theta\cos\theta & \sin\theta\cos\theta & \cos^2\theta - \sin^2\theta \end{bmatrix}, \tag{7}$$

and θ is rotated angle of new coordinate. Considering a special case of 45° rotation, where the rotated coordinate system is (x', z') and displacements are (u', w') in the rotated coordinate system, Equation 4 becomes

$$\begin{aligned} &\omega^2\rho u' + \frac{\partial}{\partial x'}\left(c_{11}'\frac{\partial u'}{\partial x'} + c_{15}'\frac{\partial w'}{\partial x'}\right) + \frac{\partial}{\partial x'}\left(c_{15}'\frac{\partial u'}{\partial z'} + c_{13}'\frac{\partial w'}{\partial z'}\right) \\ &+ \frac{\partial}{\partial z'}\left(c_{51}'\frac{\partial u'}{\partial x'} + c_{55}'\frac{\partial w'}{\partial x'}\right) + \frac{\partial}{\partial z'}\left(c_{55}'\frac{\partial u'}{\partial z'} + c_{53}'\frac{\partial w'}{\partial z'}\right) = 0, \\ &\omega^2\rho w' + \frac{\partial}{\partial x'}\left(c_{51}'\frac{\partial u'}{\partial x'} + c_{55}'\frac{\partial w'}{\partial x'}\right) + \frac{\partial}{\partial x'}\left(c_{55}'\frac{\partial u'}{\partial z'} + c_{53}'\frac{\partial w'}{\partial z'}\right) \\ &+ \frac{\partial}{\partial z'}\left(c_{31}'\frac{\partial u'}{\partial x'} + c_{35}'\frac{\partial w'}{\partial x'}\right) + \frac{\partial}{\partial z'}\left(c_{35}'\frac{\partial u'}{\partial z'} + c_{33}'\frac{\partial w'}{\partial z'}\right) = 0. \end{aligned} \tag{8}$$

The relationship between displacements (u', w') in the rotated coordinate system and (u, w) in the original coordinate system is given by

$$u = \frac{1}{\sqrt{2}}(u' - w'),\ w = \frac{1}{\sqrt{2}}(u' + w'). \tag{9}$$

By subtracting and adding two equations in Equation 8 and then using relations in Equation 9, one can obtain

$$\begin{aligned} &\omega^2\rho u + \frac{1}{2}\frac{\partial}{\partial x'}\left((c_{11}' - c_{51}' - c_{15}' + c_{55}')\frac{\partial u}{\partial x'}\right. \\ &\quad\left. + (c_{11}' - c_{51}' + c_{15}' - c_{55}')\frac{\partial w}{\partial x'}\right) \\ &\quad + \frac{1}{2}\frac{\partial}{\partial x'}\left((c_{15}' - c_{55}' - c_{13}' + c_{53}')\frac{\partial u}{\partial z'}\right. \\ &\quad\left. + (c_{15}' - c_{55}' + c_{13}' - c_{53}')\frac{\partial w}{\partial z'}\right) \\ &\quad + \frac{1}{2}\frac{\partial}{\partial z'}\left((c_{51}' - c_{31}' - c_{55}' + c_{35}')\frac{\partial u}{\partial x'}\right. \\ &\quad\left. + (c_{51}' - c_{31}' + c_{55}' - c_{35}')\frac{\partial w}{\partial x'}\right) \\ &\quad + \frac{1}{2}\frac{\partial}{\partial z'}\left((c_{55}' - c_{35}' - c_{53}' + c_{33}')\frac{\partial u}{\partial z'}\right. \\ &\quad\left. + (c_{55}' - c_{35}' + c_{53}' - c_{33}')\frac{\partial w}{\partial z'}\right) = 0, \\ &\omega^2\rho w + \frac{1}{2}\frac{\partial}{\partial x'}\left((c_{11}' + c_{51}' - c_{15}' - c_{55}')\frac{\partial u}{\partial x'}\right. \\ &\quad\left. + (c_{11}' + c_{51}' + c_{15}' + c_{55}')\frac{\partial w}{\partial x'}\right) \\ &\quad + \frac{1}{2}\frac{\partial}{\partial x'}\left((c_{15}' + c_{55}' - c_{13}' - c_{53}')\frac{\partial u}{\partial z'}\right. \\ &\quad\left. + (c_{15}' + c_{55}' + c_{13}' + c_{53}')\frac{\partial w}{\partial z'}\right) \\ &\quad + \frac{1}{2}\frac{\partial}{\partial z'}\left((c_{51}' + c_{31}' - c_{55}' - c_{35}')\frac{\partial u}{\partial x'}\right. \\ &\quad\left. + (c_{51}' + c_{31}' + c_{55}' + c_{35}')\frac{\partial w}{\partial x'}\right) \\ &\quad + \frac{1}{2}\frac{\partial}{\partial z'}\left((c_{55}' + c_{35}' - c_{53}' - c_{33}')\frac{\partial u}{\partial z'}\right. \\ &\quad\left. + (c_{55}' + c_{35}' + c_{53}' + c_{33}')\frac{\partial w}{\partial z'}\right) = 0. \end{aligned} \tag{10}$$

Now there are two systems of partial differential equations:

$$\begin{aligned} &\rho\omega^2 u + A_1 = 0, \\ &\rho\omega^2 w + B_1 = 0, \end{aligned} \quad \text{and} \quad \begin{aligned} &\rho\omega^2 u + A_2 = 0, \\ &\rho\omega^2 w + B_2 = 0, \end{aligned} \tag{11}$$

where A_1 and B_1 are the partial differential parts of Equation 4 in the original coordinate system, and A_2 and B_2 are the partial differential parts of equation system 10 in the rotated coordinate system. The two systems can be combined to as

$$\begin{aligned} &\rho\omega^2 u + aA_1 + (1-a)A_2 = 0, \\ &\rho\omega^2 w + aB_1 + (1-a)B_2 = 0, \end{aligned} \tag{12}$$

where a is an optimal real-valued coefficient that must be sought to maximize the accuracy of the solution for all propagation directions.

While it is feasible to calculate the inverse matrix $\mathbf{M}^{-1}$ for a small, 2D model, in many practical problems such as 3D seismic modeling, the system $\mathbf{Mu} = \mathbf{s}$ with a large, sparse matrix is solved iteratively rather than directly. An approximate solution $\tilde{\mathbf{u}}$ is first obtained by solving using a related, preconditioning matrix. This preconditioner is chosen such that it is related to the true matrix, but is much faster to invert. The approximate solution $\tilde{\mathbf{u}}$ from the preconditioner is then substituted into the true matrix equation to discover the effective source, $\tilde{\mathbf{s}} = \mathbf{M}\tilde{\mathbf{u}}$. The difference $\mathbf{s} - \tilde{\mathbf{s}}$ between this effective source and the true source can then be treated as an actual source for the next iteration. The final solution $\mathbf{u}$ to the original matrix equation is found if the difference is sufficiently small (Warner et al., 2008).

Figure 2 shows the snapshot of wave propagation through an anisotropic medium. The elastic coefficients are summarized as

$$c_{11} = 29.26, \quad c_{33} = 23.98, \quad c_{55} = 4.29,$$
$$c_{13} = c_{31} = 17.27, \quad c_{15} = c_{51} = 1.98,$$
$$c_{35} = c_{53} = 2.53(\text{Gpa}),$$

and the density of the medium is a constant of 2.5 g/cm^3. The source is a Ricker wavelet with dominate frequency of 10 Hz. The snapshot shows clearly qP-wave front propagating in difference directions with different velocities. S-wave front can also been observed in these figures.

Waveform modeling uses an elastic wave equation, for the generality, and produces a wavefield presented as particle-velocity components. In a homogeneous, isotropic media, $c_{11} = c_{33} = \lambda + 2\mu$, $c_{13} = c_{31} = \lambda$ and $c_{55} = \mu$, where λ and μ are Lamé parameters, and $c_{15} = c_{51} = c_{35} = c_{53} = 0$. In an acoustic assumption, $c_{11} = c_{33} = c_{13} = \lambda$ and $c_{55} = c_{15} = c_{35} = 0$. With the acoustic assumption, most field seismic records are pressure data. According to Newton's second law, the relationship between particle velocity (u, w) and pressure P is

$$\rho \frac{\partial u}{\partial t} = -\frac{\partial P}{\partial x}, \rho \frac{\partial w}{\partial t} = -\frac{\partial P}{\partial z}. \tag{13}$$

Inverse method

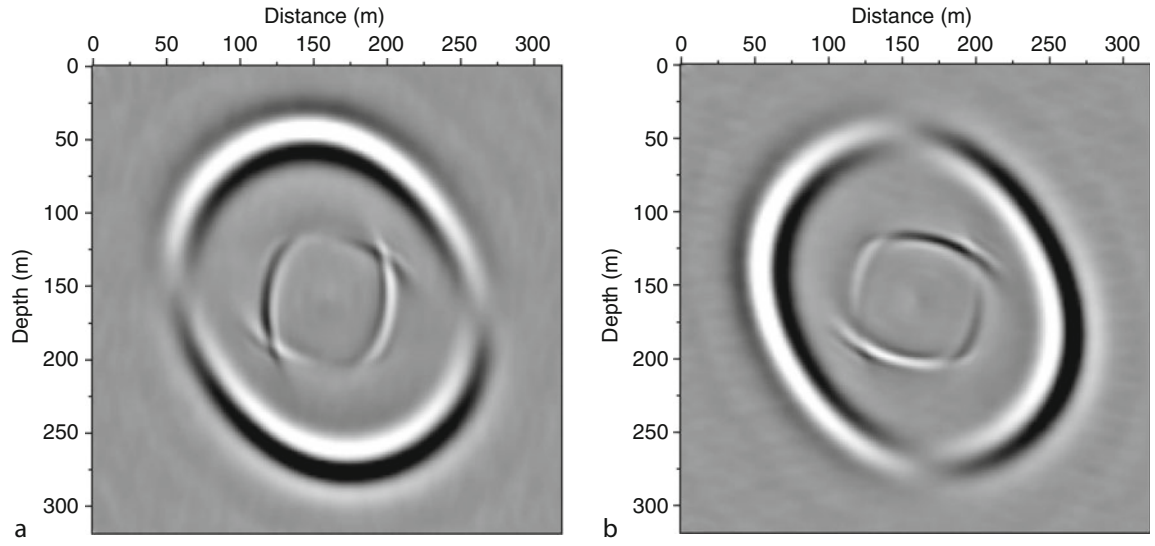
Seismic waveform tomography is an inverse problem with an objective function defined by

$$\varphi(\mathbf{m}) = [\mathbf{P}(\mathbf{m}) - \mathbf{P}_{\text{obs}}]^{\text{H}} \mathbf{C}_{\text{D}}^{-1} [\mathbf{P}(\mathbf{m}) - \mathbf{P}_{\text{obs}}] + \mu[\mathbf{m} - \mathbf{m}_0]^{\text{H}} \mathbf{C}_{\text{M}}^{-1} [\mathbf{m} - \mathbf{m}_0], \tag{14}$$

where $\mathbf{m}_0$ is a reference model, $\mathbf{m}$ is the model to invert for, $\mathbf{P}(\mathbf{m})$ is a modeled data set based on model $\mathbf{m}$, $\mathbf{P}_{\text{obs}}$ is an observed data set, $\mathbf{C}_{\text{D}}$ is the data covariance matrix with units of $(data)^2$, defining the uncertainties in the data set, $\mathbf{C}_{\text{M}}$ is the model covariance matrix with units of $(model\ parameter)^2$, and μ is a scalar that controls the relative weights of the data fitting term and the model constraint in the objective function. In Equation 14, the superscript $^{\text{H}}$ denotes the complex conjugate transpose.

For minimizing the objective function, a gradient method can be used. It starts with the differentiation of the objective function with respect to the model parameters:

$$\frac{\partial \varphi}{\partial \mathbf{m}} = 2(\mathbf{F}^{\text{H}} \mathbf{C}_{\text{D}}^{-1} \delta\mathbf{P} + \mu \mathbf{C}_{\text{M}}^{-1} \delta\mathbf{m}), \tag{15}$$



Seismic, Waveform Modeling and Tomography, Figure 2 Snap shot of wave propagation in an anisotropic medium: (**a**) horizontal component and (**b**) vertical component.

where $\delta\mathbf{m} = \mathbf{m} - \mathbf{m}_0$ is the model perturbation, $\delta\mathbf{P} = \mathbf{P}(\mathbf{m}) - \mathbf{P}_{\text{obs}}$ is the data residual, and $\mathbf{F}$ is a matrix of the Fréchet derivative of $\mathbf{P}(\mathbf{m})$ with respect to the model $\mathbf{m}$. The first term in Equation 15 is the gradient direction of the data misfit:

$$\hat{\boldsymbol{\gamma}} = \mathbf{F}^{\mathrm{H}}\mathbf{C}_{\mathrm{D}}^{-1}\delta\mathbf{P} = \mathbf{F}^{\mathrm{H}}\delta\hat{\mathbf{P}}, \tag{16}$$

where $\delta\hat{\mathbf{P}} = \mathbf{C}_{\mathrm{D}}^{-1}\delta\mathbf{P}$ is a weighted data residual. Set $\partial\varphi/\partial\mathbf{m} = 0$ in Equation 15, one obtains the following equation

$$\delta\mathbf{m} = -\alpha\mathbf{C}_{\mathrm{M}}\hat{\boldsymbol{\gamma}}, \tag{17}$$

where α is a update step length that needs to be determined.

In order to evaluate the gradient $\hat{\boldsymbol{\gamma}}$ using Equation 16, one needs to know the Fréchet matrix $\mathbf{F}$, which is obtained from the following linear formula,

$$\delta\mathbf{P} = \mathbf{F}\delta\mathbf{m}. \tag{18}$$

This is the first term in a Taylor's series for $\delta\mathbf{P}$ and relates the data perturbation $\delta\mathbf{P}$ to the model perturbation $\delta\mathbf{m}$. However, a direct computation of $[\mathbf{F}]_{ij} = \partial P_i/\partial m_j$ is a formidable task when P_i are seismic waveforms. The action of matrix $\mathbf{F}^{\mathrm{H}}$ on the weighted data residual vector $\delta\hat{\mathbf{P}}$ (Equation 16) can be computed by a series of forward modeling steps, summarized as follows (Lailly, 1984; Tarantola, 1984, 1987).

The frequency-domain acoustic wave equation for a constant density medium with velocity $c_0(\mathbf{r})$ is

$$\left(\nabla^2 + \frac{\omega^2}{c_0^2(\mathbf{r})}\right)P_0(\mathbf{r},\omega) = -S(\omega)\delta(\mathbf{r}-\mathbf{r}_0), \tag{19}$$

where $\mathbf{r}$ is the position vector, $\mathbf{r}_0$ locates the source position, $S(\omega)$ is the source signature of frequency ω, and $P_0(\mathbf{r},\omega)$ is the (pressure) wavefield of this frequency and we drop off the frequency reference in the following discussion. If the velocity is perturbed by a small amount, $\delta c(\mathbf{r}) << c_0(\mathbf{r})$, that is, $c_0(\mathbf{r}) \to c(\mathbf{r}) = c_0(\mathbf{r}) + \delta c(\mathbf{r})$, then the total wavefield is correspondingly perturbed to $P_0(\mathbf{r}) \to P(\mathbf{r}) = P_0(\mathbf{r}) + \delta P(\mathbf{r})$. Following wave Equation 19, δP approximately satisfies

$$\left(\nabla^2 + \frac{\omega^2}{c_0^2(\mathbf{r})}\right)\delta P(\mathbf{r}) = 2\omega^2 P_0(\mathbf{r})\frac{\delta c(\mathbf{r})}{c_0^3(\mathbf{r})}. \tag{20}$$

Considering $2\omega^2 P_0(\mathbf{r})\delta c(\mathbf{r})/c_0^3(\mathbf{r})$, the term on the right-hand side, as a series of "virtual sources" over $\mathbf{r}$, the integral solution for $\delta P(\mathbf{r})$ can be expressed as

$$\delta P(\mathbf{r}) = -\int_{\Omega} \delta c(\mathbf{r}')\frac{2\omega^2}{c_0^3(\mathbf{r}')}P_0(\mathbf{r}')G(\mathbf{r},\mathbf{r}')\mathrm{d}\mathbf{r}', \tag{21}$$

where $G(\mathbf{r},\mathbf{r}')$ is the Green's function for the response at $\mathbf{r}$ to a point source at $\mathbf{r}'$ for the original velocity field. Note that in the acoustic case where one assumes density to be constant, and defines the model by the velocity field only, $\mathbf{m} \equiv \mathbf{c}$. Then comparing Equation 21 against the matrix-vector form of Equation 18, we see that the Fréchet matrix is defined with element, $F(\mathbf{r},\mathbf{r}') = -[2\omega^2/c_0^3(\mathbf{r}')]P_0(\mathbf{r}')G(\mathbf{r},\mathbf{r}')$. Substituting this Fréchet kernel into Equation 16, one obtains

$$\hat{\boldsymbol{\gamma}}(\mathbf{r}) = \left(\frac{2\omega^2}{c_0^3(\mathbf{r})}\right)^* \int_D P_0^*(\mathbf{r}')G^*(\mathbf{r},\mathbf{r}')\delta\hat{P}(\mathbf{r}'). \tag{22}$$

Replacing the integral over the data space with a summation over source and receiver pairs, denoted by s and g respectively, as the source and receiver position are inherently discrete and finite in number, one can obtain

$$\begin{aligned}\hat{\boldsymbol{\gamma}}(\mathbf{r}) &= \left(\frac{2\omega^2}{c_0^3(\mathbf{r})}\right)^* \sum_s \left(P_0^*(\mathbf{r};\mathbf{r}_s)\sum_g G^*(\mathbf{r},\mathbf{r}_g)\delta\hat{P}(\mathbf{r}_g;\mathbf{r}_s)\right)\\ &= \left(\frac{2\omega^2}{c_0^3(\mathbf{r})}\right)^* \sum_s \left(P_0^*(\mathbf{r};\mathbf{r}_s)P_b^*(\mathbf{r};\mathbf{r}_s)\right),\end{aligned} \tag{23}$$

where

$$P_b(\mathbf{r};\mathbf{r}_s) = \sum_g G(\mathbf{r},\mathbf{r}_g)\delta\hat{P}^*(\mathbf{r}_g;\mathbf{r}_s) \tag{24}$$

representing the wavefield generated by a series of virtual sources $\delta\hat{P}^*(\mathbf{r}_g)$, corresponding to a single source $\mathbf{r}_s$. Note that wavefield $P_b(\mathbf{r};\mathbf{r}_s)$ is not calculated directly from Equation 24, but is computed using the same forward modeling scheme as used for the wave Equation 19 with the virtual sources $\delta\hat{P}^*(\mathbf{r}_g)$, a procedure often referred to as data residual back-propagation.

In summary, waveform tomography is performed iteratively. For each iteration, the inversion procedure may be divided into four steps:

1. For a given model estimate, calculating the synthetic wavefield $P_0(\mathbf{r};\mathbf{r}_s)$ at space position $\mathbf{r}$ corresponding to a source point at $\mathbf{r}_s$
2. Using the weighted data residual $\delta\hat{\mathbf{P}} = \mathbf{C}_{\mathrm{D}}^{-1}\delta\mathbf{P}$ as virtual sources to generate a so-called back-propagation wavefield $P_{\mathrm{b}}(\mathbf{r};\mathbf{r}_s)$
3. Crosscorrelation of the original wavefield $P_0(\mathbf{r};\mathbf{r}_s)$ and the back-propagation wavefield $P_{\mathrm{b}}(\mathbf{r};\mathbf{r}_s)$ to get the gradient direction $\boldsymbol{\gamma} = \mathbf{C}_{\mathrm{M}}\hat{\boldsymbol{\gamma}}$, where $\mathbf{C}_{\mathrm{M}}$ is the model covariance matrix with units of $(\textit{model parameter})^2$
4. Estimating the model update $\delta\mathbf{m} = -\alpha\boldsymbol{\gamma}$, where α is the optimal step length that can be found by using the linear approximation or simply line search for a minimum of the objective function.

In a time-domain implementation of waveform inversion, the gradient direction is calculated by

$$\hat{\boldsymbol{\gamma}}(\mathbf{r}) = \frac{2}{c_0^3(\mathbf{r})}\sum_s \int \left(\frac{\partial}{\partial t}p_0(\mathbf{r};\mathbf{r}_s)\frac{\partial}{\partial t}p_b(\mathbf{r};\mathbf{r}_s)\right)\mathrm{d}t, \tag{25}$$

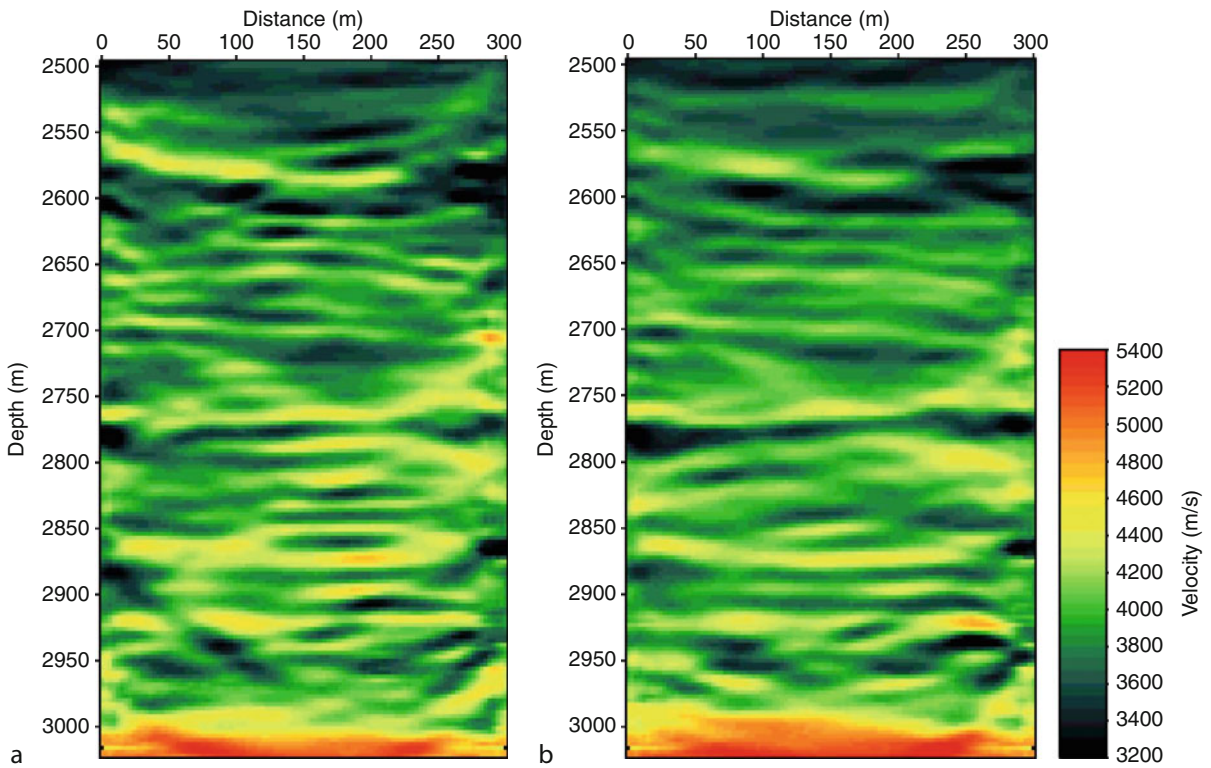
where $\partial p_0(\mathbf{r};\mathbf{r}_s)/\partial t$ denotes the time derivative of the time-domain forward propagated wavefields, and $\partial p_b(\mathbf{r};\mathbf{r}_s)/\partial t$ is the time derivative of the back-projected residual waveforms in the time domain. Zero-leg correlation of these two wavefields provides us the gradient direction for model updating.

Strategies for choosing frequencies

When dealing with real seismic data, poor signal-to-noise ratio of data slices in the frequency domain is a problem at least that affects waveform tomography. For real data application, a group of frequencies is necessarily used simultaneously for each individual iteration in the inversion procedure (Pratt and Shipp, 1999; Wang and Rao, 2006). Simultaneously using neighboring frequencies from the same spatial imaging position may have an averaging effect that suppresses the data noise to the input of the inversion. For a fixed number of model parameters to invert for, using many more data samples in the inversion means that the inverse problem becomes much better determined.

Here is a real data example of crosshole geometry, with two experiments to combat the noise in real data. In the first experiment, we use all selected frequencies consecutively (190, 195, 200, 205, . . . , 485 Hz). We start with the initial model generated from travel time inversion, and invert the 190 Hz data component first. Then, we switch to a higher frequency component (195 Hz) of the data as the inversion progresses. The result from each lower frequency is used as the starting model for the next higher frequency inversion. At each frequency stage, three iterations are carried out. Figure 3a shows the reconstructed image after using all 60 selected frequencies between 190 and 480 Hz with 5 Hz interval.

In the second experiment, we use a group of five neighboring frequencies simultaneously in the inversion. The 60 selected frequencies are assigned into 12 groups with increasing frequency contents. The result from each lower frequency group is used as the starting model for the inversion of the next higher frequency group. This strategy might mitigate the nonlinearity of the problem: For lower frequencies, the method is more tolerant of velocity errors, as these are less likely to lead to errors of more than a half-cycle in the waveforms (Pratt and Shipp, 1999). For each group, three iterations are carried out, proceeding through all groups. For each iteration, the gradient of each frequency group is computed using all five frequencies simultaneously. Figure 3b shows the tomographic image after using all of 12 frequency groups consecutively.



Seismic, Waveform Modeling and Tomography, Figure 3 Waveform tomography with a single frequency at a time or a group of frequencies simultaneously. (**a**) The inversion is executed by each frequency consecutively. The velocity image after using all 60 selected frequencies between 190 and 480 Hz (with 5 Hz interval). The image has strong X-shaped artifacts. (**b**) The inversion is executed one group by one group in sequence. The image after using all 12 frequency groups, each has five frequency components.

Comparing the images of those two experiments, one can observe that Figure 3a is marked by the presence of some X-shaped artifacts that cross the image. Such artifacts are quite often obtained in crosshole tomography, especially when waveform inversion is attempted. It is due to the non-uniform coverage of the object spectrum and the lack of information about the object spectrum in certain directions (Wu and Toksöz, 1987). When using multiple frequencies simultaneously, the inherent filtering (smoothing) effect might have an extrapolation effect of the object spectrum to the blind area. The second experiment has much fewer artifacts, and the image is smoother and more continuous than that of experiment one, especially at the 2,800–2,950 m portions. We recommend using the strategy of the second experiment in practice so that we can also mitigate the data noise effectively in the input of waveform tomography.

The computation time of the frequency-domain waveform tomography is linearly proportional to the total number of temporal frequencies used in the inversion. In reflection geometry, source-receiver pairs vary with different offsets and move along the surface. One may explore this coverage of a variety of plane-wave imaging directions, to reduce the number of frequencies needed in waveform tomography (Sirgue and Pratt, 2004). The frequency selection strategy is

$$f_{n+1} = \frac{f_n}{\cos\phi}, \tag{26}$$

where f_n is the frequency previously used, f_{n+1} is a new frequency for inversion, and $\cos\phi = z/\sqrt{h_{\max}^2 + z^2}$ is the cosine of the reflection angle, corresponding to the maximum value of the half source-receiver offset $h_{\max}$ and the target depth z. This strategy eliminates some frequencies but has a continuous coverage on vertical wave number k_z, as

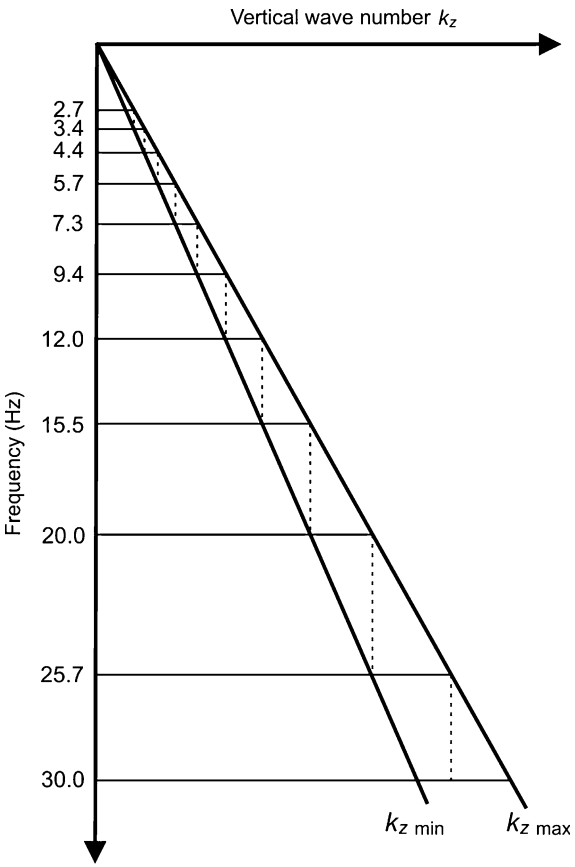
$$k_{z\min}(f_{n+1}) = k_{z\max}(f_n), \tag{27}$$

where

$$k_{z\max} = \frac{2f}{c_0}, \quad k_{z\min} = \frac{2f}{c_0}\cos\phi, \tag{28}$$

and $c_0/2$ are half of the background velocity. Figure 4 illustrates the wave number coverage corresponding to the selected frequencies in a waveform tomography example (Wang and Rao, 2009). For a model with $h_{\max} = 2,275\,\text{m}$, $z = 2,800\,\text{m}$ and $\cos\phi \approx 0.776$, the following 11 frequencies, 2.7, 3.4, 4.4, 5.7, 7.3, 9.4, 12.0, 15.5, 20.0, 25.7, and 30.0 Hz, may be selected. In this list, the dominant frequency 20 Hz is included in the inversion, and the last frequency is $f_{\max} = 30\,\text{Hz}$, due to the limitation of finite-difference modeling, which is less than the predicted value of 33.1 Hz.

On the other hand, if assuming there is no source-receiver offset coverage to be exploited, adequate wave number samples are needed in order to obtain a sufficiently good image in the model space. In this case,



Seismic, Waveform Modeling and Tomography, Figure 4 The coverage of the vertical wave numbers corresponding to 11 selected frequencies for an example waveform tomography.

the sampling rate should satisfy the anti-aliasing condition, $\Delta k_z \le 1/z_{\max}$, where $z_{\max}$ is the maximum depth to be imaged, and Δk_z is the sampling rate of the vertical wave number. Given the minimum value of the vertical wave number at frequency f by Equation 28, the difference between two neighboring wave number samples is

$$\Delta k_z \equiv k_{z\min}(f + \Delta f) - k_{z\min}(f) = 2\cos\phi\frac{\Delta f}{c_0} \le \frac{1}{z_{\max}}. \tag{29}$$

Finally, the following anti-aliasing condition for frequency sampling is obtained:

$$\Delta f \le \frac{c_0}{2z_{\max}\cos\phi}. \tag{30}$$

In this example, $\Delta f = 0.3\,\text{Hz}$.

Waveform tomography proceeds sequentially from low to high frequencies (Sirgue and Pratt, 2004; Pratt, 2008). This is because the nonlinearity of the inverse problem depends on the frequency of the data. As the misfit function at low frequencies is more linear than at high frequencies, the low-frequency inversion will have a better chance to

be successful and can accurately recover the low wave number components of the velocity model. The velocity model with accurate low wave number components is a good initial model for higher frequency inversions. At low frequencies, coarser grids can be used for computing numerical solutions of the wave equation than at high frequencies, resulting in a computational efficiency. This is a so-called multiscale approach (Bunks et al., 1995).

The multiscale approach can also be implemented in the time-domain waveform tomography (Boonyasiriwat et al., 2009). In the frequency domain, it is straightforward to apply the multiscale method, as a single frequency component is used at a time in the inversion. In the time-domain inversion, a frequency band is used instead of single frequency. For each band-passed filtered data set, the finite-difference grid size is determined by the maximum frequency of the band. Assuming a square grid, $\Delta x = \Delta z$, the grid size allowed to use in the forward modeling is determined by

$$\Delta x \le \frac{\lambda_{\min}}{N} = \frac{c_{\min}}{N f_{\max}}, \tag{31}$$

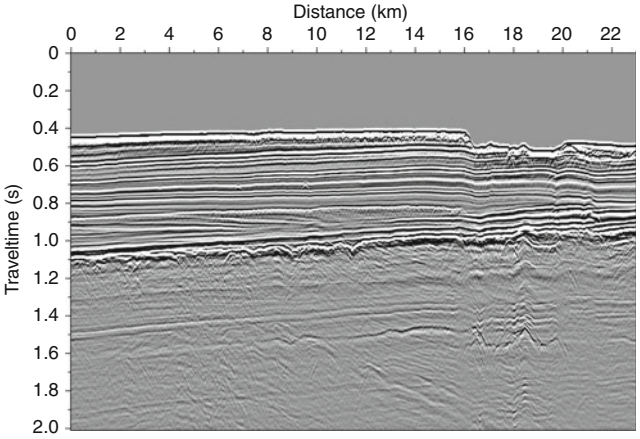
where $\lambda_{\min}$ is the minimum wavelength, $c_{\min}$ is the minimum velocity, $f_{\max}$ is the maximum frequency of the band, and N is the number of grid points at least per minimum wavelength, required by the numerical dispersion condition for the finite-difference scheme (Levander, 1988). At low frequencies, coarser grids can be used than at high frequencies. Therefore, low-frequency inversions will be considerably fast and efficient compared to high-frequency inversions, and can afford to take a large number of iterations in order to obtain an accurate estimate of low wave number components of the velocity model. Higher wave number components are progressively recovered through the sequential uses of higher frequency-band data in the inversions. Therefore, the multiscale approach should have a steady convergence than a single-scale method that tends to recover both low and high wave numbers simultaneously.

Preprocessing for field data waveform tomography

Let us demonstrate the application of waveform tomography to a real, marine seismic data set, as shown in Figure 5. This example profile shows the structure along the section. However, the input to waveform tomography is a series of shot records, which should be subject to some necessary processing, such as multiple attenuation and transformation of the original point sources to equivalent line sources, before they are used in tomography.

Figure 6a is a sample shot record, consisting of 120 traces with a minimum source-receiver offset of 337.5 m and a maximum source-receiver offset of 1,825 m. We investigate the feasibility of reflection-seismic tomography within such a narrow source-receiver offset range.

Waveform tomography usually does not include free-surface multiples. Otherwise including seismic multiples



Seismic, Waveform Modeling and Tomography, Figure 5 An example profile of marine seismic data, which shows the structure along a section. However, the input to waveform tomography is a series of shot records, which should be subject to some necessary processing.

bouncing back and forward within the water layer in the tomographic inversion will increase the nonlinearity of the problem. As the number of multiples increases, the errors in model (and in turn in synthetics) will also increase. We use a narrow-offset shot record in multiple attenuation, also to avoid the wide-angle refraction of the water bottom and their multiples, as the current methodology for free-surface multiple prediction cannot properly model the refraction multiples. As marked in Figure 6a, the most difficult part of multiple attenuation is where the refraction wave just starts appearing. Figure 6b displays a shot record and the one after free-surface multiple attenuation, using a multiple prediction through inversion (MPI) method (Wang, 2004, 2007). The real shot record is generated by a point source, but Figure 6c is an equivalent line-source shot gather, after partial compensation as follows.

Before input to waveform tomography, a shot record of real seismic data needs to be partially compensated, to become a gather generated from a line source. For a two-dimensional (2-D) case, the 2-D Green's function is

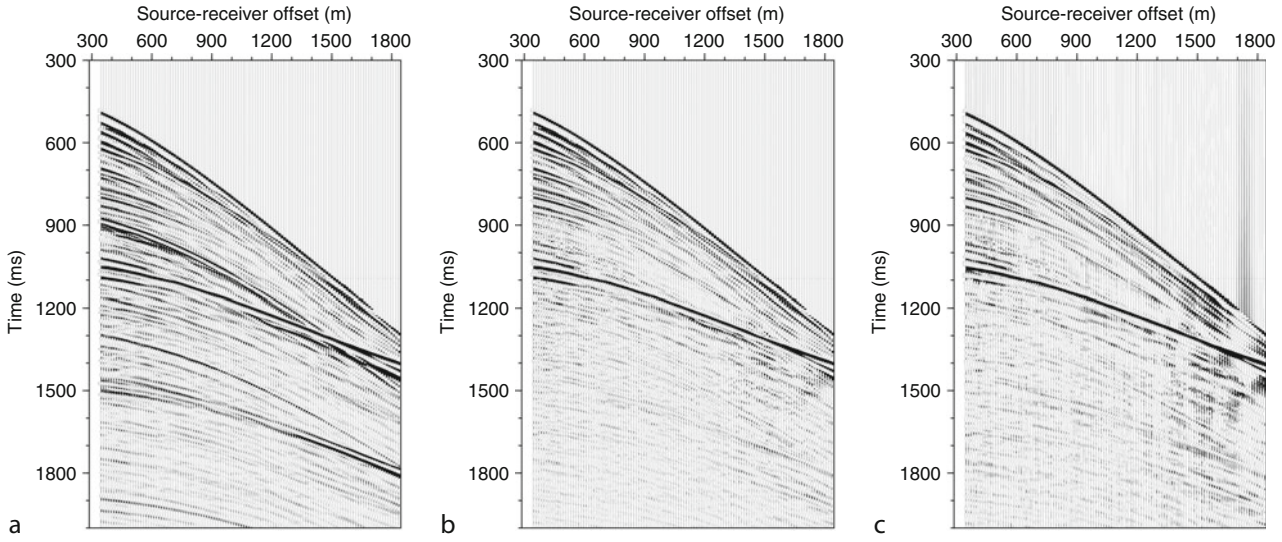
$$G_{2D}(\mathbf{r}, \mathbf{r}_s, \omega) = \left(\frac{c}{i8\pi\omega R}\right)^{1/2} \exp\left(-i\omega\frac{R}{c}\right), \tag{32}$$

where $R = |\mathbf{r} - \mathbf{r}_s|$ is the distance from the source, and c is the acoustic velocity of the medium. For a three-dimensional (3-D) case, Green's function is

$$G_{3D}(\mathbf{r}, \mathbf{r}_s, \omega) = \frac{1}{4\pi R} \exp\left(-i\omega\frac{R}{c}\right). \tag{33}$$

Comparing the 2-D and 3-D Green's functions produces a partial compensation operator as

$$W = \sqrt{\frac{2\pi R c}{i\omega}}. \tag{34}$$



Seismic, Waveform Modeling and Tomography, Figure 6 A marine seismic data example. (**a**) A sample shot record with 120 traces. (**b**) The shot record after multiple attenuation. (**c**) The same shot record after partial compensation.

In the time domain, the operator W shows the following behavior (in the far-field),

$$W(t) = D_{-1/2}(t)\sqrt{2\pi Rc}, \tag{35}$$

where $D_{-1/2}(t)$ is a half-integrator, defined as the inverse Fourier transform of $(i\omega)^{-1/2}$ (Deregowski and Brown, 1983). For a narrow-offset, reflection geometry, one can assume $2R \propto ct$ and obtain

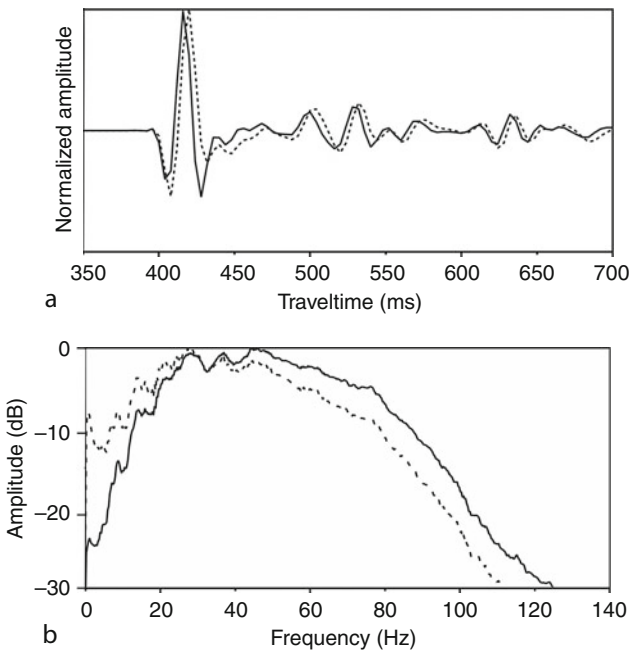
$$W(t) \propto D_{-1/2}(t)c(t)\sqrt{t}. \tag{36}$$

Therefore, this partial compensation may be implemented in two steps: scaling $c(t)\sqrt{t}$ in the time domain, and multiplication $(i\omega)^{-1/2}$ in the frequency domain.

Figure 7 closely compares a seismic trace from a point source (solid curve) and a trace after partial compensation (dash curve). With the application of the operator $(i\omega)^{-1/2}$, wavelets in the trace from an equivalent line source are broader than those actually generated from a point source.

Alternatively, one could use a 2.5-D wave modeling and inversion scheme for waveform tomography. But as it involves integration along the infinite line-source direction that is perpendicular to the source-receiver profile, a 2.5-D scheme would take a much longer running time, compared to the 2-D wave modeling and inversion (Song and Williamson, 1995).

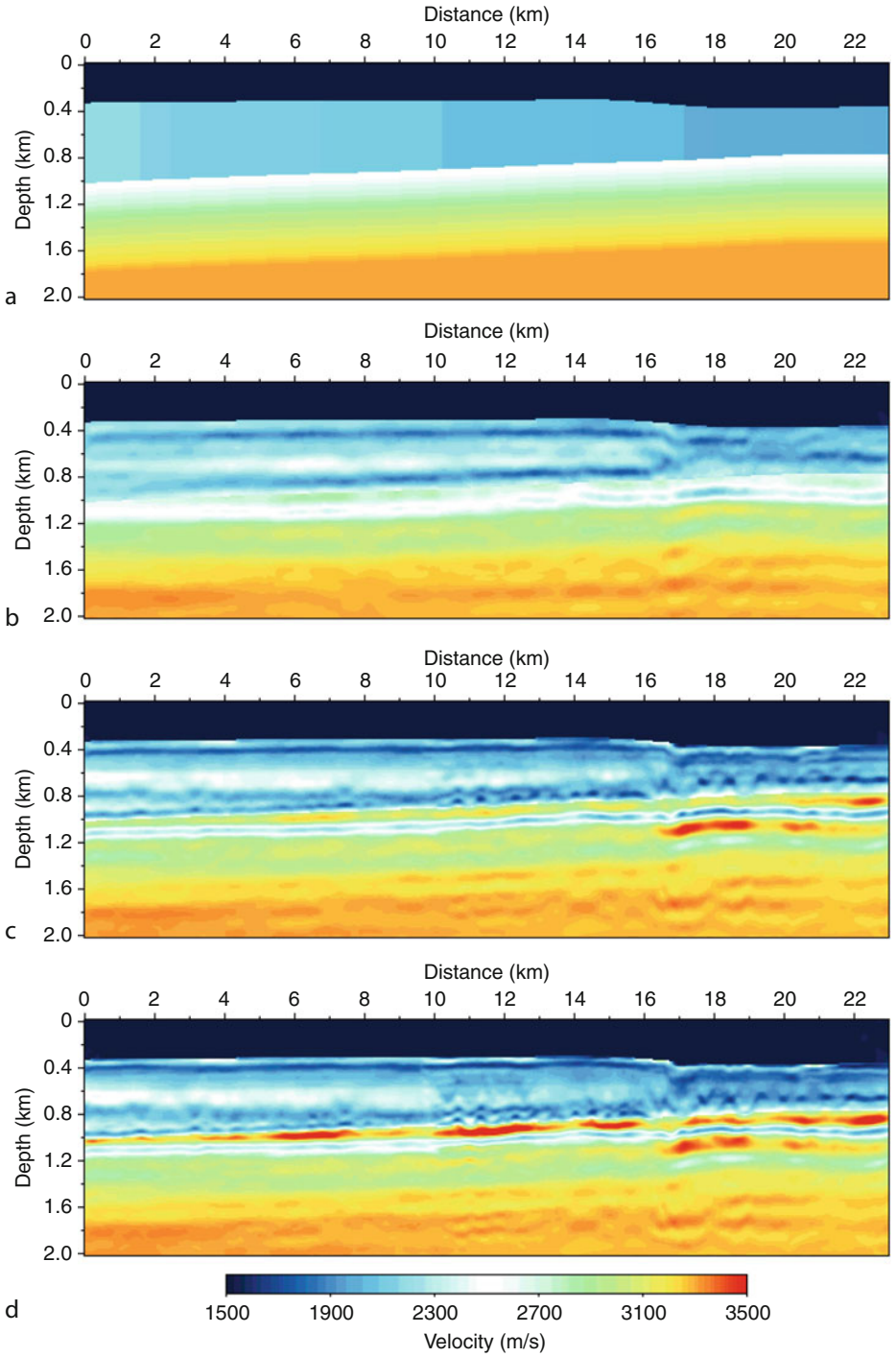
The frequency ω can also be complex-valued, to include an exponential function either for the amplitude attenuation or compensation. The attenuation used in wavefield $P(t)$ is for suppressing the wrap-around effect in the Fourier transform domain. An opposite sign with an exponential increase in $P(t)$ could also be used to boost the energy of deep reflections. In the latter case, the real data should be balanced with an automatic gain control.



Seismic, Waveform Modeling and Tomography, Figure 7 (**a**) Comparison between a seismic trace from an actual point source (*solid*) and the trace after partial compensation (*dash curve*). Wavelets in a trace from a line source (i.e., after partial compensation) are broader than those from a point source. (**b**) Comparison of amplitude spectra of seismic traces from a point source and a line source.

Reflection-seismic waveform tomography

In the reflection-seismic data, there is no significant energy recorded at low frequencies less than 6 Hz. A travel time inversion is often performed to generate the initial velocity model for the iterative waveform tomography (Pratt et al., 2002). Referring to the stack section (Figure 5), we pick the travel times of two reflections from the pre-stack seismic data, and run a travel time



Seismic, Waveform Modeling and Tomography, Figure 8 Waveform tomography. (**a**) The initial velocity model built from travel time tomography. (**b**) The velocity model of waveform tomography using frequencies in the range of 6.9–7.5 Hz. (**c**) The velocity model of waveform tomography using frequencies in the range of 6.9–13.8 Hz. (**d**) The final velocity model obtained from the waveform tomography using all frequencies in the range of 6.9–30 Hz.

inversion to generate a layered velocity model with two interfaces (Figure 8a). The first layer is a water layer with velocity 1,500 m/s. The second layer has a linear velocity of 2,200 m/s at one end and 2,000 m/s at distance 23 km.

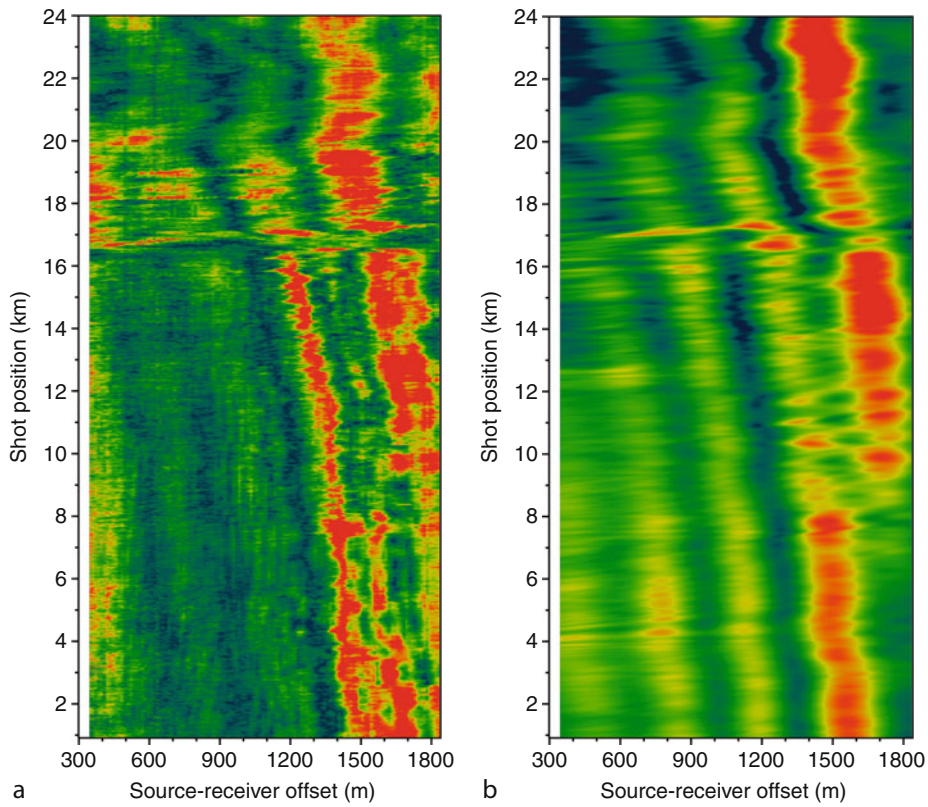
In the frequency-domain waveform tomography, we use a group of three frequencies simultaneously in each iterative inversion. Using a group of neighboring frequencies in the input can suppress the noise effect in the real data, and also more data samples used in an inversion means a much better determined inverse problem. In this example, there are 26 groups of frequencies in total in the range of 6.9–30 Hz with an interval of 0.3 Hz. The first group includes frequencies 6.9, 7.2, and 7.5 Hz, and the last group includes frequencies 29.4, 29.7, and 30 Hz. Figures 8b and 8c are the velocity models from waveform tomography using frequencies in ranges of 6.9–7.5 Hz and 6.9–13.8 Hz, respectively, where Figure 8d is the final velocity model obtained from waveform tomography using all frequencies in the range.

Although the starting model generated by travel time tomography generates a smooth boundary for the water bottom, waveform tomography produces a sharp geometry with a spatial variation close to that shown in seismic sections (Figure 5). The tomography image shows clearly a stratified structure underneath the water layer. In the second layer, at the distance between 0 and 10 km, there is a high-low-high vertical velocity variation immediately underneath the water bottom, between depth of 350 and 700 m at the left end of the profile, and a high-low-high-low vertical variation above the second interface. At the distance between 17 and 23 km, there is also a low-high-low vertical velocity variation within the second layer.

Most significantly, in the third layer, high-low-high velocity channels appear immediately underneath the second interface and cross the entire section. These thin-layers separation could not be generated by any conventional travel time tomography. Beneath that, the velocity pattern in the depth between 1.2 and 1.6 km varies laterally between 0–5, 5–17, and 17–23 km in distance.

Seismic reflection is generated due to the contrast in impedance, which is the product of density and velocity. In the inversion example here however, the density is assumed to be constant. Therefore, for a more accurate quantitative interpretation of the high velocity variation within the topmost thin layer immediately underneath the water layer, a density contrast between the water and water-bottom sediment needs be compensated.

Figure 9 compares a sample frequency slice (at 9.3 Hz) of real data input to waveform tomography and synthetic data generated from the inversion result. The vertical axis is the shot position in the surface, and the horizontal axis is the source-receiver offset. The data displayed are the



Seismic, Waveform Modeling and Tomography, Figure 9 (**a**) A sample amplitude slice of real seismic data at frequency 9.3 Hz. (**b**) The amplitude slice of synthetics at the same frequency, obtained from the waveform tomography result.

amplitudes of complex-valued samples in the frequency domain, and are normalized based on their RMS values. For the far-offset seismic traces, there are strong amplitudes for post-critical angle reflections, which play an important role in the inversion for the shallow part of model. There is also a good fit at near-offset traces, which influence the high wave number perturbations of the velocity field.

Summary

Waveform tomography uses the original waveform recordings and wave equation modeling to extract high-resolution tomographic images from seismic data. Travel time inversion may be used first to provide a reliable initial model for the waveform inversion. The latter uses a group of frequencies simultaneously in an iterative inversion, and proceeds from low to high frequencies. Even for reflection-seismic data with limited source-receiver offset, the waveform tomography also can potentially image subsurface features with detailed spatial variation at sub-wavelength scales.

Bibliography

Alford, R. M., Kelly, K. R., and Boore, D. M., 1974. Accuracy of finite-difference modelling of the acoustic wave equation. *Geophysics*, **39**, 834–842.

Boonyasiriwat, C., Valasek, P., Routh, P., Cao, W., Schuster, G. T., and Macy, B., 2009. An efficient multiscale method for time-domain waveform tomography. *Geophysics*, **74**, WCC59.

Bunks, C., Saleck, F. M., Zaleski, S., and Chavent, G., 1995. Multiscale seismic waveform inversion. *Geophysics*, **60**, 1457–1473.

Deregowski, S. M., and Brown, S. M., 1983. A theory of acoustic diffractors applied to 2-D models. *Geophysical Prospecting*, **31**, 293–333.

Kelly, K. R., Ward, R. W., Treitel, S., and Alford, R. M., 1976. Synthetic seismograms: a finite difference approach. *Geophysics*, **41**, 2–27.

Lailly, P., 1984. Migration methods: partial but efficient solutions to the seismic inverse problem. In Santosa, F., Pao, Y. H., Symes, W., and Holland, Ch. (eds.), *Inverse Problems of Acoustic and Elastic Waves*. Philadelphia, PA: SIAM.

Levander, A., 1988. Fourth-order finite-difference P-SV seismograms. *Geophysics*, **53**, 1425–1437.

Pratt, R. G., 2008. Waveform tomography – successes, cautionary tales, and future directions. In workshop 11, "Full Waveform Inversion – Current Status and Perspectives," *70th EAGE Conference & Exhibition*, Rome.

Pratt, R. G., and Shipp, R. M., 1999. Seismic waveform inversion in the frequency domain, Part 2: fault delineation in sediments using crosshole data. *Geophysics*, **64**, 902–914.

Pratt, R. G., and Worthington, M. H., 1990. Inverse theory applied to multisource crosshole tomography. Part I: acoustic wave-equation method. *Geophysical Prospecting*, **38**, 287–310.

Pratt, R. G., Shin, C., and Hicks, G. J., 1998. Gauss-Newton and full Newton methods in frequency-space seismic waveform inversion. *Geophysical Journal International*, **133**, 341–362.

Pratt, R. G., Gao, F., Zelt, C., and Levander, A., 2002. A comparison of ray-based and waveform tomography: implications for migration. In *64th EAGE Conference & Exhibition*, Florence.

Rao, Y., and Wang, Y., 2009. Fracture effects in seismic attenuation images reconstructed by waveform tomography. *Geophysics*, **74**, R25–R34.

Sirgue, L., and Pratt, R. G., 2004. Efficient waveform inversion and imaging: a strategy for selecting temporal frequencies and waveform inversion. *Geophysics*, **69**, 231–248.

Song, Z. M., and Williamson, P. R., 1995. Frequency-domain acoustic wave modeling and inversion of crosshole data, I: 2.5-D modeling method. *Geophysics*, **60**, 784–795.

Štekl, I., and Pratt, R. G., 1998. Accurate viscoelastic modeling by frequency-domain finite differences using rotated operators. *Geophysics*, **63**, 1779–1794.

Tarantola, A., 1984. Inversion of seismic reflection data in the acoustic approximation:. *Geophysics*, **49**, 1259–1266.

Tarantola, A., 1987. *Inverse Problem Theory: Methods for Data Fitting and Parameter Estimation*. Amsterdam: Elsevier.

Virieux, J., 1986. P-SV wave propagation in heterogeneous media: velocity-stress finite-difference method. *Geophysics*, **51**, 889–901.

Wang, Y., 2004. Multiple prediction through inversion: a fully data-driven concept for surface-related multiple attenuation. *Geophysics*, **69**, 547–553.

Wang, Y., 2007. Multiple prediction through inversion: theoretical advancements and real data application. *Geophysics*, **72**, V33–V39.

Wang, Y., and Rao, Y., 2006. Crosshole seismic waveform tomography, I: Strategy for real data application. *Geophysical Journal International*, **166**, 1237–1248.

Wang, Y., and Rao, Y., 2009. Reflection seismic waveform tomography. *Journal of Geophysical Research*, **114**, B03304.

Warner, M., Stekl, I., Umpleby, A., Morgan, J., Pain, C., and Wang, Y., 2008. 3D wavefield tomography: problems, opportunities and future directions. In *70th EAGE Conference & Exhibition*, Rome.

Wu, R., and Toksöz, M. N., 1987. Diffraction tomography and multisource holography applied to seismic imaging. *Geophysics*, **52**, 11–25.

Cross-references

Seismic Data Acquisition and Processing
Seismic Imaging, Overview
Seismic, Ray Theory
Traveltime Tomography Using Controlled-Source Seismic Data

SEISMICITY, INTRAPLATE

Paul Bodin
Pacific Northwest Seismic Network (PNSN), University of Washington, Seattle, WA, USA

Definition

Intraplate Seismicity refers to earthquakes that take place on faults other than those that form the boundaries between Earth's tectonic plates.

Historical context

From the inception of plate tectonics as a simple set of unifying concepts, Earth scientists recognized that the notion of completely rigid plates should not be taken to be rigorous or complete. Deformations of Earth's crust and upper mantle, and the earthquakes that generally accompany them, must take place off of recognized plate interfaces. This fact was evident theoretically because in order for

plate tectonics to operate on timescales greater than one Wilson cycle, new plate boundaries must form within plates, so deformation must take place off of existing plate boundaries. In addition, it was unassailable observationally because seismicity off of recognized plate boundaries clearly *does* take place, albeit at a relatively trivial pace compared to Earth's total earthquake production.

Two generations later, despite the fact that they obviously exist, intraplate earthquakes remain enigmatic and challenging; fitting subjects for special study. The dominant questions about these earthquakes not only reveals our incomplete understanding of earthquake physics in general, but also highlights a number of issues specific to active tectonics and seismic hazard in intraplate settings.

Definitions and classification

We may define interplate earthquakes as being those that rupture parts of faults that connect directly mechanically to the globe-girdling system of recognized boundaries between tectonic plates. Intraplate earthquakes break faults that do not so connect. The difference is one of ultimate slip. Because interplate faults form a network within which new plates may be created and old plates consumed, slip on them is unbounded. On the other hand, the total slip on intraplate faults must terminate at the ends of the fault. So increasing slip comes at the cost of increasing strain in the surrounding rocks which, itself being presumably limited, caps how much any given fault can slip. Therefore, intraplate faults may be expected to slip generally more slowly, and less far, than interplate faults, and this possibly to have some consequence on the characteristics of earthquakes on the two types of faults.

It is customary for Earth scientists to distinguish further between intraplate earthquakes that break faults within *mobile zones* and those that take place within *stable zones*. Mobile zones are regions that, although not directly on recognized plate boundary fault, have been experiencing regional deformation distributed over a broad area near a plate boundary. An example is the active faulting across China, Tibet, and Mongolia, the deformations of which may be clearly related to three-dimensional complexities in the plate boundary framework. Stable zones are not only sufficiently remote from plate boundaries to be insensitive to their geometric complexities, but also are stable in the sense that significant tectonic events have not affected them during the Cenozoic Era ($\sim$66 million years). They are generally expected to be areas of cold, thick continental lithosphere. Earthquakes in stable continental regions, often designated SCR earthquakes, are particularly enigmatic.

Reasons to distinguish intraplate earthquakes

One reason to study intraplate earthquakes (and particularly SCR earthquakes) as a class involves a practical consequence: they strike the most unwary victims. Because of the long recurrence intervals between intraplate earthquakes, the threatened populace is generally unprepared and the structures they build are particularly vulnerable. It has recently been noted that this argument is becoming less restrictive to intraplate earthquakes as population growth within plate boundary regions has taken place so quickly of late that even these regions are becoming less prepared as a whole (Bilham, 2009). Nevertheless, the argument remains applicable in plate interiors where the building stock may be not only not up to modern seismic codes, but is generally older.

In part, SCR earthquakes remain especially enigmatic not merely because they are rare, but also because their recurrence may be so long that (1) the sizes and sometimes locations of previous earthquakes are not well known, even for those that have occurred in historical times, and (2) it is usually not possible to ascertain what their recurrence is with certainty. This leads to great uncertainty about the seismic hazards in intraplate regions. Is damaging earthquake shaking possible anywhere (albeit with small probability within any 50-year time period)? Or are even SCR earthquakes restricted to occur in places with discernible (and hence predictable) characteristics? This is tantamount to a question of whether the entirety of Earth's crust at critical stress levels such that only small fluctuations or differences in fault strength control the occurrence of earthquakes.

With such long recurrence intervals, seismological observations are sparse and we rely on the paleoseismic evidence contained within the geological record to try to understand them. It is clear that plate tectonics is accurate to first order and that intraplate deformation takes place very slowly, which makes them a challenge for even modern high-precision geodetic techniques. It seems likely that progress in understanding intraplate earthquakes will remain slow for the time being.

Another reason they are of special interest is that they may reveal clues about elusive earthquake physics. While Earth scientists believe that nearly all tectonic earthquakes are caused by sudden slip on a fault driven by stresses in the surrounding rock (the elastic rebound theory), the specific mechanisms that localize rupture initiation and growth of a rupture remain obscure. Systematic differences between fault, stress, and geological characteristics of intraplate faults discussed below may be exploited to test hypotheses of the important controls of earthquake fault rupture.

Do intraplate and interplate earthquakes scale differently?

Because intraplate earthquakes, and particularly those in stable zones, have longer recurrence intervals, it is often hypothesized that their host faults will have longer time to "heal" between earthquakes (e.g., Scholz, 2002). Fault healing here refers to a set of hypothesized fault processes such as the formation of minerals that bind the sides of the

fault together making them stronger over time. Thus, intraplate earthquakes may be expected to break under higher loads – at higher stress levels – than their cousins that break more active plate boundary faults. The anticipated result is that they "scale" differently: that is, big earthquakes differ from small earthquakes in systematically different ways in interplate as opposed to intraplate regions. Specifically, intraplate earthquake ruptures will slip more in an earthquake per unit of fault area, relieve more stress (have a larger "stress drop") and produce more seismic energy for their size, than the earthquakes we are more familiar with at plate boundaries. Thus seismologists expect, and arguably observe, that intraplate earthquakes tend to have higher stress drops than interplate earthquakes. Seismological evidence suggests that as a class their stress drops may be a factor of 2–5 higher than plate boundary earthquakes (Kanamori and Allen, 1986). But because observations are so few, much uncertainty remains about this issue.

Source scaling studies are further hindered by systematic differences between seismic wave propagation in intraplate regions and that in more mobile or active zones. Because rock surrounding plate boundaries is often less intact and hotter than in plate interiors, seismic wave energy is more strongly absorbed, or attenuated, than in the cold hard hearts of plates. So, per unit of energy released at their source, the effects of intraplate earthquakes are more widely distributed – they have a longer reach.

What drives intraplate earthquakes?

Probably the most critical enigma of intraplate seismicity concerns the source of driving stresses. It is easy to invoke "left over" or residual stresses originating from geometric irregularities or variations in frictional properties on the master plate boundary faults to explain the origin of mobile zone intraplate earthquakes. Remote from plate boundaries, occasionally sources of localized stresses may be recognized. One clear example is from intraplate regions experiencing uplift from glacial rebound, as ice sheets retreat with global warming. This mechanism well explains the seismicity pattern of Sweden, for example (e.g., Mörner, 2009). Swarms of seismicity often accompany geothermal activity, even at sites in the remote interior of tectonic plates (e.g., Ibs-von Seht et al., 2008). Where such sources of localized stress are not evident, however, it is more problematic to assign the provenance of seismic stress. Modeling studies suggest that stresses within plates may be best explained as a combination of resolved stresses originating at plate boundaries (although they may be thousands of kilometers distant) plate bending due to a plate's travel over a non-spherical Earth, or forces arising from convective patterns in Earth's mantle.

Many scientists presume that intraplate earthquakes take place at sites where crustal rocks have been significantly weakened compared to those in surrounding rocks. In this model, seismicity should be associated with reactivated fault zones. Big intraplate earthquakes are thought to be restricted to areas that have previously been faulted, say within failed continental rift systems (e.g., Johnson, 1996).

One intriguing idea is that the strength of faults could be reduced by high pore-fluid pressure. Fluids (which could include water or even carbon dioxide or hydrocarbons) that occupied the internal portion of a fault could become pressurized and reduce the frictional resistance of the fault to slipping. That would imply that seismically active faults in intraplate regions are those that contain significant quantities of fluids at high confining stress. Sufficient stress to drive earthquakes could indeed be present everywhere within plate interiors, and the conditions to release the stress controlled by something external.

Examples

The moment magnitude (Mw) = 6.8 "Nisqually" earthquake of January 23, 2001 in Washington state, USA, provides an example of an intraplate earthquakes breaking a fault within the crust of the subducting ocean plate (Ichinose et al., 2006). The earthquake resulted from slip on a normal fault driven by bending stresses as the downgoing slab flexes in its descent. On a larger scale, the Cascadia subduction zone provides an interesting example of the interplay between inter- and intraplate earthquakes in that while the master plate boundary megathrust fault is broken about every 500 years with large subduction thrust events (e.g., Satake and Atwater, 2007), there appears to be no current seismicity on the plate boundary fault. Rather all earthquakes either historical or instrumentally recorded are either intra-slab events or crustal earthquakes within the overriding North America plate. These events are interplate by our definition. However, they are clearly associated with a nearby subduction interface and are driven by residual stresses that the plate boundary is unable to completely relieve.

The May 13, 2008 Mw = 7.9 Wenchuan earthquake in the Szechuan Province of China is an earthquake in the tectonically mobile zone of Asia, and is part of the escape of material from Asia and the uplifting of the Himalaya due to the collision of the Indian subcontinent with Asia. About 300 km of a fault on the boundary between the Longmenshan orogenic belt and the Szechuan basin broke in the earthquake (Burchfiel et al., 2008). With a high overall stress drop of 17 MPa and localized patches of up to 53 MPa (Zhang et al., 2009), the Wenchuan earthquake had large slip and disastrous effects on a populace that was largely unaware of the potential for, and unprepared for, this long-recurrence-interval (2,000–10,000 years) intraplate earthquake.

The January 26, 2001 Mw = 7.6 Bhuj earthquake took place ~400 km? from the rather diffuse onshore portion of the left-lateral strike-slip Makran boundary between India and the Arabian and Eurasian plates, but within the

western part of the peninsular Indian shield. It broke the deeper part (~5 to ~40 km) of a previously unknown reverse fault. It was located within ~100 km of a similar-sized (though shallower) earthquake in 1835 that generated a surface rupture – the Allah Bundh in the Great Rann of Kachchh. The stress drop was a high, but not extreme ~16–20 MPa, but the earthquake was very powerful, with the source lasting only about 7 s. Widespread liquefaction and secondary effects were observed, but the large death toll (~30,000) was apparently not the result of extreme ground motions, which could have been engineered for, but rather of poor construction and preparation.

Controversy

One particularly illustrative and significant example of SCR interplate seismicity concerns the New Madrid Seismic Zone that threatens a 7-state region surrounding the lower Mississippi River valley in the south central USA (e.g., Johnson and Schweig, 1996). The region was the site of at least three major earthquakes during the winter of 1811–1812. While at the time thinly populated and rural, a return of these earthquakes today would affect ten million people and put at risk numerous cities, commerce, and transportation and communication infrastructure. The 1811–1812 earthquakes were accompanied by a host of surface effects (liquefaction, landslides, surface deformation) that remain visible in the landscape today. Ground motions from the earthquakes were also reported to be felt as far away as Charleston, South Carolina (~950 km) and even Boston, Massachusetts (~1,600 km). There is a complex zone of active microearthquakes producing several $M < 3$ earthquakes each week. Early research based on ground-shaking intensities derived from historical reports proposed that the three largest earthquakes in the series exceeded M8. The ~100-km long zone of active microseismicity, which many presume to represent the active faults, seems much too small to host even a single M8 earthquake, let alone three, unless stress drops, and therefore slip per unit length, were extremely large. The earthquakes occupy a shallow basin formed atop a failed continental rift, formed in latest Precambrian time and briefly reactivated (but once again failed) in the Cretaceous. The faults are oriented such that they are nearly optimally oriented to be reactivated as thrust and strike-slip fault in the current compressional mid-plate stress regime.

Subsequent reexamination of the historic intensity data, and taking into account the feeble attenuation of Seismic waves within rocks in the plate interior, have yielded generally smaller estimates of magnitudes, as low as M7 for the three largest 1811–1812 earthquakes. In the absence of new information, however, it might be best to admit that while these earthquakes were undoubtedly large, their magnitudes will remain forever moot. Paleoseismic studies of earthquake-induced liquefaction features in surface deposits reveal at least four repeated episodes of strong shaking (as strong as what took place in the early nineteenth century), often in clusters like the 1811–1812 events, with an average recurrence interval of ~500 years. The geological evidence for recurrent large earthquakes is balanced by the lack of primary surface faulting, or even the prominent topographic relief one would expect if much slip had taken place at the rates paleoseismic evidence infers for the past several thousand years. So either the big events are really a good deal smaller than the largest estimates of their size or seismic activity in the zone was initiated only a few thousand years ago.

Geodetic evidence, from repeated campaign surveys and continuous GPS studies show that deformation surrounding the active seismic zone (Smalley et al., 2005) is very slow, with some (e.g., Newman et al., 1999) arguing that it is not different from zero. However, given the lack of a mid-plate fault model it is not clear what deformation pattern one should expect. So there is no general agreement about what the lack of observed active deformation means, and the interpretations remain contentious. Some suggest that either the largest earthquakes to be expected from the seismic zone are not so terrifyingly large as previously thought, or perhaps that, however large they were, they are not being recharged with stress and are not to be feared. Others hold that the paleoseismic, historical, and microseismic evidence point to the likelihood of future damaging earthquakes just like the historical ones, albeit driven by processes that are still beyond our ken.

Summary or conclusions

Intraplate seismicity is a catchall phrase that encompasses a wide variety of seismogenic phenomena; essentially all earthquakes that do not occur on plate bounding faults. Locales hosting intraplate earthquakes range from broad deformation zones that surround some plate boundaries to areas that lie within the cores of continents. Larger earthquakes – especially those in what are otherwise tectonically stable cold continental crust – appear to be associated with faults that formed in an earlier tectonic setting but are reactivated by the stress field applied currently. Not all faults are so reactivated, so it is likely that some as-yet-unclear weakening mechanism causes only certain faults to fail. Localized deformation, say glacial rebound or geothermal stresses can be responsible for interplate earthquakes, but does not explain all of them. As a class of faults, intraplate faults usually slip more slowly and less far than interplate faults, and this leads to longer intraplate earthquake recurrence intervals. Also, and probably related to this, intraplate earthquakes may generally be expected to have larger stress drops. Although less frequent than plate boundary earthquakes, intaplate seismicity is very dangerous because the population and infrastructure at risk from them is less prepared than in more seismically active areas. The study of intraplate seismicity is an endeavor that, because of its low rate is best done by integrating across different source regions and combining different Earth science disciplines.

Bibliography

Bilham, R., 2009. The seismic future of cities. *Bulletin of Earthquake Engineering*, doi:10.1007/s10518-009-9147.

Burchfiel, B. C., Royden, L. H., van der Hilst, R. D., Hager, B. H., King, R. W., Li, C., Lü, J., Yao, H., and Kirby, E., 2008. A geological and geophysical context for the Wenchuan earthquake of 12 May 2008, Sichuan, People's Republic of China. *Geological Society of America Today*, **18**(7), doi:10.1130/GSATG18A.1.

Ibs-von Seht, M., Plenefisch, T., and Klinge, K., 2008. Earthquake swarms in continental rifts – a comparison of selected cases in America, Africa and Europe. *Tectonophysics*, **452**(1–4), 66–77, doi:10.1016/j.tecto.2008.02.008. ISSN 0040-1951.

Ichinose, G. A., Thio, H. K., and Somerville, P. G., 2006. Moment tensor and rupture model for the 1949 Olympia, Washington, earthquake and scaling relations for Cascadia and Global Intraslab Earthquakes. *Bulletin. Seismological Society of America*, **96**(3), 1029–1037, doi:10.1785/0120050132.

Johnson, A. C., 1996. Seismic moment assessment of earthquakes in stable continental regions II: historical seismicity. *Geophysical Journal International*, **125**(3), 639–678.

Johnson, A. C., and Schweig, E. S., 1996. The enigma of the New Madrid Earthquakes of 1811–1812. *Annual Review of Earth and Planetary Sciences*, **24**, 339–384.

Kanamori, H., and Allen, C. R., 1986. Earthquake repeat time and average stress drop. In Das, S., et al. (eds.), *Earthquake Source Mechanics*. Washington: American Geophysical Union. AGU Geophysics Monograph, Vol. 37, pp. 227–236.

Mörner, N-A., 2009. Late Holocene earthquake geology in Sweden, *Geological Society, London*, Special publications, 316, pp. 179–188, doi:10.1144/SP316.11

Newman, A., Stein, S., Weber, J., Engeln, J., Mao, A., and Dixon, T., 1999. Slow deformation and lower seismic hazard at the New Madrid Seismic zone. *Science*, **284**, 619–621, doi:10.1126/science.284.5414.619.

Satake, K., and Atwater, B. F., 2007. Long-term perspectives on giant earthquakes and tsunamis at subduction zones. *Annual Review of Earth and Planetary Sciences*, **35**, 349–374, doi:10.1146/annurev.earth.35.031306.140302.

Scholz, C. H., 2002. *The Mechanics of Earthquakes and Faulting*. Cambridge, UK: Cambridge University Press.

Smalley, R., Ellis, M. A., Paul, J., and Vanarsdale, R., 2005. Space geodetic evidence for rapid strain rates in the New Madrid seismic zone of central USA. *Nature*, **435**, 1088–1090, doi:10.1038/nature03642.

Zhang, Y., Feng, W., Xu, L., Zhou, C., and Chen, Y., 2009. Spatio-temporal rupture process of the 2008 great Wenchuan earthquake. *Science in China. Series D: Earth Sciences*, **52**(2), 145–154.

Cross-references

Earthquake, Focal Mechanism
Earthquake, Magnitude
Earthquakes and Crustal Deformation
Earthquakes, Strong-Ground Motion
Geodesy, Figure of the Earth
Geodesy, Ground Positioning and Leveling
Geodynamics
Mantle Viscosity
Paleoseismology
Plate Driving Forces
Plate Motions in Time: Inferences on Driving and Resisting Forces
Plates and Paleoreconstructions
Plate Tectonics, Precambrian

SEISMICITY, SUBDUCTION ZONE

Akira Hasegawa
RCPEV, Graduate School of Science, Tohoku University, Sendai, Miyagi, Japan

Definition

Subduction zone. Consuming plate boundary where the oceanic plate subducts into the mantle beneath the continental plate.

Interplate earthquakes. Earthquakes that occur along the plate boundary. In case of the subduction zone, they occur at the shallow portion of the boundary between the subducting plate and the overriding plate.

Intraslab earthquakes. Earthquakes that occur within the subducted plate, that is, within the slab.

Shallow earthquakes. Earthquakes that occur at depths shallower than 60 km. Focal depths of interplate earthquakes in subduction zones are usually shallower than 60 km, and so they are shallow earthquakes.

Intermediate-depth earthquakes. Earthquakes that occur at depths of 60–300 km.

Deep earthquakes. Earthquakes that occur at depths deeper than 300 km. Both intermediate-depth and deep earthquakes occur only within the subducted plate, and so they are intraslab earthquakes.

Introduction

There are seismically active and inactive areas on the Earth. Most of Earth's seismic activity is concentrated in narrow continuous belts that wrap around it. Plate tectonics proposes that the Earth's entire surface is comprised of somewhat more than ten large, rigid, mosaic-like plates, and that the relative movement of these plates causes tectonic activity at their boundaries. Consequently, most earthquakes are concentrated at plate boundaries, forming narrow, belt-shaped seismic zones.

Oceanic plates are generated at mid-ocean ridges by the upwelling of mantle material. Since the Earth's surface area is constant, the surface area does not expand with the generation of the oceanic plate. Instead, the same amount of material moves back down into the mantle. The generation of plates is balanced by the consumption of other plates elsewhere at consuming plate boundaries. *Subduction zones* are where oceanic plates plunge downward into the mantle. At subduction zones, two plates collide with each other; the heavier plate is overridden by the lighter one and sinks down into the mantle. Thus, the heavier oceanic plate subducts beneath the lighter continental plate. The subduction of the oceanic plate causes high seismic activity. In fact, it is in these subduction zones that most of the world's large earthquakes occur.

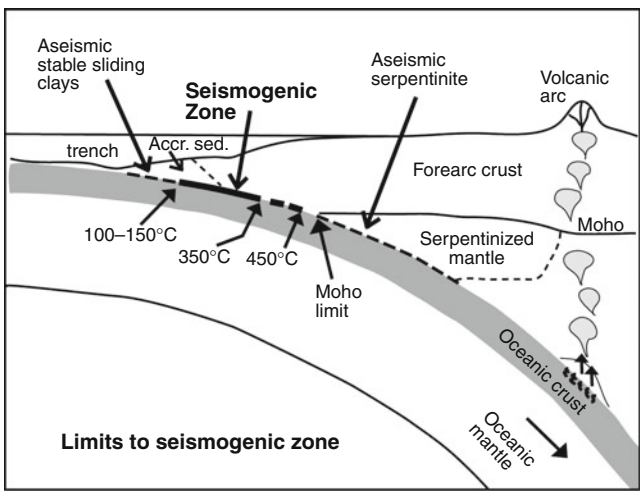
Interplate earthquakes

In subduction zones, the denser oceanic plate sinks down into the mantle underneath the lighter continental plate.

The shallow portion of the plate boundary is locked by friction, and in order to release accumulated stress, occasional slips at the plate boundary occur and cause earthquakes. Most of the world's large, shallow earthquakes are interplate earthquakes caused in this way. Interplate coupling occurs only within a specific range of depths, which seems to be primarily determined by the temperature of the plate interface.

Oleskevich et al. (1999) estimated the thermal structures of plate interfaces in several subduction zones, and compared them with the updip and downdip depth limits of source areas of large interplate earthquakes. The results show: (1) the updip limit of the depth of interplate earthquakes corresponds with the depth at which the temperature of the plate boundary reaches 100–150°C and (2) the downdip limit approximately matches the depth at which the temperature reaches 350°C (Figure 1). In the shallowest part of plate interfaces, smectite clays, which are clay minerals distributed in the sediments there, prevent interplate coupling by producing stable sliding. The temperature rises with increasing depth and when it reaches 100–150°C, dehydration decomposition occurs and the smectite clay turns into illite and chlorite. This transformation is considered the onset of interplate coupling.

Then, deeper, where the temperature exceeds 350°C, interplate coupling ceases and stable sliding takes place again. Figure 1 shows that the transition zone from unstable sliding to stable sliding occurs at depths where the temperature ranges from 350°C to 450°C. In the case where the depth at which the temperature of the plate interface reaches 350°C is deeper than the depth at which it contacts the Moho of the upper plate, the contact zone with the Moho is considered to be the downdip limit. The reason the downdip is limited to the Moho contact zone is that stable sliding prevails where the mantle of the upper plate contacts the oceanic plate because the mantle may be serpentinized by water supplied by the dehydration of the subducting oceanic plate. It is considered that interplate earthquakes do not occur any more under those conditions (Hyndman et al., 1995). In some subduction zones, such as in northeast Japan, the upper plate's mantle is not serpentinized, since the seismic wave velocities there show normal values of non-serpentinized mantle materials. In such cases, the downdip limit is once again determined by temperature, although the specific temperature at which the transition occurs is not yet known.

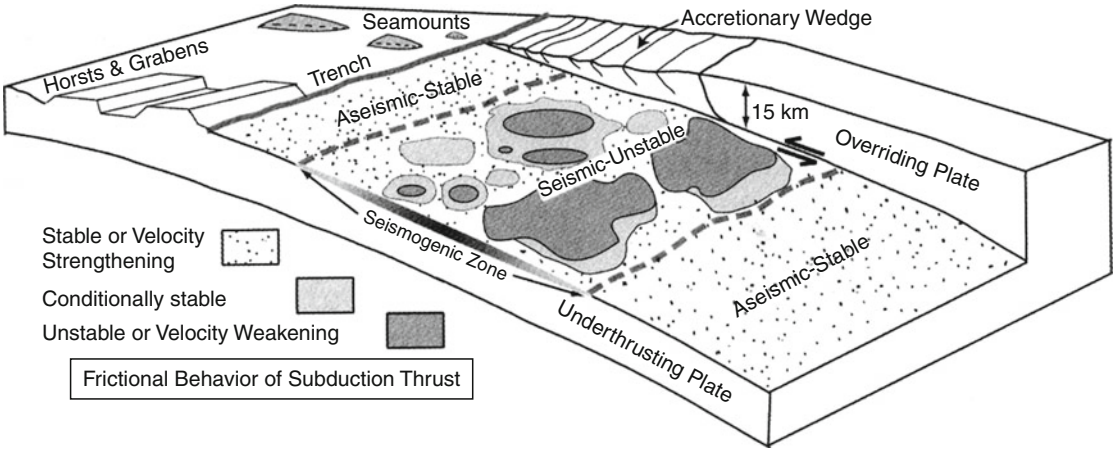


Seismicity, Subduction Zone, Figure 1 Sketch of the subduction plate boundary seismogenic zone (Oleskevich et al., 1999, copyright by the American Geophysical Union).

The degree of interplate coupling can be indicated by how much interplate slip is accounted for by earthquakes. This index is called "seismic coupling" and denotes the percentage of the seismic slip among all slip amounts occurring at the "seismogenic zone" in Figure 1. If seismic coupling is 1, all slip occurring in that area is caused by earthquakes. In contrast, if it is 0, it means that all slip occurring in that area is aseismic stable sliding. Seismic coupling differs from one subduction zone to another, and the larger the maximum magnitude of interplate earthquakes is in a subduction zone, the larger is its seismic coupling (Lay and Kanamori, 1981).

The younger the age of the subducting plate and the faster the plates are converging, the larger the maximum earthquake magnitude (Ruff and Kanamori, 1980) and so the stronger is the interplate coupling. A faster convergence velocity results in a higher rate of stress accumulation, which would accompany a higher rate of earthquake activity. Young plates are generally less dense than older plates, and being relatively light, the gravitational force pulling them downward is weaker and their subduction angle is lower than with older, denser plates. When the subduction angle is low, the contact area between the subducting plate and the upper plate is large. The large contact area and relatively weak downward force produce strong interplate coupling between the plates. Moreover, younger plates have relatively smoother plate surfaces because they do not have long alteration histories after their generation at the midoceanic ridge, which probably allow the formation of larger focal areas of large maximum earthquakes. For these reasons, if two plates have the same subduction velocity, the younger plate will have stronger interplate coupling and larger maximum earthquake magnitude (Kanamori, 1977).

Interplate coupling along the seismogenic zone is heterogeneous both in space and time; the seismogenic zone is not fully locked, as mentioned above, and some portions of it will slide stably during interseismic periods. Seismic waveform inversions of large interplate earthquakes have revealed spatially heterogeneous seismic slip along the seismogenic zone, and backslip inversions of geodetic data, which estimate spatial distribution of interplate coupling, have confirmed spatially heterogeneous strain accumulation along the zone during the interseismic period. A conceptual model of frictional properties on the fault surface based on laboratory-derived fault constitutive laws



Seismicity, Subduction Zone, Figure 2 A conceptual model of frictional properties on the subduction plate boundary (originally from Bilek et al. (2002) and modified by Dixon and Moore (2007), copyright by the Columbia University Press).

(Figure 2) explains the observed heterogeneous interplate coupling along the seismogenic zone. It involves patches of frictionally unstable fault surface surrounded by stable and conditionally stable fault surfaces. Frictionally stable portions of the fault are subject to steady aseismic sliding, whereas unstable patches (asperities) remain locked during the interseismic period. Then they finally fail, as evidenced by the occurrence of earthquakes. Conditionally stable portions typically slide stably, but may fail and slide dynamically if loaded at high strain rates by the seismic slip of a neighboring asperity. Earthquakes nucleate only in the unstable patches, and once an earthquake occurs in an unstable patch its seismic slip propagates throughout the patch, with some penetration into the neighboring conditionally stable areas. This conceptual model indicates that the larger the area occupied by the frictionally unstable patches in the seismogenic zone, the larger the seismic coupling of that subduction zone is.

Shallow inland intraplate earthquakes

Convergence of two plates at subduction zones causes stress accumulation also in the overriding continental plate, producing deformation in it well beyond the trench axis. Usually, this deformation is compressional crustal shortening, which partially accommodates the plate convergence. Shallow earthquakes occur in the crust of the overriding plate, reflecting this ongoing compression. Focal mechanisms of these earthquakes are thrust fault or strike-slip fault with compressional axis oriented nearly in the plate convergence direction. Recurrence intervals of these types of earthquakes are much longer than those of interplate thrust earthquakes. However, once they occur, it often causes serious damages in lives and societies due to the shallow hypocenter and proximity to inhabited areas.

In most cases, these shallow inland earthquakes occur in the depth range between 1–3 and 10–20 km, forming the seismogenic zone mainly in the upper crust of the overriding plate. Deformation is made by aseismic ductile flow at shallow depths above the upper bound because of the presence of unconsolidated granular material, and at large depths below the lower bound because of the onset of plasticity at a critical temperature (Scholz, 1998). This critical temperature is considered to be about 300–400°C, and so the lower bound of the seismogenic zone varies depending on the local thermal gradient.

Active faults, which are surface traces of repeated fault movements by earthquakes, show locations where in the seismogenic zone of the overriding plate large shallow earthquakes have occurred in the recent past, the Quaternary Period. Surveys of active faults further provide information on history of earthquakes, including their magnitudes and the time of the last event, which is basic and important for long-term earthquake forecast of shallow inland intraplate earthquakes.

Although we can learn from active faults where large earthquakes have occurred, it is difficult to understand why earthquakes occur in those locations. Recent investigations based on dense GPS and seismic observation network data in Japan have provided some information on this problem. Analyses of GPS data showed the existence of belt-like zones with concentrated crustal shortening in the overriding plate, along which many large shallow earthquakes have occurred or active faults are distributed (Sagiya et al., 2000; Miura et al., 2004). Seismic tomography studies revealed prominent low seismic velocity zones in the lower crust to the uppermost mantle right beneath these concentrated deformation zones (Nakajima et al., 2001; Nakajima and Hasegawa, 2007). The low seismic velocities are inferred to reflect the existence of aqueous fluids, which are perhaps supplied from the subducted slabs. These observations suggest that the fluids weaken the surrounding crustal materials and cause the concentrated deformation there, leading to large shallow inland earthquakes (Iio et al., 2004; Hasegawa et al., 2005). If this is the case, places to which aqueous fluids

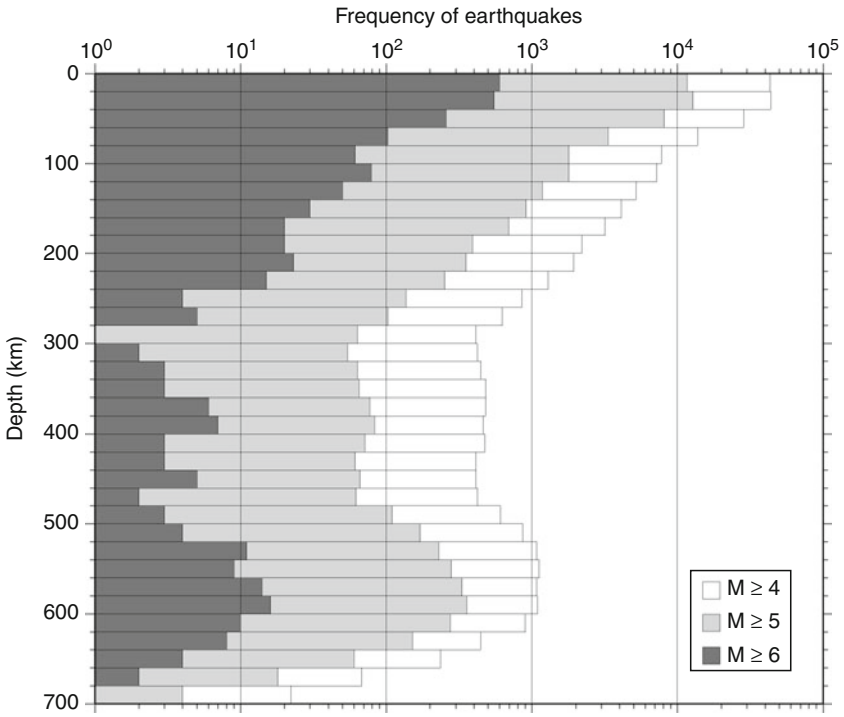
are supplied intensively from the mantle wedge below are considered to be locations where large shallow earthquakes occur.

Intraslab earthquakes

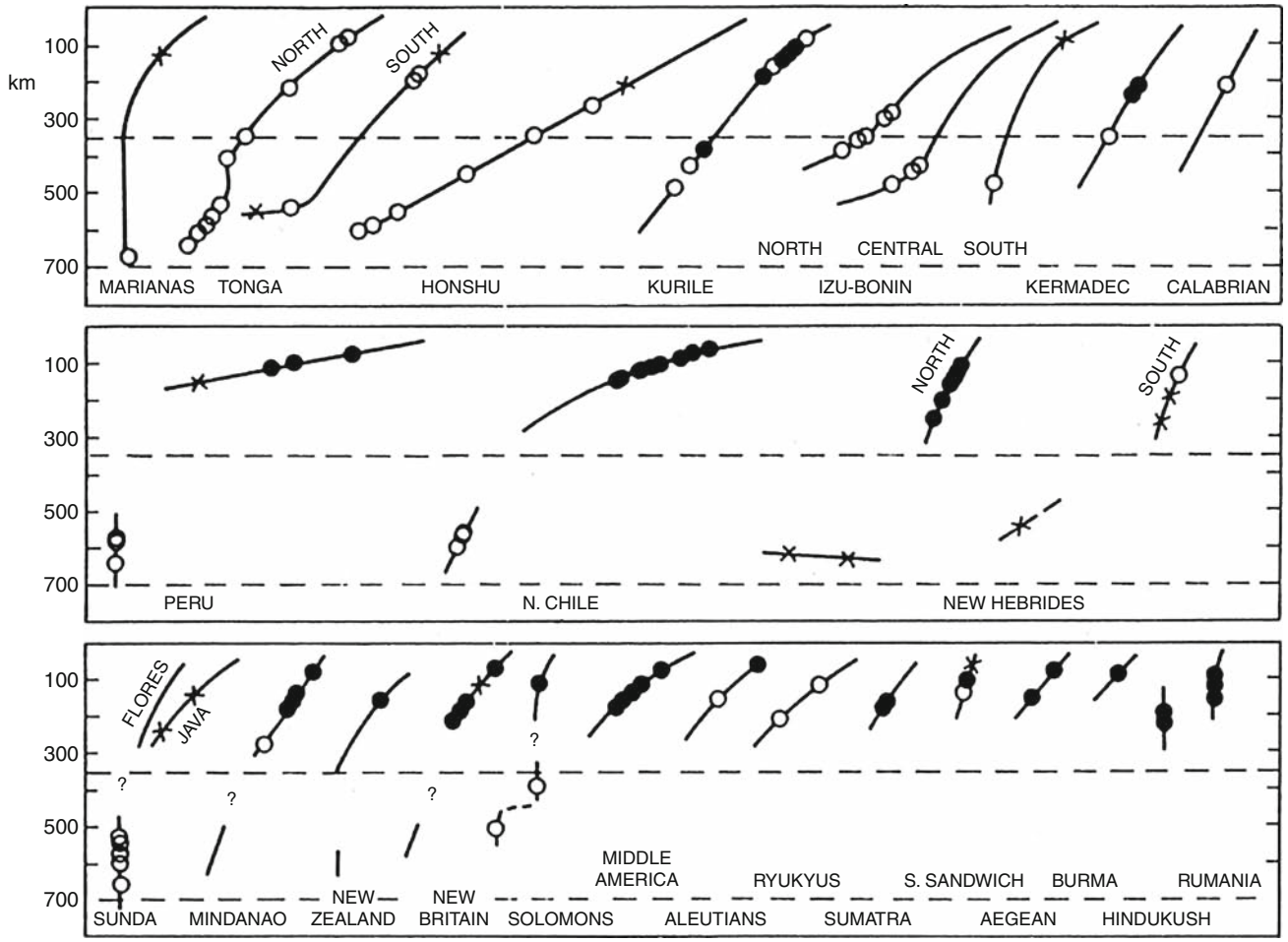
Earthquakes that occur at depths of more than about 60 km are found only in subduction zones. That is because these earthquakes occur in subducting oceanic plates or slabs. They are distributed in an inclined plane in the mantle, which is called "deep seismic zone" or "Wadati–Benioff zone," after its discoverers. The frequency distribution of earthquakes for each depth range throughout the world (Figure 3) shows that depths from 0 to 50–60 km have the greatest occurrence frequency. Proceeding deeper, the frequency decreases nearly monotonically until a depth of about 300 km, then starts to increase with depth from that point, reaching a local maximum at 500–600 km deep. The earthquake depth range can be divided into three zones: shallow (0–60 km), intermediate depth (60–300 km), and deep (deeper than 300 km). Of these, intermediate-depth earthquakes and deep earthquakes occur within the slab so they are called intraslab earthquakes (or slab earthquakes).

In many cases, the focal mechanisms of intraslab earthquakes are classified as either downdip compression (DC) earthquakes, in which the compressional axis is oriented in the dip direction of the slab, or downdip extension (DE) earthquakes, in which, on the contrary, the tensional axis is oriented in the dip direction of the slab. There is a systematic tendency in the depth distribution of focal mechanisms that depends on the age of the subducting plate. That is, as indicated in Figure 4: (1) DE-type earthquakes dominate in young plate subduction zones such as in Central America, in which intraslab earthquakes occur only to a depth of about 250 km. (2) Intraslab earthquakes occur to a depth of about 650 km in old plate subduction zones such as Tonga and the Izu-Bonin arc, and the DC-type focal mechanism dominates at all depths in those areas. (3) Subduction zones such as the Kermadec and Kuriles areas have characteristics intermediate between (1) and (2); DE-type earthquakes dominate in shallow reaches while DC-type earthquakes dominate deeper. (4) In subduction zones like Chile and the New Hebrides areas, in which intraslab earthquakes do not occur at depths between 300 and 500 km, the DE-type dominates at shallow depths while the DC-type dominates deeper.

Isacks and Molnar (1971) explain the formation of these earthquake-generating stress fields in the slab as follows. The depths to which the downdip ends of slabs reach constitute four different conditions (Figure 5). In one configuration, the end of the slab reaches the top of the high strength lower mantle (Figure 5c), which corresponds to the conditions described in (2) above. In two other configurations, it ends within the upper mantle (Figures 5a and b) corresponding to (1) and (3) above. In the fourth case, the slab is not continuous, but is divided into upper and lower segments (Figure 5d) corresponding to (4) above. When slabs subduct into the mantle, the resistive force becomes



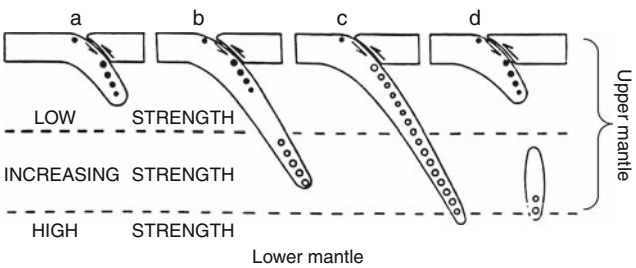
Seismicity, Subduction Zone, Figure 3 Focal depth distribution of global seismicity. Earthquakes with magnitudes greater than 4, 5, and 6 in the EHB catalog (Engdahl et al., 1998) for the period from 1964 to 2004 are plotted.



Seismicity, Subduction Zone, Figure 4 Global summary of stress states within subducting slabs (Isacks and Molnar, 1971, copyright by the American Geophysical Union). Open and solid circles show downdip compression and downdip extension, respectively. Crosses indicate that neither the compression nor tension axis is approximately parallel to the slab dip.

stronger with increasing depth, and the earthquake-generating stress fields discussed above are formed. Although the above explanation by Isacks and Molnar seems plausible at first glance, recent seismic tomographic studies are now revealing that the end of the slabs reach the top of the lower mantle in most subduction zones. Therefore, the interpretation provided in Figure 5 is not always correct.

Slab mineralogy studies have shown that depth perturbations of phase transformation boundaries resulting from thermal anomalies within the slabs also generate positive and negative buoyancy forces depending on depth, and so are major contributors to stress fields in the slabs (e.g., Bina et al, 2001). Figure 6 shows a diagram of the mineralogy in a subducting slab. Since the slab is colder than the surrounding mantle, the α-olivine to β-spinel phase transformation at 410 km depth and the β-spinel to γ-spinel phase transformation at 550 km depth become shallower, whereas decomposition of γ-spinel to perovskite + magnesiowustite at 660 km depth becomes deeper

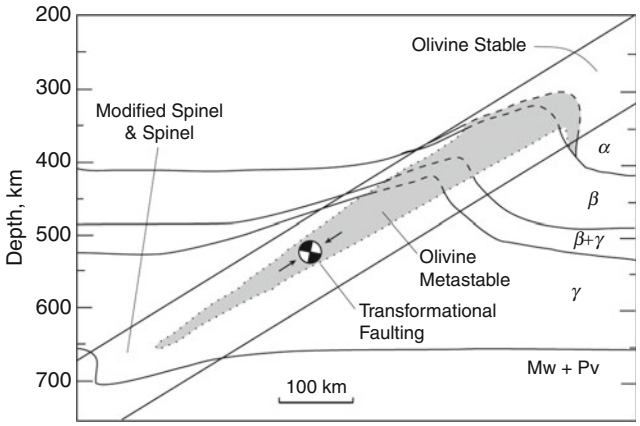


Seismicity, Subduction Zone, Figure 5 Sketch of a model explaining the stress state within the subducting slab where gravitational forces act on excess mass within the slab (Isacks and Molnar, 1971, copyright by the American Geophysical Union).

than in the ambient mantle. Moreover, the olivine to β-spinel transition must be kinetically hindered within the cold slab, since the relatively colder temperatures in the slab inhibit reaction rates. A wedge-shaped zone of

metastable peridotite thus formed in the cold core of the slab persists to the depth shown by the shaded area in Figure 6 (Kirby, 1995).

Upward deflections of the olivine to β-spinel transition at 410 km and the β-spinel to γ-spinel transition at 550 km yield negative buoyancy forces by stabilizing the dense higher pressure phases in the slab surrounded by the less dense lower pressure phases in the ambient mantle. On the contrary, the downward deflection of the γ-spinel to perovskite + magnesiowustite transition at about 660 km yields positive buoyancy forces. Moreover, the metastable olivine wedge yields positive buoyancy forces. Downdip compression at depths deeper than about 300 km, observed for all the subduction zones (Figure 4), can be well explained by the compressional stress generated by the combination of the buoyancy forces produced by the metastable olivine wedge, the downward deflection of the perovskite-forming transition, and the force resisting slab penetration into the lower mantle (Chen et al., 2004). It is expected that the downdip compressional stress caused by the metastable olivine wedge is smaller for warm (young) and slow slabs. Chen et al. (2004) found a good correlation between the thermal parameter of the slab and the nature of the downdip stress at depths shallower than about 300 km in accordance with this expectation: warm, slow slabs tend to have tensional stress; and cold, fast slabs tend to have compressional stress.

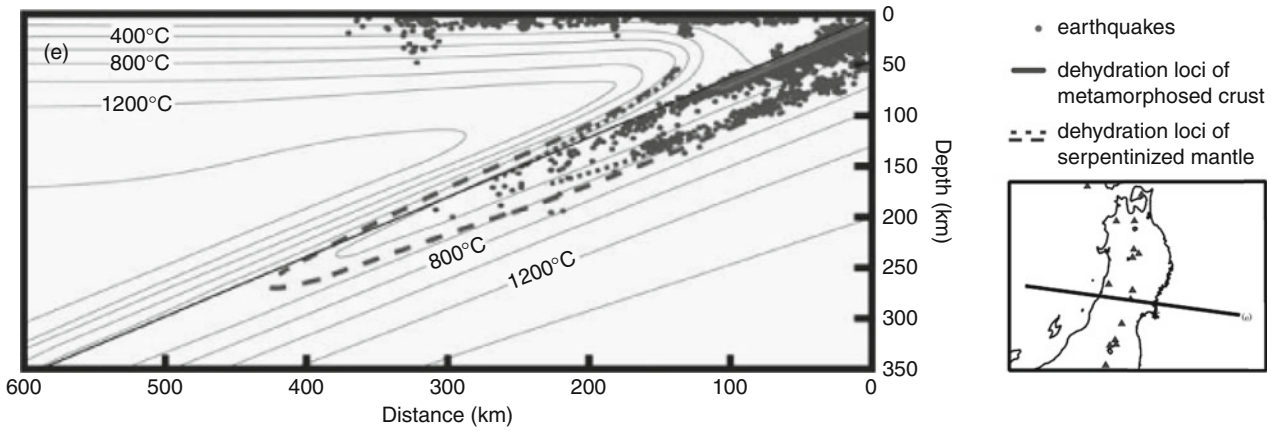


Seismicity, Subduction Zone, Figure 6 Schematic diagram of the mineralogy within a subducting slab (Kirby, 1995, copyright by the American Geophysical Union).

Intermediate-depth earthquakes: double seismic zone

It is known that intermediate-depth intraslab earthquakes form a double seismic zone in several subduction zones. This was first clearly found beneath northern Honshu, in northeast Japan (Figure 7), where the deep seismic zone is composed of two planes at depths of 70–150 km that are parallel to each other and separated by about 30 km (Hasegawa et al., 1978). Upper-plane earthquakes, most of which occur in the crust of the slab, have DC-type focal mechanisms, while lower-plane ones, occurring in the mantle of the slab, are characterized as DE-type earthquakes. Engdahl and Scholz (1977) explained the formation of such a double-planed deep seismic zone by the unbending of the subducting oceanic plate at this depth range. Later studies showed, however, that DC-type upper-plane and DE-type lower-plane events are not always seen and that in some subduction zones both the upper and lower planes have DC stress. These observations indicate that the unbending force may not be the main cause of the formation of the double seismic zone, although it must be acting on the plate.

Similar to ordinary shallow earthquakes, most intermediate-depth and deep earthquakes have double-couple radiation patterns and little or no isotropic



Seismicity, Subduction Zone, Figure 7 Double seismic zone and dehydration loci of oceanic crust and mantle for central Tohoku, northeast Japan, projected on an across-arc vertical cross section (Yamasaki and Seno, 2003, copyright by the American Geophysical Union). The solid line and broken and dotted lines show the dehydration loci of the crust and mantle, respectively. Dots denote earthquake hypocenters. Medium thick and thin lines are the upper plate interface and isotherm contours, respectively.

component (Frohlich, 2006). However, it has been thought that intermediate-depth and deep earthquakes are mechanically different from shallow earthquakes that are generated by brittle fracture. This is because normal stress on the fault becomes too high to cause ordinary brittle shear faulting. At depths of 60–650 km, the overburden produces confining pressures of 2–23 GPa, yielding the strength of the fault of about the same order of magnitude. Earthquake occurrence requires a shear stress exceeding that strength, but it is unlikely that such a high shear stress is actually produced at such depths in the slab. Even if some process in the slab can produce such a high shear stress, it cannot be maintained because ductile flow will occur. Therefore, some special mechanism is necessary to generate earthquakes at such great depths.

One possible mechanism that decreases the strength of the fault for intermediate-depth earthquakes is dehydration embrittlement (Raleigh and Paterson, 1965; Green and Houston, 1995). Before its subduction, the oceanic plate contains a considerable amount of water as hydrated minerals. The temperature and pressure of the plate rise as it subducts and decompose the hydrated minerals and discharge water. The water reduces the effective normal stress and enables brittle fracture. Dehydration embrittlement also seems to explain the formation of the double seismic zone as discussed below.

If dehydration embrittlement is the mechanism generating intermediate-depth intraslab earthquakes, intraslab earthquakes should not be expected to occur everywhere in the slab, but only where hydrated minerals exist, particularly at facies boundaries where the water content changes. It has become increasingly clear that intermediate-depth intraslab earthquakes do seem to occur in such areas.

In order to examine whether or not dehydration embrittlement is the cause of intraslab earthquakes, we need to know the location within the slab where the hydrated minerals are decomposed and the water is discharged. Yamasaki and Seno (2003) determined the locations in the slab where the dehydration decompositions of serpentinized slab mantle and metamorphosed slab crust occur based on experimentally obtained phase diagrams of rocks. They compared these locations with the spatial distribution of intraslab earthquakes. The result for the Pacific slab under northeast Japan is shown in Figure 7. The estimated dehydration loci of the crust and mantle roughly correspond to the location of the upper and lower planes of the double seismic zone. They obtained the same results for the other five subduction zones. This explains well why double seismic zones are formed and why earthquakes in the lower plane have a planar distribution near the center of the slab mantle.

The above results support the dehydration embrittlement model of intraslab earthquakes. However, whether hydrated minerals actually exist at depths where lower-plane earthquakes occur (the deeper part of the mantle that lies at a maximum of about 40 km below the upper slab boundary) is still a point of contention. In this regard, further validation is necessary for the generation of lower-plane events.

Formation of a belt of intraslab seismicity in the slab crust beneath northeast Japan

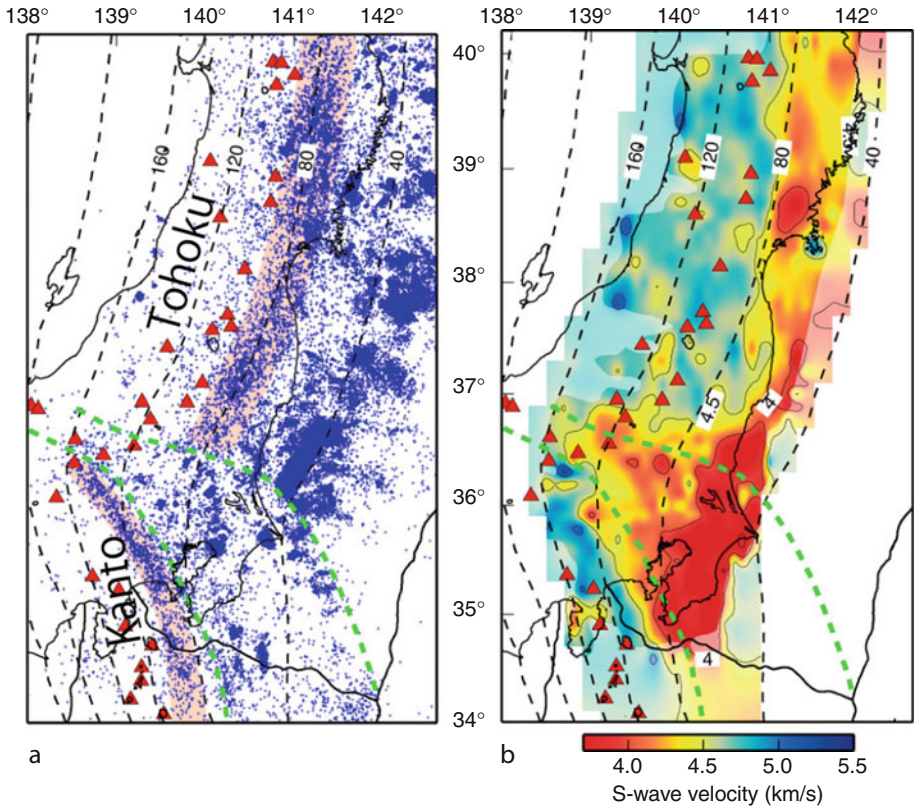
Recent studies have shown a more detailed spatial distribution of upper-plane earthquakes and revealed its clear relationship with the slab structure. This, as described below, can be explained well if the cause of intraslab earthquakes is dehydration embrittlement.

As indicated by the pale-pinked zones in Figure 8a, earthquakes in the upper plane of the double seismic zone beneath northeast Japan form a remarkable belt-like seismic zone at depths between 70 and 90 km, parallel to iso-depth contours of the upper surface of the Pacific slab. The across-arc vertical cross section of earthquakes in Figure 9 also shows that there is a concentration of earthquakes at depths of 70–90 km, which corresponds to the belt-like seismicity indicated in Figure 8a. If the cause of intraslab earthquakes is dehydration embrittlement, this concentration of seismicity is expected to correspond to the location in the slab crust where hydrated minerals are present, particularly around the facies boundary where dehydration occurs.

This belt-like seismicity indicated by the pale-pinked zones is not parallel to iso-depth contours, but is oblique to them and deepens locally in Kanto (Figure 8a). This local deepening is thought to be caused by the shield effect of the Philippine Sea slab, which is located immediately above the Pacific slab. The two slabs are in direct contact under Kanto, and the contact zone between the two is indicated as an area enclosed by two broken lines in the figure. The downdip end of the contact zone is parallel to the belt-like seismicity beneath Kanto, which is oblique to the iso-depth contours of the upper plate interface. Such a clear correspondence indicates that under Kanto, the Philippine Sea slab prevents the Pacific slab from being heated up by the mantle wedge and delays the phase transformation in the slab crust, resulting in the locally deepened belt-like seismicity in that area (Hasegawa et al., 2007).

Phase transformation of the slab crust causes an increase in seismic wave velocity. Therefore, we can verify from seismic tomography whether the phase transformation is the cause of the belt-like seismicity. Distribution of S-wave velocity in the slab crust (Figure 8b) shows that the low-velocity layer persists down to about 80 km depth in Tohoku; deeper than that, S-wave velocity becomes high. On the other hand, in Kanto, the depth range of the low-velocity layer deepens locally in the zone of contact with the Philippine Sea slab. This means the distribution of S-wave velocity also confirms the prediction that the phase transformation is delayed.

Direct comparison with mineralogical research results does not allow us to verify whether the phase transformation accompanying dehydration actually occurs at the depth of this belt-like seismicity. Though accurate



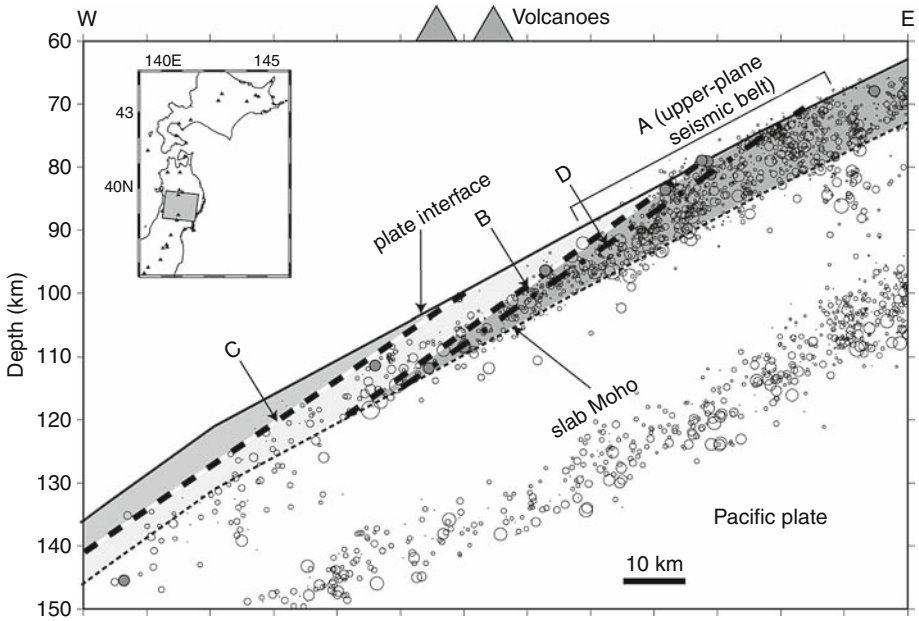
Seismicity, Subduction Zone, Figure 8 (**a**) Earthquake epicenter distribution and (**b**) S-wave velocity distribution in the subducting Pacific slab crust beneath northeast Japan (Nakajima et al., 2009, copyright by the American Geophysical Union). Pale-pinked zones in (**a**) show the belt-like seismic zone. S-wave velocity perturbations on a curved plane 5 km below the upper surface of the slab are shown in (**b**) by the color scale at the bottom. Iso-depth contours of the upper surface of the slab are shown by broken curves. The zone of contact with the overlying Philippine Sea slab is enclosed by green broken lines.

temperature distribution in the slab is necessary for such a detailed examination, unfortunately, many uncertainties remain regarding the estimation of the temperature in the slab. However, it is possible to make at least a rough estimate. In the vertical cross section in Figure 9, the facies boundaries estimated based on the phase diagram of midoceanic ridge basalt (MORB) by Hacker et al. (2003) and the geotherm by Peacock and Wang (1999) are shown as broken lines. The concentration of earthquakes forming the belt-like seismicity is found around the facies boundary in the shallower areas (B in the figure). That approximately agrees with the prediction based on the dehydration embrittlement model. As another example, the facies boundary determined by using the MORB phase diagram by Omori et al. (2009) and the subducted slab geotherm by van Keken et al. (2002) are indicated as a dashed-dotted line (D) in the figure. In this case also, the facies boundary approximately agrees with the location of the belt-like seismicity. Since many uncertainties remain in the estimation of the slab temperature, we need further validation for the location of the facies boundary. At any rate, the facies boundary must be strongly dependent not on pressure but on temperature.

Deep earthquakes

As described in previous sections, seismic observations seem to support the dehydration embrittlement hypothesis for the generation of intermediate-depth earthquakes, at least those in the crust of the slab. Focal depth distribution of global seismicity (Figure 3) shows that the frequency of earthquakes, after a nearly monotonic decrease from shallow depths, recovers at a depth of about 300 km and gradually increases until the local maximum at 500–600 km. This may suggest that a different mechanism is working for the generation of deep earthquakes.

Several hypotheses have been proposed for the generation of deep earthquakes. There exists the difficulty in generating deep earthquakes as already described in section "Intermediate-depth earthquakes: double seismic zone." Three mechanisms out of those proposed so far seem to overcome the difficulty. One is the shear heating



Seismicity, Subduction Zone, Figure 9 Across-arc vertical cross section of intraslab earthquakes in central Tohoku, northeast Japan (Kita et al., 2006, copyright by the American Geophysical Union). Earthquakes are shown by open circles. The estimated jadeite lawsonite blueschist–lawsonite amphibole eclogite facies boundary (B) and the lawsonite amphibole eclogite–eclogite facies boundary (C) of the MORB (Hacker et al., 2003) are denoted by thick broken lines. The facies boundary (D) based on Omori et al. (2009) is also shown by a dashed-dotted line.

instability of creep (Griggs and Handin, 1960; Ogawa, 1987). Plastic creep of minerals produces heat, while high temperatures promote creep. Therefore, positive feedback promoting creep would arise if the creep of a thin layer occurs at a higher rate than the heat produced by the creep diffuses into the surrounding area. Thus, the creep accelerates, and finally the thin layer of material begins to melt, producing dynamic slip on the layer, that is, occurrence of an earthquake.

The second hypothesis is the transformational faulting (Kirby et al., 1987; Green and Burnley, 1989). Olivine in the mantle of the slab transforms to β-spinel associated with the subduction of the slab. The phase transformation does not occur simultaneously but originates sporadically here and there, forming small lens-shaped spinel phases. The phase transformation proceeds as these spinel lenses expand. The spinel lenses will be formed in a direction perpendicular to the maximum compressive stress axis due to volume reduction occurring in the phase transformation. At the same time, compressive stress will arise at the tip of each spinel lens due to volume reduction as the phase transformation proceeds. This compressive stress promotes the phase transformation. In this way, the spinel lens expands as if it were a crack. Green and Burnley called this spinel lens anticrack. At some critical anticrack density, those anticracks that are aligned on a plane directed obliquely to the maximum compressive axis at an acute angle are connected with each other and become unified, producing a dynamic slip on the plane, that is, the occurrence of an earthquake.

The last one is the dehydration embrittlement. Although this hypothesis seems to work for the generation of intermediate-depth earthquakes as already explained, some researchers argue that the dehydration reaction continues to occur in the mantle of the slab at greater depths and that dehydration embrittlement is responsible also for deep earthquakes (Omori et al., 2004).

It is not clear which of the three mechanisms is actually working, or if another completely different mechanism is responsible for generating deep earthquakes. More investigations are required for an exact understanding of the mechanism by which deep earthquakes are generated.

Summary

Recent investigations have significantly developed our understanding of generation mechanism for earthquakes in subduction zones. Asperity model seems to be applicable to interplate earthquakes. This gives us a theoretical background for long-term earthquake forecast: Place and magnitude of future earthquakes can be estimated from information on locations and sizes of asperities on the plate boundary. Dehydration embrittlement model seems to work as the generation mechanism for intraslab intermediate-depth earthquakes, particularly for earthquakes in the slab crust. Aqueous fluids expelled by the

dehydration reaction in the subducted plate migrate upward, finally reaching the arc crust. Recent investigations have shown that the fluids thus supplied originally from the subducting plate play an important role in generating shallow inland intraplate earthquakes. Generation mechanism for deep earthquakes is still an open question. More systematic and intensified studies are expected to resolve it.

Bibliography

Bilek, S. L., and Lay, T., 2002. Tsunami earthquakes possibly widespread manifestations of frictional conditional stability. *Geophysical Research Letters*, **29**(14), 1673, doi:10.1029/2002GL015215.

Bina, C. R., Stein, S., Marton, F. C., and Van Ark, E. M., 2001. Implications of slab mineralogy for subduction dynamics. *Physics of the Earth and Planetary Interiors*, **127**, 51–66.

Chen, Po-Fei, Bina, C. R., and Okal, E. A., 2004. A global survey of stress orientations in subducting slabs as revealed by intermediate-depth earthquakes. *Geophysical Journal International*, **159**, 721–733.

Dixon, T. H., and Moore, J. C., 2007. *The Seismogenic Zone of Subduction Thrust Faults, Margins Theoretical Institute and Experimental Earth Science Series*. New York: Columbia University Press.

Engdahl, E. R., and Scholz, C. H., 1977. A double Benioff zone beneath the central Aleutians: an unbending of the lithosphere. *Geophysical Research Letters*, **4**, 473–476.

Engdahl, E. R., van der Hilst, R., and Buland, R., 1998. Global teleseismic earthquake relocation with improved travel times and procedures for depth determination. *Bulletin of the Seismological Society of America*, **88**, 722–743.

Frohlich, C., 2006. *Deep Earthquakes*. New York: Cambridge University Press.

Green, H. W., and Burnley, P. C., 1989. A new self-organizing mechanism for deep-focus earthquakes. *Nature*, **341**, 733–737, doi:10.1038/341733a0.

Green, H. W., and Houston, H., 1995. The mechanics of deep earthquakes. *Annual Review of Earth and Planetary Sciences*, **23**, 169–213.

Griggs, D. T., and Handin, J., 1960. Observations on fracture and a hypothesis of earthquakes. In Griggs, D. T., and Handin, J. (eds.), *Rock Deformation: Geological Society of America. Memoirs, 79*. New York: Geological Society of America, pp. 347–373.

Hacker, B. R., Peacock, S. M., Abers, G. A., and Holloway, S. D., 2003. Subduction factory 2. Are intermediate-depth earthquakes in subducting slabs linked to metamorphic dehydration reactions? *Journal of Geophysical Research*, **108**(B1), 2030, doi:10.1029/2001JB001129.

Hasegawa, A., Umino, N., and Takagi, A., 1978. Double-planed structure of the deep seismic zone in the northeastern Japan arc. *Tectonophysics*, **47**, 43–58.

Hasegawa, A., Nakajima, J., Umino, N., and Miura, S., 2005. Deep structure of the northeastern Japan arc and its implications for crustal deformationand shallow seismic activity. *Tectonophysics*, **403**(1–4), 59–75.

Hasegawa, A., Nakajima, J., Kita, S., Okada, T., Matsuzawa, T., and Kirby, S., 2007. Anomalous deepening of a belt of intraslab earthquakes in the pacific slab crust under Kanto, central Japan: possible anomalous thermal shielding, dehydration reactions, and seismicity caused by shallower cold slab material. *Geophysical Research Letters*, **34**, L09305, doi:10.1029/2007GL029616.

Hyndman, R. D., Wang, K., et al., 1995. Thermal constraints on the seismogenic portion of the southwestern Japan subduction thrust. *Journal of Geophysical Research*, **100**, 15373–15392.

Iio, Y., Sagiya, T., and Kobayashi, Y., 2004. Origin of the concentrated deformation zone in the Japanese Islands and stress accumulation process of intraplate earthquakes. *Earth Plants Space*, **56**, 831–842.

Isacks, B. L., and Molnar, P., 1971. Distribution of stresses in the descending lithosphere from a global survey of focal mechanism solutions of mantle earthquakes. *Reviews of Geophysics and Space Physics*, **9**, 103–174.

Kanamori, H., 1977. Seismic and aseismic slip along subduction zones and their tectonic implications. In Ewing, M. (ed.), *Island Arcs, Deep Sea Trenches and Back-Arc Basins*. Washington, DC: American Geophysical Union, Vol. I, pp. 163–174.

Kirby, S., 1987. Localized polymorphic phase transformations in high-pressure faults and applications to the physical mechanism of deep earthquakes. *Journal of Geophysical Research*, **92**(B13), 789–800, doi:10.1029/JB092iB13p13789.

Kirby, S., 1995. Intraslab earthquakes and phase changes in subducting lithosphere. *Reviews of Geophysics*, **33**(S1), 287–297.

Kita, S., Okada, T., Nakajima, J., Matsuzawa, T., and Hasegawa, A., 2006. Existence of a seismic belt in the upper plane of the double seismic zone extending in the along-arc direction at depths of 70–100 km beneath NE Japan. *Geophysical Research Letters*, **33**, L24310, doi:10.1029/2006GL028239.

Lay, T., and Kanamori, H., 1981. Fundamental studies, laboratory investigations and models. In Ewing, M., Simpson, D. W., and Richards, P. G. (eds.), *An International Review, Earthquake Prediction*. Washington, DC: AGU, pp. 579–592.

Miura, M., Sato, T., Hasegawa, A., Suwa, Y., Tachibana, K., and Yui, S., 2004. Strain concentration zone along the volcanic front derived by GPS observations in NE Japan arc. *Earth Planets Space*, **56**, 1347–355.

Nakajima, J., and Hasegawa, A., 2007. Deep crustal structure along the Niigata-Kobe tectonic zone, Japan: its origin and segmentation. *Earth Planets and Space*, **59**, e5–e8.

Nakajima, J., Matsuzawa, T., Hasegawa, A., and Zhao, D., 2001. Three-dimensional structure of Vp, Vs, and Vp/Vs beneath northeastern Japan: implications for arc magmatism and fluids. *Journal of Geophysical Research*, **106**, 843–857.

Nakajima, J., Tsuji, Y., and Hasegawa, A., 2009. Seismic evidence for thermally-controlled dehydration reaction in subducting oceanic crust. *Geophysical Research Letters*, **36**, L03303, doi:10.1029/2008GL036865.

Ogawa, M., 1987. Shear instability in a viscoelastic material as the cause of deep focus earthquakes. *Journal of Geophysical Research*, **92**, 13801–13810.

Oleskevich, D. A., Hyndman, R. D., and Wang, K., 1999. The updip and downdip limits to great subduction earthquakes: thermal and structural models of Cascadia, south Alaska, SW Japan, and Chile. *Journal of Geophysical Research*, **104**, 14965–14991.

Omori, S., Komabayashi, T., and Maruyama, S., 2004. Dehydration and earthquakes in the subducting slab: empirical link in intermediate and deep seismic zones. *Physics of the Earth and Planetary Interiors*, **146**, 297–311.

Omori, S., Kita, S., Maruyama, S., and Santosh, M., 2009. Pressure – temperature conditions of ongoing regional metamorphism beneath the Japanese islands. *Gondwana Research*, **16**, 458–469.

Peacock, S. M., and Wang, K., 1999. Seismic consequences of warm versus cool subduction metamorphism: examples from southwest and northeast Japan. *Science*, **286**, 937–939.

Raleigh, C. B., and Paterson, M. S., 1965. Experimental deformation of serpentinite and its tectonic implications. *Journal of Geophysical Research*, **70**, 3965–3985.

Ruff, L. J., and Kanamori, H., 1980. Seismicity and the subduction process. *Physics of the Earth and Planetary Interiors*, **23**, 240–252.

Sagiya, T., Miyazaki, S., and Tada, T., 2000. Continuous GPS array and presentday crustal deformation of Japan. *Pure and Applied Geophysics*, **157**, 2303–2322.
Scholz, C. H., 1998. Earthquakes and friction laws. *Nature*, **391**, 37–42.
van Keken, P. E., Kiefer, B., and Peacock, S. M., 2002. High-resolution models of subduction zones: implications for mineral dehydration reactions and the transport of water into the deep mantle. *Geochemistry, Geophysics, Geosystems*, **3**, 10, doi:10.1029/2001GC000256.
Yamasaki, T., and Seno, T., 2003. Double seismic zone and dehydration embrittlement of the subducting slab. *Journal of Geophysical Research*, **108**(B4), ESE9–1, doi:10.1029/2002JB001918.

Cross-references

Earthquake, Focal Mechanism
Great Earthquakes
Seismicity, Intraplate

SEISMOGRAM INTERPRETATION

Ota Kulhanek[1], Leif Persson[2]
[1]Department of Earth Sciences, Section of Seismology, Uppsala University, Uppsala, Sweden
[2]Department of Underwater Research, FOI, Stockholm, Sweden

Synonyms

Anatomy of seismograms; Reading seismic records; Structure of seismograms

Definition

Seismogram interpretation. Identification of various phases appearing on seismograms (measurable changes in frequency and/or amplitude), including the recognition of the waves with respect to possible travel paths through the Earth.

Introduction

A seismogram is the recording of earth motion (displacement, velocity, or acceleration) as a function of time. Its appearance reflects the combined effects of the source, the propagation path, the characteristics of the recording instrument (see *Seismic Instrumentation*), and the ambient noise due to the specific conditions at the particular recording site. To decipher the often complicated form of seismogram traces requires an extensive knowledge of source physics, structure of the Earth, wave propagation, and earthquake geography. In the early years of observational seismology, say, prior to 1960, the analyst was essentially reading short- and long-period analog seismograms made on smoked, photographic, or heat-sensitive paper or by ink-pen recorders. Approximately, in the middle of the 1980s, the situation changed dramatically. The advent and extensive deployment of broadband digital seismometers with large dynamic ranges complemented with an access to fast, powerful computers opened for seismogram interpreters new possibilities never expected before. Today, the interpreter performs easily a set of operations that significantly improve and accelerate the process of phase identification on digital seismic records. These include, for example, rotation of seismogram components and particle motion techniques. Digital multichannel data serve as input for frequency-wave number analysis and polarization filtering. Synthetic seismograms model observations and phase onsets revealed on seismograms can be associated within narrow time windows with theoretically predicted onset arrival times.

Velocity models for radially stratified Earth have been developed to provide travel times for major seismic waves for the purpose of earthquake location and phase identification. In 1987, IASPEI initiated an effort to update the JB Tables (Jeffreys and Bullen, 1940), which have served for 50 years, by constructing new global travel-time tables. With access to an extensive ISC data set (1964–1987) this effort resulted in two velocity models: *iasp91* (Kennett and Engdahl, 1991) and *sp6* (Morelli and Dziewonski, 1993) deduced from empirical smoothed travel-time curves of the main phases. The two models reveal only small differences in predicted travel times. With respect to the older JB Tables, the most significant differences were found in the upper mantle and core (see *Earth's Structure, Global*). New, *iasp91*, travel-time tables, which include also the ancillary phases, were calculated from the velocity model (Kennett, 1991). For teleseismic *P* waves, the new tables are in average 1.8–1.9 s faster than the JB Tables. The *iasp91* model was slightly modified in 1995 and a new model *ak135* was proposed (Kennett et al., 1995). The model is suited for predicting the arrival times of a wide variety of seismic phases. It is routinely used by major agencies (ISC, NEIC) in event locations and phase identifications.

At first sight, it seems that there is a contradiction between the duration of the rupture at the source, which takes between a fraction of a second and a few minutes (depending on magnitude), and the length of the observed seismogram, which for large and distant shocks can extend over several hours. The length of the seismogram depends primarily on various wave propagation effects such as reflection, refraction, mode conversion, dispersion, etc., and has little to do with the duration of the quake. A seismogram, especially from a distant earthquake, will often show a number of more or less distinct waves, commonly called phases, distributed in time, which have traveled along different propagation paths and, which were subjected to different mode conversions. Numerous manuals for seismogram interpretation have been available since the early 1950s. Some of them have been of local/regional importance, but some have gained worldwide recognition. The latter include works of Neumann (1951), Simon (1968), Willmore (1979), Payo (1986), and Kulhanek (1990). The most recent is the manual edited by Borman (2002).

Essentially, there are two types of seismic waves, *P* and *S* body waves (see *Body Waves*) and Love (*LQ*) and Rayleigh (*LR*) surface waves (see *Surface Waves*). The most important characteristics of body and surface waves, which are invaluable in any seismogram interpretation, are the following:

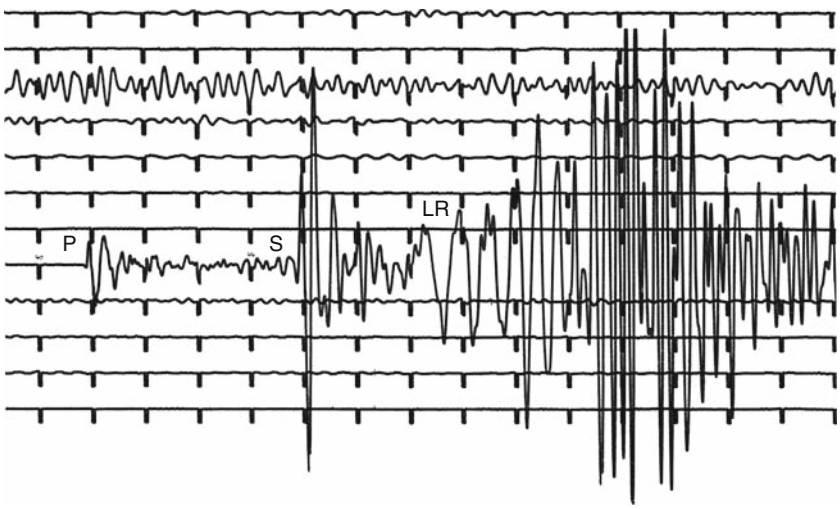
- Different waves travel with different velocities. At any epicentral distance, *P* is recorded first, followed by *S*, *LQ*, and *LR*.
- Body waves produce records in the high-frequency range, from 0.1 to 10 Hz, while surface waves occupy low frequencies of about 0.1 Hz and lower. Exceptions will be described later.
- Due to different polarization, *P*-waves are usually better displayed on vertical-component seismograms while *S*-waves are often best on horizontal-component records. Fluids do not sustain shear strain and therefore *S*-waves do not travel through liquid parts of Earth's interior. *LQ*-waves are exhibited only on horizontal-component records, while *LR*-waves are received by both the vertical and horizontal seismometers. For shallow shocks, surface waves usually dominate the seismogram.
- Surface waves exhibit an important property called velocity dispersion (see *Surface Waves*). In practice, this means that the long period surface waves approach the station first and are recorded ahead of the "slower" shorter waves. Ideally, the seismogram of *LQ*- or *LR*-waves will start with long-period motion, which gradually, as time increases, will turn into shorter and shorter periods.

Some of these features are demonstrated on the seismogram displayed in Figure 1. There is a sharp *P* onset followed after approximately 4 min by a clear *S*. About 2 min after *S* we observe the arriving *LR* (vertical-component seismogram). The clearly dispersive character of the recorded *LR* in this case manifests a continental propagation path signature. The large *LR* amplitudes indicate a shallow-focus depth. Except for the first *P* onset, all later arrivals are contaminated by codas of preceding phases so that, on the records, there is virtually no interval of quiescence between individual phases.

In the following description, we introduce several categories of seismic events. The classification is based on the distance between the event and the recording site, that is, on the epicentral distance, which in turn governs propagation paths along which seismic waves travel through the Earth's interior. Events recorded at distances shorter than about 1° will be called local events. Regional earthquakes are shocks recorded at distances between approximately 1° and 10°. For epicentral distances 0–10°, seismic waves propagate through the crust and/or along Moho. Accordingly, we call these waves, crustal waves. At recording distances 10–103°, the waves travel mainly through the mantle. For distances larger than about 103°, seismic waves enter the core or are diffracted along the core-mantle boundary (so-called core waves). Events observed at distances larger than 20° (or 30°) are collectively called teleseisms.

Crustal waves; recording distances 0–10°

In a one-layer continental crust, a source located within the crust radiates both *P* and *S* waves, which will be recorded along the Earth's surface. These waves are encoded as *Pg* and *Sg*. The subscript *g* refers to granitic layer. In continental earthquakes it has been often observed that *Sg* has the largest amplitude. The nomenclature of seismic phases used in the present article is that of the IASPEI Standard (Storchak et al., 2003 or *Seismic*

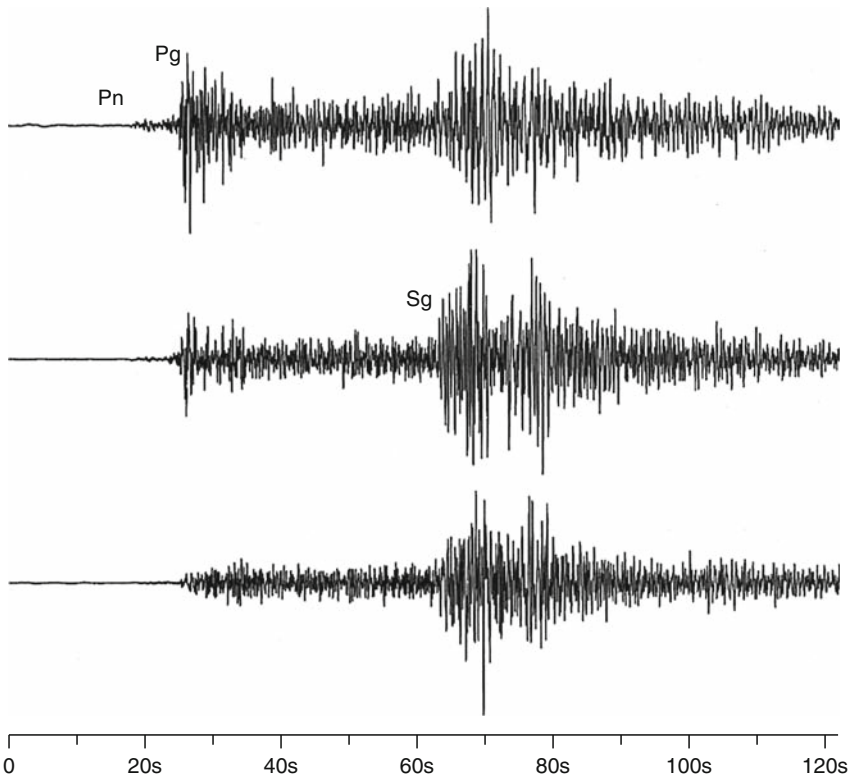


Seismogram Interpretation, Figure 1 Long-period, vertical-component, ink-pen seismogram of the Greek earthquake of February 25, 1981 (M = 6.3, h = 33 km) made at UPP (Uppsala), Sweden, at an epicentral distance of 22°. There is 1 min between successive time marks (small downward offsets).

Phase Names: IASPEI Standard). A reflected wave is also possible from the outer side of the Moho, and the corresponding reflected *P* and *S* waves are labeled as *PmP* and *SmS*, respectively. At larger distances (beyond the critical distance), we shall record so-called head waves, that is, *P* and *S* waves refracted along the Moho discontinuity and labeled *Pn* and *Sn*. Head waves propagate with velocities of the uppermost mantle. Note that *Pg* and *Sg* exist for all epicentral distances from Δ=0 and outwards, while *Pn* and *Sn* cannot be observed at distances shorter than the critical distance, which for a continental crust is about 100 km. Waves reflected from Moho are rather scarce and difficult to identify. The best chance to discern *PmP* or *SmS* on records is at very short epicentral distances where the contamination by *Pg* and *Pn* (or *Sg* and *Sn*) is not severe.

At epicentral distances shorter than about 10°, records are complicated because of lateral heterogeneities in the crust. In a two-layer crustal model, which includes the Conrad discontinuity, we record new phases. For epicentral distances larger than about 100 km, these are the refracted *P* and *S*, traveling along the Conrad discontinuity. An asterisk in the superscript position, P^* and S^*, indicates this phase. An alternative code is *Pb* and *Sb*. The subscript *b* refers to basaltic layer.

Let us assume a continental crust with propagation velocities for *Pg*, P^*, and *Pn* of 6.0, 6.6, and 8.0 km/s, respectively. At short distances, less than about 150 km or so, the first seismic wave arriving at the recording station is *Pg*. For distances larger than critical but less than about 150 km/s, *Pg* is followed by P^* and *Pn*, in this order. P^* and *Pn* travel with velocities significantly higher than *Pg*. Therefore, at distances larger than approximately 150–200 km crustal waves change their order of arrival. For distances larger than about 200 km, the first arriving phase is *Pn*, next arrives P^*, and then *Pg*. This is true only for continental travel paths. Seismograms from earthquakes beneath the sea bottom, made at islands or coastal stations, will not show *Pg* or *Sg* since there is no granitic layer in the oceanic crust. Similarly, quakes originating in the lower crust, beneath the Conrad discontinuity, do not produce *Pg* or *Sg* phases. Hence, first arrivals on records from these events will be *Pn* or P^*. With a certain time delay following the *P* phases, proportional to the epicentral distance, the crustal *S* waves arrive in the same order as *P* waves. For local events the order of *S* onsets will be *Sg*, S^*, *Sn* while for events from distances larger than about 300 km, we observe first *Sn* followed by S^* and *Sg*. As an example, a record from a regional earthquake is displayed in Figure 2. The first discernible phase



Seismogram Interpretation, Figure 2 The Kaliningrad, Russia, earthquake of September 21, 2004 (M = 5.0, h = 15 km) recorded at broadband station GOT (Gotland), Sweden, at an epicentral distance of 335 km. From top to bottom: vertical, N–S and E–W components. There are 10 s between successive time marks.

is *Pn*, weakly recorded on the vertical and N-S channels. It is followed after about 6 s by strong *Pg*, and after another 39 s by clear *Sg*, visible on all three components. *Sn* expected to precede *Sg* by 9 s cannot be identified on the seismograms.

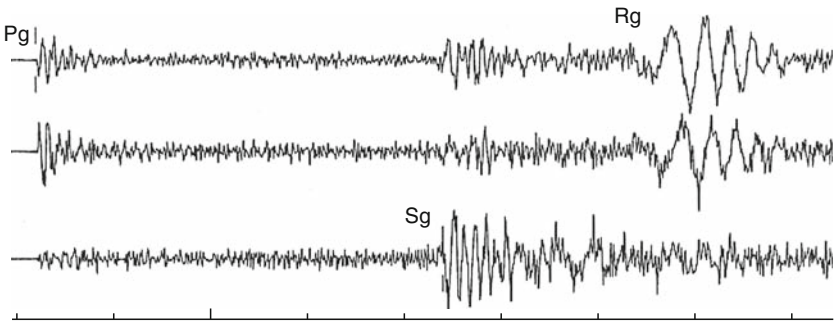
Short-period *S* waves, multiply reflected between the free surface and Moho or other crustal velocity discontinuities, interfere with each other and give rise to a wave group labeled *Lg*. There is no clear distinction between *Sg* and *Lg*. *Lg* waves propagate as guided waves, supercritically incident on the Moho and multiply reflected within the (continental) crust with a velocity of about 3.5 km/s. They may dominate the seismograms made usually at a distance of 5° and larger. *Lg* are best seen on horizontal-component records. Some rare cases are known where *Lg* propagated over distances of several thousand kilometers (Kulhanek, 1990).

Near-surface local and regional events (earthquakes, industrial explosions, rockbursts, etc.) also generate short-period surface wave of Rayleigh type, labeled *Rg*. The presence of *Rg* in the seismogram is a reliable indicator of a very shallow event with focal depth of the order of one or a few kilometers. On the other hand, if *Rg* waves are absent (epicentral distance of several hundred kilometers or less), we are concerned with a deeper natural event, that is, with a crustal quake at a depth most likely between 5 and 25 km, since all types of man-made events can be excluded. *Rg* waves travel with a velocity of 3 km/s or slightly higher and are usually well recorded up to distances of about 600 km (Båth, 1983). These waves are best, but not exclusively, seen on vertical-component records. An example of a recorded *Rg* wave is displayed in Figure 3. The *Rg* phase is well developed (large amplitudes, clear dispersion) indicating a focal depth of 2 km or less.

Island and coastal seismographic stations frequently record so-called *T* waves (tertiary waves) arriving after *P* and *S* (see *T Waves*). They are characterized by propagation within the oceans as ordinary sound waves emitted by earthquakes near the sea bottom or by submarine volcanic eruptions. The propagation of *T* waves is very efficient and observations at distances as large as about 80° have been reported. Observations of *T* phases have proved useful in discriminating between underground nuclear explosions, detonated beneath oceanic islands and tectonic earthquakes (Adams, 1979). On records, *T* phases often exhibit rather monochromatic oscillations (periods usually less than 1 s) with a gradual increase and decrease of amplitudes of total duration up to several minutes. There is no sharp onset in the *T* phase group, which creates difficulties when reading the *T* arrival times. In general, there is a great variety in the appearance of *T* phases due to the dependence upon the bottom topography in the vicinity of generation, oceanic stratification, and water-land conversion and transmission.

Appearance of seismograms made at local and/or regional distances varies from place to place mainly due to lateral variations in crustal structure. It is, therefore, difficult to list generally valid clues for record interpretation. Nevertheless, following principles may guide the analyst to read seismograms and evaluate software solutions of local and regional earthquakes.

- Predominant periods of recorded crustal waves *Pg*, P^*, *Pn*, *Sg*, S^*, *Sn*, etc., are normally less than 1 s. *Rg* periods are usually not longer than several seconds.
- Often, *Sg* has the largest amplitude (when large short-period *Rg* is missing), best on horizontal-component records. At larger distances, *Lg* may dominate the records.
- For epicentral distances less than about 200 km (depending upon the crustal structure and focal depth), the first arriving phase is *Pg*. For larger distances, *Pn* arrives first.
- Near-surface events, from distances less than about 600 km, often generate short-period *Rg* with clear dispersion, best seen on vertical channels.
- Local and regional events of low or moderate magnitude are characterized by short total record duration, usually not exceeding several minutes.
- Island and coastal seismographic stations frequently record various types of *T*-phases.



Seismogram Interpretation, Figure 3 Weak, near-surface event ($M \sim 2$) in southern Sweden, recorded at broadband station NYN (Nynäshamn) at an epicentral distance of 36 km. From top to bottom: vertical, N–S and E–W components. There are strong onsets of *Pg* (vertical, N–S) and *Sg* (E–W) and a clear *Rg*-phase with distinct dispersion, about 2 s after *Sg*, best developed on the vertical and N–S components. There is 1 s between successive time marks at the bottom. (Note: epicenter almost due north of NYN).

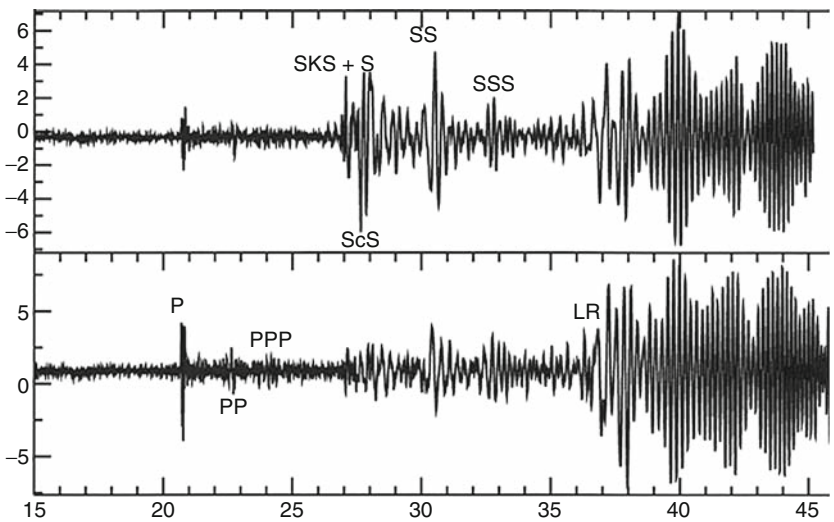
Body waves traversing the mantle; recording distances 10–103°

Mantle, in contrast to the overlying crust, may, in the first approximation, be considered as laterally homogeneous. At distances around 10°, *Pn* and *Sn* become difficult to identify in the records. Instead, *P* and *S* phases show on seismograms. Some workers consider the distance range between 10° and 103° as ideal to record not only the direct *P* and *S* waves, but also the whole family of reflected and converted waves. Travel paths of these waves are dominated by the mantle and corresponding seismograms are relatively simple. *P* is usually stronger on the vertical component, while *S* is more clearly seen on horizontal components. *S* often exhibits wave trains with longer periods when compared with corresponding *P*. Large-amplitude *S* waves are often observed at distances up to about 100°.

A direct *P* reflected from the free surface once or twice is called *PP* or *PPP*, respectively. In the same way we have also *SS* or *SSS*. For distances larger than about 40°, the free-surface reflected phases become very distinct. At distances around 100° and larger, *PP* and *SS* often belong to the largest recorded body waves. Converted waves like *PS* and *SP* appear only at distances larger than 40°. Examples of seismograms with free-surface reflections are displayed in Figure 4.

Core-reflected phases, that is, waves reflected back into the mantle from the outer core-mantle boundary are labeled *PcP* and *ScS*. Including mode converted waves, we have also *PcS* and *ScP*. Because core-reflected phases emerge steeply, *ScP* is usually stronger on vertical component records than *PcS*. Large core-reflected phases are usually recorded at shorter epicentral distances, say at 40° or less. At distances around 39°, *ScP* and *PcS* (surface foci) are often contaminated with the arrival of direct *S* and the phase separation is difficult. When the ray path of *PcP* grazes the outer core boundary, the combination of direct *P* and *PcP* is called *P* diffracted and labeled *Pdif* (older *Pdiff*).

Waves ascending from the focus to the free surface (in continents) or to the sea bottom (under oceans), where they are reflected back into the mantle, are commonly called depth phases. We can list the four possibilities of reflections near the epicenter, which are *pP*, *sP*, *pS*, and *sS*. The first case, for example, denotes the wave that traveled upward from the focus as *P* (short leg) and had been reflected back off the free surface again as *P* (long leg). Depth phases, primarily *pP*, are the most important phases routinely used in focal-depth estimations. The deeper the focus, the later is the *pP* phase in relation to direct *P*. Thus, accurately measured arrival-time differences, *pP-P*, are reliable indicators of the depth of the focus. Most teleseismic shallow events (depth of focus around 33 km) will reveal a *pP* 9–11 s after *P* (see, e.g., Herrin et al., 1968). In case of a deeper focus, it is sometimes possible to recognize several different reflections/conversions from the free surface, for example, *pPP*, *sPP*, *pPS*, *sPS*, *pSP*, etc. Interpretation of depth phases must be done with utmost care since, for example, *pP* from a deep event can easily be erroneously interpreted as *P* when the first arrival (*P*) is weak. Similar difficulty emerges for multiple shocks. Depending on focal orientation and other factors, *sP* may be stronger than *pP* and may be mistaken for it. Depth phases are sometimes stronger than the direct *P*, and may be the first readable phase. We usually require several records made at different epicentral distances to make a reliable identification of the depth phase. We can



Seismogram Interpretation, Figure 4 The Costa Rica earthquake of August 20, 1999, ($M = 6.9$, $h = 40$ km). Exhibited seismograms are long-period, E–W (*upper trace*) and vertical-component (*lower trace*), records made at KONO (Kongsberg), Norway. The epicentral distance is 84°. There are 100 s between two successive time marks at the bottom of the figure. For this distance range, *SKS* arrives ahead of *S*, but the separation is difficult. Both displayed channels are dominated by a fundamental *LR* wave, developed along a mainly oceanic propagation path.

also mention less common phases like *pmP* and *pwP* (first identified by Mendiguren, 1971), corresponding to Moho underside reflection and to water surface reflection, respectively (Schenk et al., 1989). A thin layer of low-velocity sediments at the sea bottom may have small impedance contrast. Hence, short-period records may show large *pwP* when compared with *pP* reflecting at the sea bottom (Engdahl et al., 1998). *pwP* can easily be misinterpreted as *pP* leading to an overestimation of the focal depth.

Bolt (1982) introduced symbols like *P*400*P* or *P*650*P* to indicate seismic waves reflected at the underside of a secondary discontinuity, in the upper mantle, at a depth of 400 and 650 km, respectively. These phases arrive at the recording station ahead of the expected (calculated) arrival time for the main *PP* phase and are interpreted as reflections from upper mantle discontinuities, that is, as *PdP*. When the early arrival time cannot be explained in terms of known discontinuities as *PdP*, we call these onsets early *PP* or precursors to *PP*.

Body waves traversing the core; recording distances 103° and beyond

Due to the diffraction around the core-mantle boundary (see *Earth's Structure, Global*), amplitudes of direct *P* waves decay dramatically at distances larger than 100°. The short-period *P* reappear consistently on records first at distances of about 140°. The distance range $103° < \Delta < 140°$ is called the shadow zone. On seismograms, *Pdif* waves usually show small amplitudes, emergent or gradual onsets and the energy shifts to longer periods. Long-period *Pdif* are sometimes observed out to distances of 160° or more. *S* waves are affected at the core-mantle boundary in a similar way.

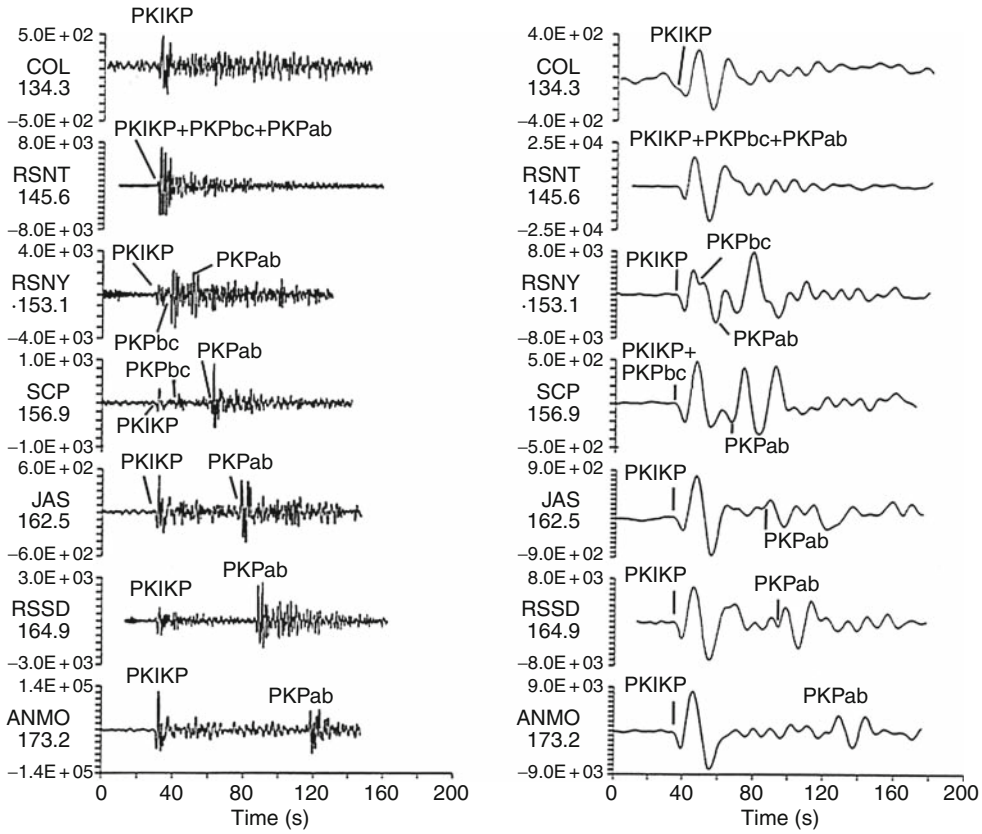
Waves that traveled through the outer core but do not enter the inner core are labeled *PKP* (also *P′*), *PKS*, *SKS* (also *S′*), and *SKP*. The *SKP* phase is stronger on vertical components than *PKS*. These phases have a caustic near 130° and close to this distance are often the only phase recorded on short-period seismograms. At epicentral distances close to 144°, *PKP* shows a distinct concentration of energy, that is, large amplitudes on records. The phenomenon may be viewed in terms of two *PKP* travel-time branches denoted *PKPbc* and *PKPab* for the first and the second arrival, respectively. At 144°, the waves from the two branches coincide, the waves reinforce one another, which results in energy concentration near that distance.

P waves that traverse the inner core are denoted by *I*, giving rise to phases *PKIKP*, *PKIKS*, *SKIKS*, and *SKIKP*, although these are often still simply referred to as *PKP*, *PKS*, etc. Phases, with an *S* leg in the inner core would include the letter *J*, such as *PKJKP*, but these have been difficult to identify on records. Cao et al. (2005) made use of high-quality, broadband seismograms from the Gräfenberg array station, Germany, and identified a phase with arrival time and slowness consistent with predictions of *PKJKP*.

If the studied event is weak, then usually no *Pdif* is observed in the entire distance interval $\Delta > 103°$ and the first arrival seen on the record will be that of *PKP*. At epicentral distances 105–120°, *PKIKP* usually provides the first onset discernible on the seismogram. In the region of the caustic, that is, around 144°, the wave train of recorded core phases becomes particularly complicated. It is first at distances beyond the caustic point where observed onsets may be separated into individual *PKP* branches. The energy distribution changes with the increasing distance. *PKPbc* is the dominant branch just beyond the caustic, up to about 153°. In records of weaker events (Δ = 144–153°), *PKPbc* is often the first visible onset since *PKIKP* (*alt. PKPdf*), theoretically preceding *PKPbc*, is too weak to be observed. As the distance increases, *PKPbc* becomes weaker and vanishes from records at distances of about 160° and larger. For distances beyond, say, 157°, *PKPab* usually dominates the seismogram. *PKIKP* in the distance range from about 125° to the caustic is often preceded by early arrivals or precursors, which can arrive many seconds ahead of the main phase. These are often explained by scattering phenomena at or near the core-mantle boundary. Subscripts *ab*, *bc*, and *df* are used in agreement with travel-time charts of Jeffreys and Bullen (1940). Recorded core phases made at a suite of seismographic stations at a distance range from 134° to 173° are shown in Figure 5.

We may form new symbols for the whole family of waves propagating through the outer core. For example, *PKKP* is a *P* wave reflected from the inside of the core-mantle boundary, often very pronounced on records made at distances between 60° and 80°. However, the striking onset may easily be misinterpreted as a first *P* arrival of another event. *P* waves trapped inside the Earth's liquid core and with multiple *K* legs are called *PNKP*, where $N-1$ is the number of reflections. Multiple reflections within Earth's outer core were first observed by Engdahl (1968). He used records from two deep earthquakes and an event in Novaya Zemlya ($h = 0$) and identified recorded seismic waves reflected as many as four times (i.e., *P5KP*) within the Earth's outer core. Cases like *P4KP* and *P7KP* have been reported by Bolt (1982). Waves that traverse the Earth's interior and are reflected at the outside or inside of the inner core are labeled *PKiKP* or *PKIIKP*, respectively.

The best chance to observe *PKPPKP*, or for short *P′P′*, is around distances $2 \times 144° = 288°$, or 72° if we take the shortest distance from source to station. *P′P′* is often well recorded, arriving about 30 min after *P*, when most of the coda amplitudes of preceding phases have already become faint. *P′P′* may in some cases be wrongly interpreted as a new *P* or *PKP*. 72° is also equivalent to $3 \times 144° = 432°$, so the phase *P′P′P′* is also strong at this distance, and may be observed, for strong earthquakes, about another 20 min after *P′P′*. In 1969, first observations of precursors to *P′P′*, sometimes called early *P′P′*, were made (Engdahl and Flinn, 1969). These were interpreted as *P′dP′* analogous to *PdP*, that is, as reflections of *PKP* at



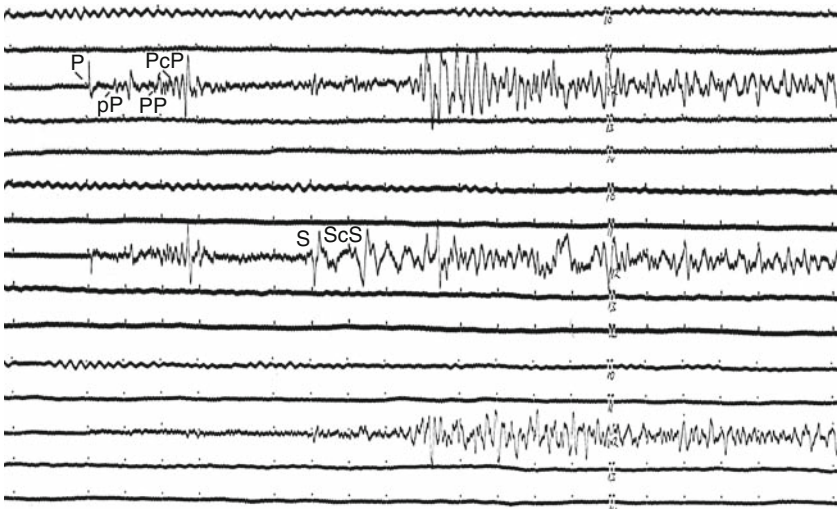
Seismogram Interpretation, Figure 5 Seismograms from a Mid-Indian-Rise earthquake of May 16, 1985 (M = 6.0, h = 10 km) made at a suite of seismographic stations that sample the epicentral distance between 134.3° and 173.2°. Short-period (*left*) and long-period (*right*), vertical-component records are exhibited. Station codes, epicentral distances, and amplitude scaling are given to the left of each trace (G. Choy, personal communication; reproduced from Kulhanek, 1990, with permission from Elsevier, Science).

secondary velocity discontinuities within Earth's upper mantle. For example, $P'650P'$ passes through the core twice and on seismograms precedes $P'P'$ by about 2 min.

Similar to *PKP* and *PNKP*, there are *SKS* and *SNKS* waves, respectively. First *SKS* waves are observed at distances between 60° and 70° and the range of observations extends to distances of 180° or so. *SKS* exhibits a caustic point at a distance of about 80° so that the best region to study *SKS* waves is that between 70° and 90°. The phase identification has to be made with utmost care since *SKS* recorded at this distance range are often contaminated with direct *S*. To mistake *S* for *SKS* and vice versa will adversely affect the epicentral location. At about 82° (depending upon details in the structural model), *SKS* begins to arrive ahead of *S* (Figure 4). For distances shorter than about 95°, *SKS* is usually smaller than *S*, however, at distances beyond 95°, *SKS* amplitudes are quite large. *S* and *SKS* are best recorded on long-period horizontal-component seismograms. The period of *SKS* phases may reach several tens of seconds. Occasionally, these body waves are also seen on short-period records, although the onset time of the later of the two phases is usually very emergent due to the contamination by the coda of the earlier phase.

Body waves from intermediate-focus and deep-focus earthquakes

Later studies confirmed conclusions from Wadati's pioneering work (Wadati, 1927) that intermediate-focus and deep-focus shocks produce simpler seismograms with exceptionally well recorded impulsive body waves while surface-wave amplitudes decrease as the quake becomes deeper. Strong depth phases, such as *pP* and *sS*, are also frequently very distinct on records from deep events. However, the duplication of principle phases by surface reflection (e.g., confusion between *pP* and *PP*) often complicates the seismogram interpretation. Weak precursors, *pmP*, (Moho underside reflections) to surface reflections, *pP*, are discussed in Schenk et al. (1989). They are best observed in long-period records and in continents often produce clear arrivals. In a more retrospective-type interpretation, the absence of aftershocks (see *Earthquake, Aftershocks*) will support the classification of the shock as



Seismogram Interpretation, Figure 6 Seismograms from an intermediate-focus earthquake (M = 6, h = 199 km) in the Hindu-Kush region on September 4, 1993, made at UPP, Sweden, at an epicentral distance of 41°. The traces show long-period channels from top to bottom: vertical, E–W and N–S components.

a deep-focus or intermediate-focus event. Figure 6 shows seismograms from an intermediate-focus earthquake made at an epicentral distance of 41°. Note the impulsive character of the *P*, *pP*, *PcP*, *S*, and *ScS* phases; virtually no surface waves have been recorded from this event.

Surface waves

With exception of very short epicentral distances, long-period surface waves, *LQ*, and *LR* dominate the seismograms of shallow events. Deep events of the same magnitude will generate abnormally small (insignificant) surface waves. This feature provides the interpreter with a viable tool to discriminate, at first glance, between shallow and deep-focus earthquakes. (Note a paradox: underground nuclear explosions, being very shallow events, generate only small, sometimes invisible, surface waves). The analyst has several clues to identify surface waves on the record and to distinguish *LQ* from *LR*. While Love waves are best displayed on horizontal-component seismograms, Rayleigh waves are best seen on vertical-component records. Both *LQ* and *LR* propagate slower than *P* or *S*, but since *LQ* propagate faster than *LR*, they are recorded ahead of *LR*. Due to dispersion, long-period surface waves advance on records toward the beginning of the wave train as it travels through the medium. An experienced analyst will distinguish between recorded surface waves that have traveled along pure oceanic or continental path. Dispersion of oceanic routes give rise to long wave trains with rather slow and sometimes hardly visible period change over relatively long (10–20 min) record segments (see Figure 4). In contrast, continental paths generate surface waves with characteristic fast period decrease with time, which is often easily recognized on records (see Figure 1). Surface waves can also travel by different modes (overtones), which are often seen on records as high-frequency components superimposed on the surface-wave train. We talk then about fundamental-mode and higher-mode surface waves (see *Surface Waves*). Higher modes are most frequently observed for waves traversing purely continental paths. Nevertheless, higher modes have in some cases also been associated with oceanic paths. Surface-wave higher modes disappear when the waves cross the transition between continental and oceanic structures. Higher modes propagate faster than the fundamental mode and are, therefore, recorded ahead of *LQ* and *LR*.

Dispersion curves (see *Surface Waves*) show a rather complicated pattern with several local minima and maxima. Surface waves traveling with these minima or maxima group velocities are called Airy phases. On seismograms, an Airy phase is characterized by a constant-frequency compact wave train, often with a remarkable amplitude buildup of dispersed surface waves traveling by fundamental-mode propagation. In the period range from approximately 50–200 s, the group velocity of *LR* (both the oceanic and continental paths) monotonically decreases with increasing period. Physically this means that in this period range, long-period *RL* waves follow the law of inverse dispersion. Observations of this phenomenon are rather scarce. An example is exhibited in Kulhanek (1990).

Extremely long-period surface waves called mantle waves, with periods from somewhat less than 1 min to several minutes, have been observed from large distant shocks. They can be of either Love- or Rayleigh-wave type. The former propagates with nearly constant speed of 4.4 km/s and shows an impulsive shape on the seismogram. The latter travels with velocity between 3.6 and 4.1 km/s. Wavelengths of mantle waves vary from several

hundreds to more than 1,000 km, so that a large part of Earth's mantle is affected by these waves. An interesting feature of mantle waves is their repeated appearance on records, which is due to their multiple travel around the Earth. The *LQ* mantle wave was given the label *G* (after B. Gutenberg) and the *LR* type mantle wave, the label *R*. The older nomenclature sometimes uses *W* instead for *R* (from German *Wiederkehrewellen*, meaning repeated waves). *G* waves that propagate the direct and anticenter routes are labeled *G1* and *G2*, respectively. Waves that have in addition traveled once around the Earth are denoted *G3* and *G4*, and so on. Similarly, we have *R1*, *R2*, *R3*, *R4*, etc. As an exceptional case we can mention records of the 1960 Chile earthquake, $M = 9.5$. Seismograms made at Uppsala, Sweden, reveal mantle waves *G20* and *R20* that have traveled a total distance equal to that from the Earth to the Moon (Båth, 1979).

Volcanic earthquakes and unusual seismic sources

During eruptive episodes, volcanoes can produce up to thousands of small earthquakes per day, recorded only at short distances, say, less than 50 km, from the volcanoes (see *Earthquakes, Volcanogenic*). Different categorization of volcanic earthquakes may be found in the literature. For example, Minakami (1961) classifies volcanic earthquakes into three groups, in accordance with different generation mechanisms and signal characteristics. (1) A-type earthquakes with focal depth between 1 and 10 km near the volcanic magma system. Corresponding records show clear high-frequency *P* and *S* phases. (2) B-type earthquakes with foci at depth of 1 km or less. Records reveal low-frequency coda, emergent onsets, and usually no distinct *S*. (3) Explosion-type earthquakes taking place at the very surface of the Earth. Close to active volcanoes, we also frequently detect so-called volcanic tremors, which are due to long-duration, more or less continuous, volcanic vibration.

Implosion earthquakes, impact earthquakes (e.g., the Tunguska, Siberia, impact in 1908), frost actions, low-magnitude icequakes generated by temperature changes in glaciers, earthquakes related to large-scale landslides, etc. are some of the types of events that, together with tectonic earthquakes, volcanic earthquakes, and oceanic microseisms, belong to the category of natural seismic sources. There is also a variety of man-made seismic sources such as industrial and military explosions, cultural noise (traffic, industry work), mining activity, high dams (seismicity, triggered/induced), fluid injections, etc. The source identification, for earthquakes other than tectonic or volcanic, is usually a task in itself. A classical example of source identification is the well-known problem of discriminating underground nuclear explosions from earthquakes (see *Seismic Monitoring of Nuclear Explosions*).

Conclusions

Seismogram interpretation, described in this article, is essentially devoted to the art of identification of various seismic "arrivals" or wave types visible on seismograms. It is usually followed by seismogram analysis, which may include determination of basic source parameters (origin time, hypocenter location, size) but may also cover more advanced studies, such as wave-form modeling, determination of velocity distribution, etc. Obviously, a large part of seismogram analysis is a domain of research. Nevertheless, phase identification is a doorway and without correct seismogram interpretation hardly any analysis would be possible. Repeated observations of peculiar phases on seismic records often led to new discoveries of details in Earth's structure and/or dynamics. Sometimes, theories were developed in advance, to be later confirmed by observations (e.g., free oscillations of the Earth). Seismogram interpretation is a fundamental, and in our view, also fascinating and rewarding part of modern seismology.

Bibliography

Adams, R. D., 1979. T-phase recordings at Rarotonga from underground nuclear explosions. *Geophysical Journal of the Royal Astronomical Society*, **58**, 361–369.

Bolt, B. A., 1982. *Inside the Earth*. Berkeley: Freeman.

Bormann, P. (ed.), 2002. *New Manual of Seismological Observatory Practice (NMSOP)*. Potsdam: Deutsches GeoForschungsZentrum.

Båth, M., 1979. *Introduction to Seismology*. Basel/Boston/Stuttgart: Birkhäuser.

Båth, M., 1983. Earthquake data analysis: An example from Sweden. *Earth-Science Review*, **19**, 1891–303.

Cao, A., Romanowicz, B., and Takeuchi, N., 2005. An observation of *PKJKP*: Inferences on inner core properties. *Science*, **308**, 1453–1455.

Engdahl, E. R., 1968. Seismic waves within Earth's outer core: Multiple reflections. *Science*, **161**, 263–264.

Engdahl, E. R., and Flinn, E. A., 1969. Seismic waves reflected from discontinuitie within Earth's upper mantle. *Science*, **163**, 177–179.

Engdahl, E. R., van der Hilst, R. D., and Buland, R., 1998. Global teleseismic earthquake relocation with improved travel times and procedures for depth determination. *Bulletin of the Seismological Society of America*, **88**, 722–743.

Herrin, E., Arnold, E. P., Bolt, B. A., Clawson, G. E., Engdal, E. R., Freedman, H. W., Gordon, D. W., Hales, A. L., Lobdell, J. L., Nuttli, O., Romney, C., Taggart, J., and Tucker, W., 1968. Seismological Tables for P phases. *Bulletin of the Seismological Society of America*, **58**, 1193–1352.

Jeffreys, H., and Bullen, K. E., 1940. *Seismological Tables*. London: British Association for the Advancement of Science.

Kennett, B. L. N., 1991. *IASPEI 1991 Seismological Tables*. Canberra: Research School of Earth Sciences, Australian National University.

Kennett, B. L. N., and Engdahl, E. R., 1991. Traveltimes for global earthquake location and phase identification. *Geophysical Journal International*, **105**, 429–465.

Kennett, B. L. N., Engdahl, E. R., and Buland, R., 1995. Constraints on seismic velocities in the Earth from traveltimes. *Geophysical Journal International*, **122**, 108–124.

Kulhanek, O., 1990. *Anatomy of Seismograms*. Amsterdam: Elsevier.

Mendiguren, J. A., 1971. Focal mechanism of a shock in the middle of the Nazca plate. *Journal of Geophysical Research*, **76**, 3861–3879.

Minakami, T., 1961. Study of eruptions and earthquakes originating from volcanoes, I. *International Geology Review*, **3**, 712–719.
Morelli, A., and Dziewonski, A. M., 1993. Body wave traveltimes and a spherically symmetric P- and S-wave velocity model. *Geophysical Journal International*, **112**(178), 194.
Neumann, F., 1951. Principles underlying the interpretation of seismograms. U.S. Department of Commerce, Coast and Geodetic Survey. Special publication number, 254.
Payo, G., 1986. *Introduccion al Analisis de Sismogramas*. Madrid: Instituto Geográfico Nacional.
Schenk, T., Muller, G., and Brustle, W., 1989. Long-period precursors to *pP* from deep-focus earthquakes: the Moho underside reflection *pMP*. *Geophysical Journal International*, **98**, 317–327.
Simon, R. B., 1968. *Earthquake Interpretations*. Golden: Colorado School of Mines.
Storchak, D. A., Borman, P., and Schweitzer, J., 2003. The IASPEI Standard Seismic Phase List. *Seismological Research Letters*, **74**, 761–772.
Wadati, K., 1927. Existence and study of deep earthquakes (in Japanese). *Journal of the Meteorological Society of Japan*, Ser 2, 5, 119–145.
Willmore, P. L., 1979. *Manual of seismological observatory practice*. Report SE-20, World Data Center A for Solid Earth Geophysics, NOAA, Boulder.

Cross-references

Body Waves
Core-Mantle Coupling
Earth's Structure, Upper Mantle
Earthquake, Aftershocks
Earthquake, Focal Mechanism
Earthquakes, Volcanogenic
Seismic Monitoring of Nuclear Explosions
Seismic Phase Names: IASPEI Standard
Seismic Velocity-Density Relationships
Surface Waves
Traveltime Tomography Using Controlled-Source Seismic Data

SEISMOLOGICAL NETWORKS

Eric Robert Engdahl[1], István Bondár[2]
[1]Center for Imaging the Earth's Interior, Department of Physics, University of Colorado at Boulder, Boulder, CO, USA
[2]International Seismological Centre, Pipers Lane, Thatcham, Berkshire, UK

Definition

Seismological Networks. Networks of seismographic stations for the recording of ground motions produced by seismic waves propagating from natural and anthropogenic sources.

Introduction

Seismology has its roots in observations of earthquakes and earthquake-generated ground motion. Seismological networks as envisioned by the pioneers in seismology are now a reality, a reality that is frequently upgraded and expanded as technology improves. A global system of broadband seismographs with high dynamic range is now in place. This is supplemented by many national, regional, and local networks capable of high-resolution monitoring of the frequent smaller earthquakes in seismically active places. Parametric data derived from these networks make a major contribution to national and international information services. Because of page limitations, we can only hope to provide the reader with a glimpse of all there is to know about seismographic networks.

Global seismic networks

Seismological recordings have been made on Earth for hundreds of years in some form or another. However, global monitoring of earthquakes only began in the 1890s when John Milne created 40 seismic observatories to measure the waves from these events (see Figure 1). Shortly after the International Geophysical Year (1957–1958), a concerted effort was made to establish and maintain a more modern standardized seismic network on the global scale.

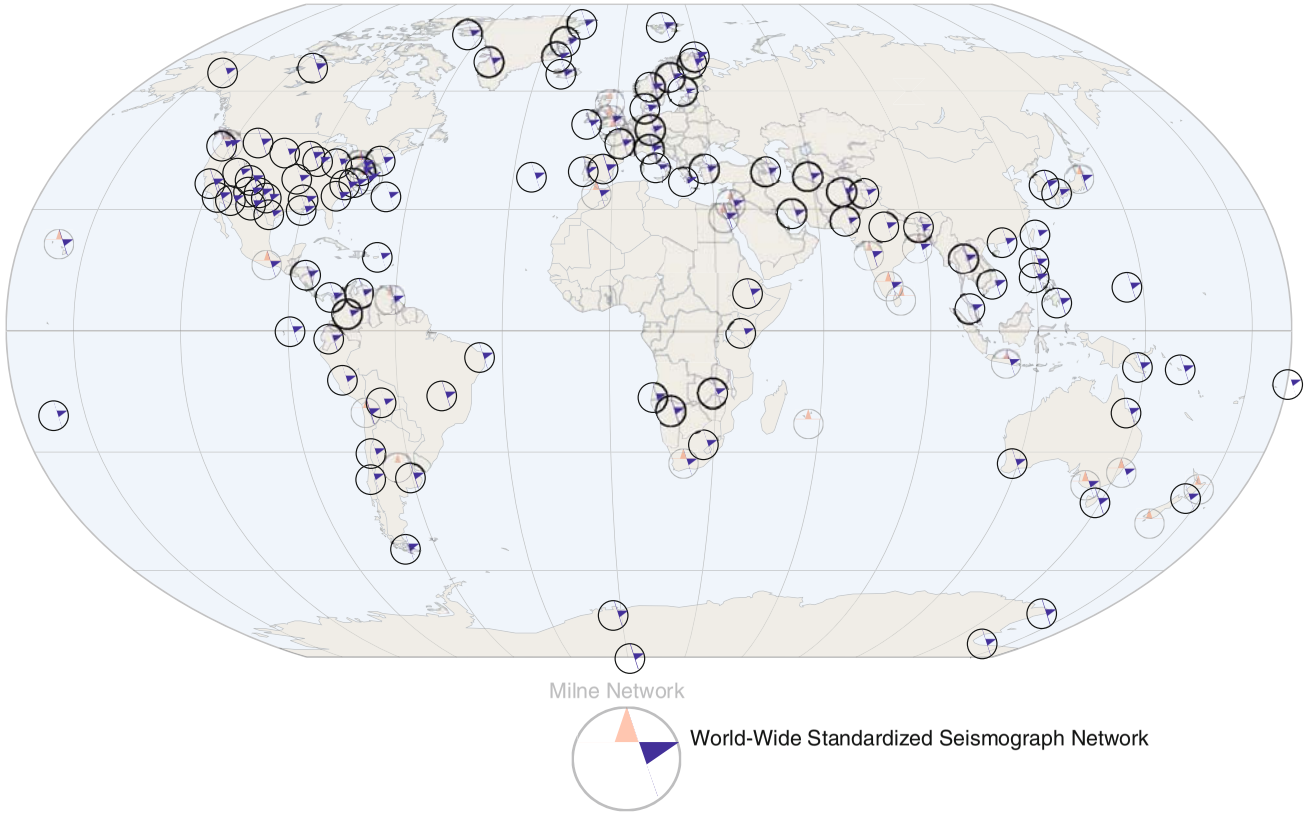
Worldwide Standardized Seismograph Network

In the early 1960s, the World-Wide Standardized Seismograph Network (WWSSN) was established. Between 1961 and 1966, 120 WWSSN stations with identical seismic instruments were deployed in more than 60 countries and islands throughout the world. The WWSSN program included the collection, review, and copying of the recorded seismograms, and the distribution of copies to researchers throughout the world. Considered one of the most important advances ever in observational seismology, the WWSSN produced the data needed to support unprecedented progress in earthquake, explosion, and tectonic research. In the years that followed the deployment of the WWSSN, many of the stations were modernized and expanded into regions not initially covered. Eighty of the original WWSSN stations are still in operation and still supported.

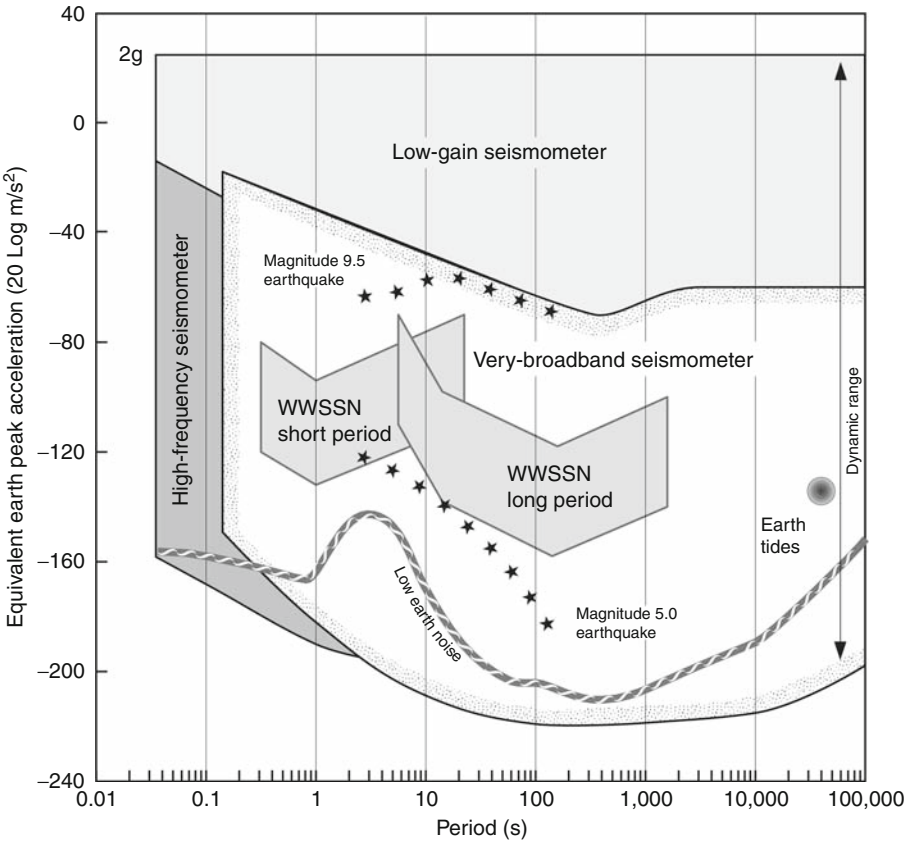
International Federation of Digital Seismograph Networks

In the 1980s, the international seismological community recognized the new opportunities within its field for improved understanding of the internal structure and dynamical properties of the Earth provided by recent developments in seismograph network technology. It also recognized that rapid access to seismic data networks of modern broadband digital instruments wherever they might be was now possible. The developments included greatly improved broadband seismographic systems that capture the entire seismic wave field with high fidelity (see Figure 2), efficient and economical data communications and storage, and widely available, powerful computing facilities.

In view of the above and to take advantage of existing developing global and regional networks, the International Federation of Digital Seismograph Networks (FDSN, http://www.fdsn.org/) was formed to provide a forum for: developing common minimum standards in seismographs (e.g., bandwidth) and recording characteristics (e.g., resolution and dynamic range); developing



Seismological Networks, Figure 1 Locations of stations in the Milne network and the World-Wide Standardized Seismograph Network. (Courtesy of KentAnderson, Incorporated Research Institutions for Seismology.)



Seismological Networks, Figure 2 Bandwidth and dynamic range of WWSSN and modern seismographic systems. (Courtesy of Rhett Butler, Incorporated Research Institutions for Seismology.)

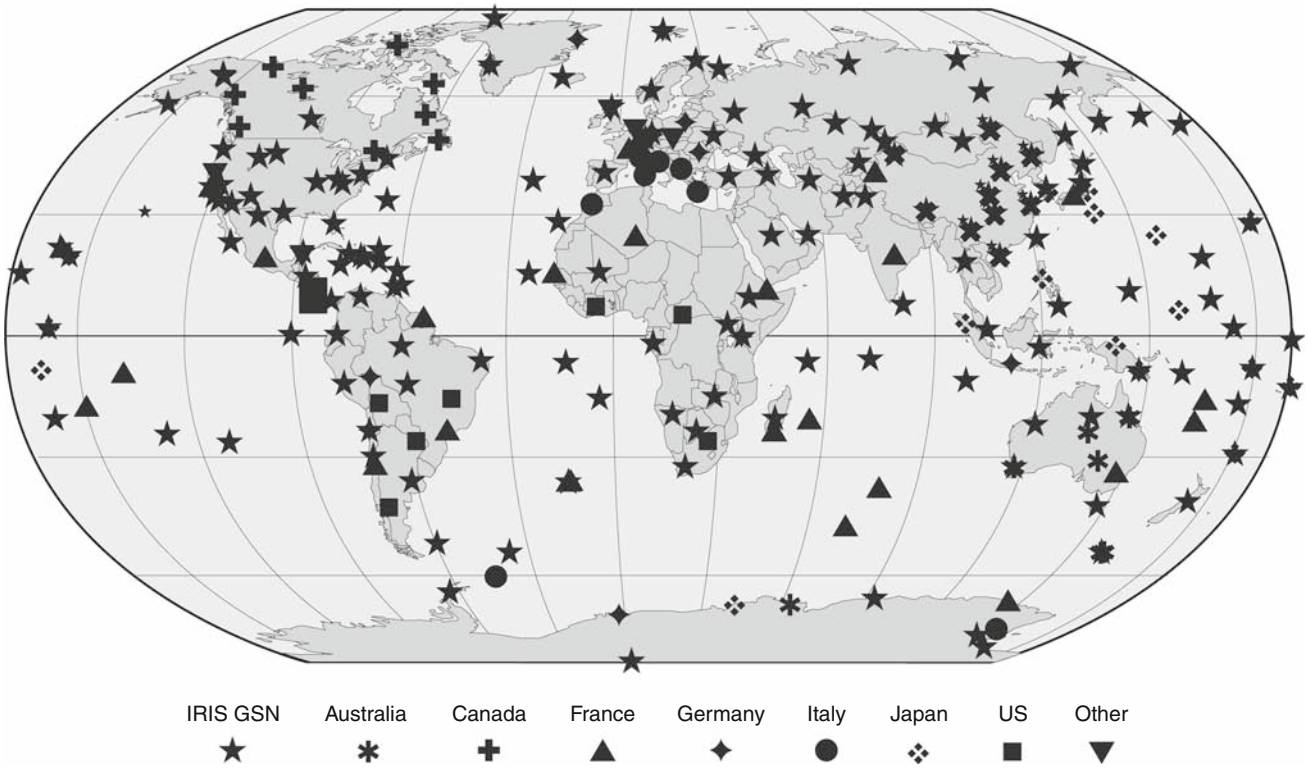
standards for quality control and procedures for archiving and exchange of data among component networks; and coordinating the siting of additional stations in locations that will provide optimum global coverage.

The FDSN was formed as a global organization, open to all national and international programs committed to the deployment of broadband seismographs and willing to contribute to the establishment of an optimum global system with timely data exchange. Its membership is comprised of groups responsible for the installation and maintenance of seismographs either within their geographic borders or globally. Membership in the FDSN is open to all organizations that operate more than one broadband station. However, among the most important contributors is the Global Seismographic Network (GSN) of the Incorporated Research Institutions for Seismology (IRIS), and the French GEOSCOPE and German GEOFON global networks. It is important to note that many stations in the FDSN were formerly part of the WWSSN program. Members agree to coordinate station siting and provide free and open access to their data. This cooperation helps scientists all over the world to further the advancement of earth science and particularly the study of global seismic activity.

The FDSN goals related to station siting and instrumentation are to provide stations with good geographic distribution, recording data with 24 bits of resolution in continuous time series with at least a 20 sample per second sampling rate. The FDSN was also instrumental in development of a universal standard (SEED) for distribution of broadband waveform data and related parametric information. The FDSN system of global network observatories includes contributions from many international partners (Figure 3), forming a large backbone of permanent seismological observatories. Developments in communications and other technological advances have expanded the role of the FDSN in rapid earthquake analysis, tsunami warning, and nuclear test monitoring. With such long-term observations, scientists are now getting a glimpse of Earth structure changes on human time scales, such as the rotation of the inner core. Continued observations for the next 50 years will enhance our image of the Earth and its processes.

International Monitoring System

Under the Comprehensive Test-Ban Treaty (CTBT, http://www.ctbto.org/), it is a requirement that there will be an International Monitoring System (IMS) to detect any clandestine nuclear weapon detonation in any environment – underground, under-water, or above ground. The aim of



Seismological Networks, Figure 3 Locations of backbone stations in the International Federation of Digital Seismograph Networks with major partners indicated by *symbols*. (Courtesy of Rhett Butler, Incorporated Research Institutions for Seismology.)

the IMS is detection, identification, and location of any such test, and the technologies involved will be seismological, hydroacoustic, infrasound, and atmospheric radionuclide monitoring. In support of these technologies, there will be appropriate means of global communication and an International Data Centre (IDC) to which monitoring data will be transmitted.

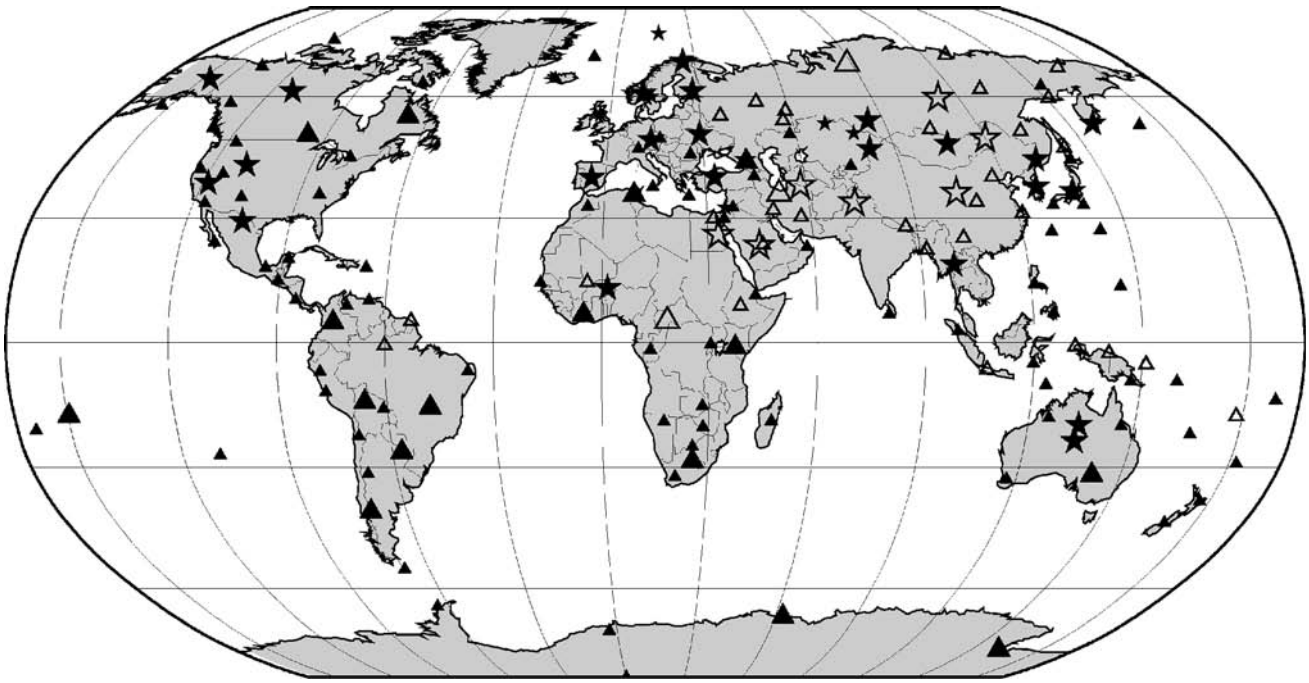
Seismological monitoring is considered the cornerstone of the IMS because of the likelihood that tests would be conducted underground and out of sight. The seismological technology required for the detection of earthquakes is well known, but the problem for the IMS is to distinguish between signals from weapons detonations, earthquakes, and non-nuclear explosions in quarrying or mining. There are so many earthquakes (>200,000 per year) and mining events worldwide, of similar magnitude to a small nuclear explosion, that data analysis becomes a key problem. The seismological verification program therefore includes intensive study of seismicity and mining operations in order to develop regional backgrounds.

A global network of 50 "primary" stations in 34 countries is planned, with two basic equipment packages: "three-component" (3-C) broadband seismometers, which act as single detection systems monitoring ground movement in 3 directions – vertically, and horizontally east-west and north-south; or "arrays" which include 3-C systems plus a cluster of narrowly spaced vertical-component short-period sensors distributed to gain optimum signal-to-noise ratios by waveform correlation. Arrays may contain many detectors located over large areas (up to 200 km^2). The advantage of an array station is that it allows approximate event location on a stand-alone basis. Important historical developments such as the LASA, NORSAR, Graefenberg arrays, and more recently the wide but densely distributed USArray, have marked important steps in seismological network development that have facilitated new research fields and services in seismology. As Figure 4 indicates, approximately half of the proposed stations will be arrays.

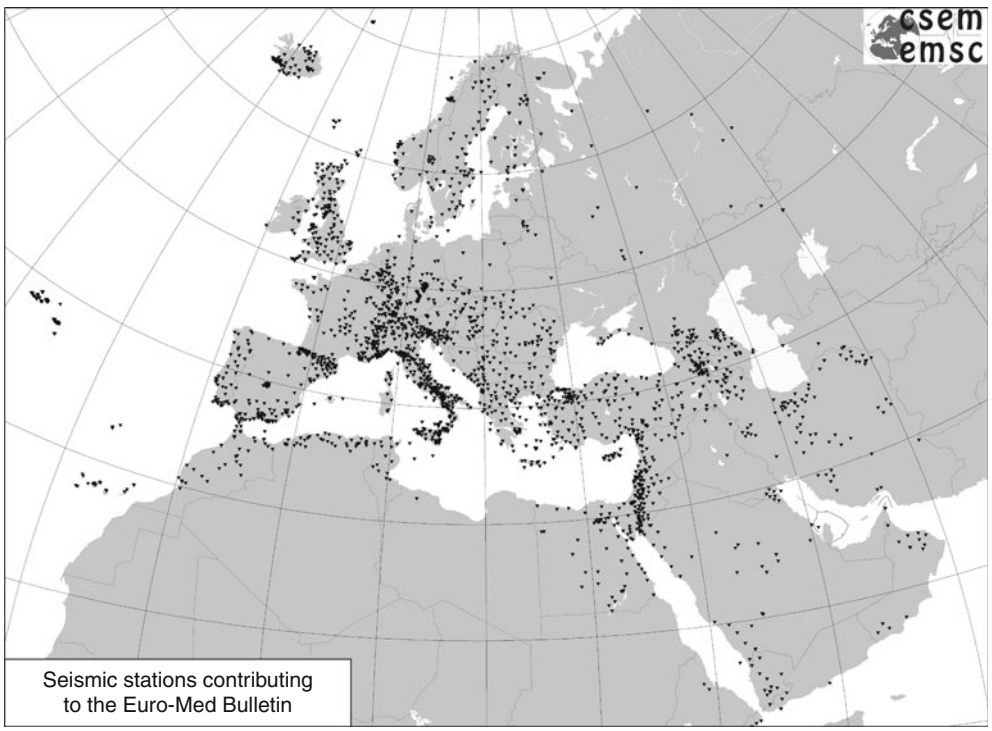
In addition to the primary network, there is to be an "auxiliary" network comprising 120 stations distributed among 61 countries (Figure 4). Stations in this network will mainly consist of existing 3-C stations, which already form part of the host countries' seismological monitoring operations. The purpose of the auxiliary network is simply to provide additional data to support that from the primary network in order to facilitate signal discrimination and hence event detection and location.

Regional seismic networks

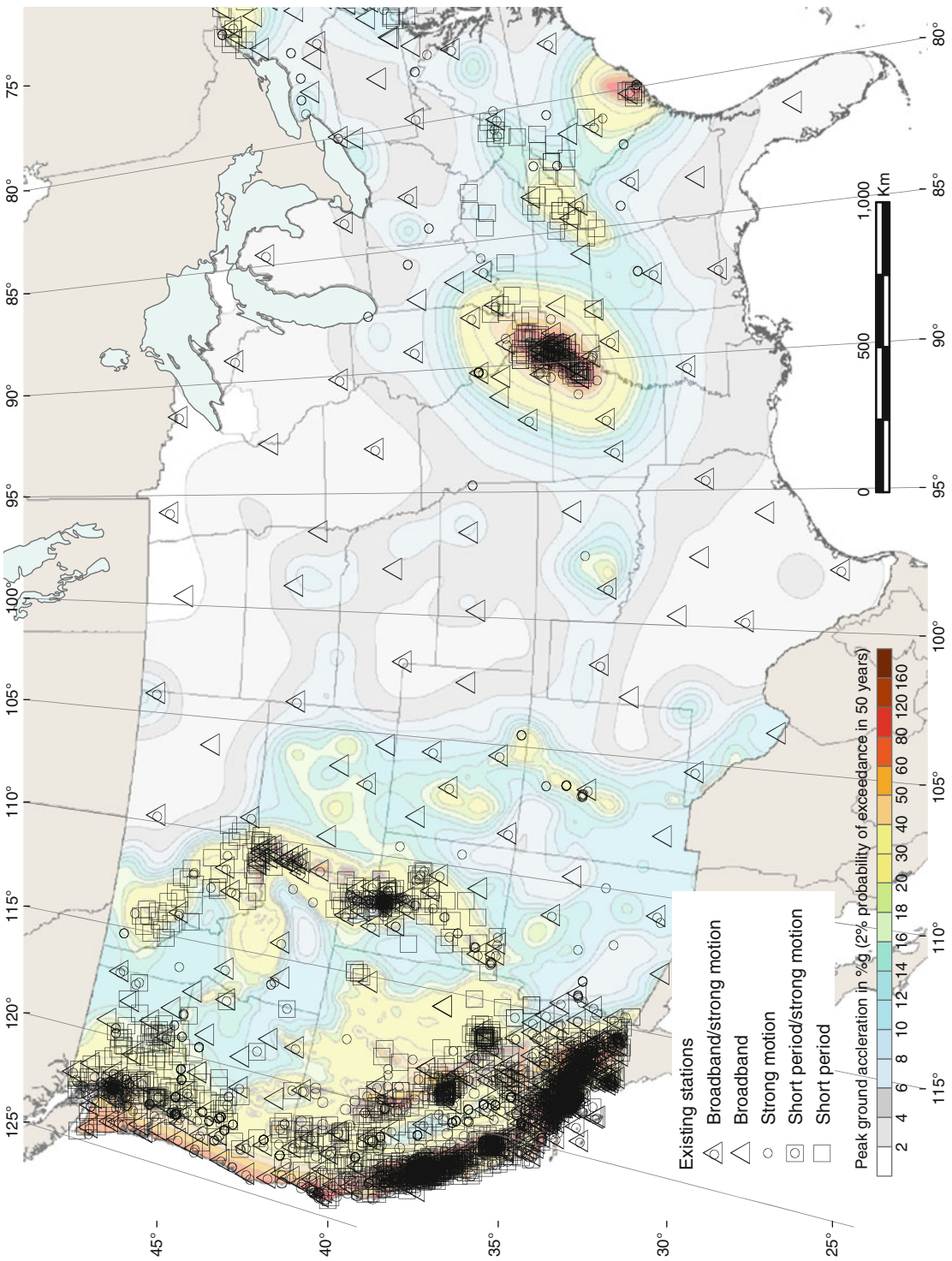
National seismographic networks operate within the political boundaries of a country, and their primary mission is to issue rapid alerts to government agencies and the general public for potentially damaging earthquakes. National



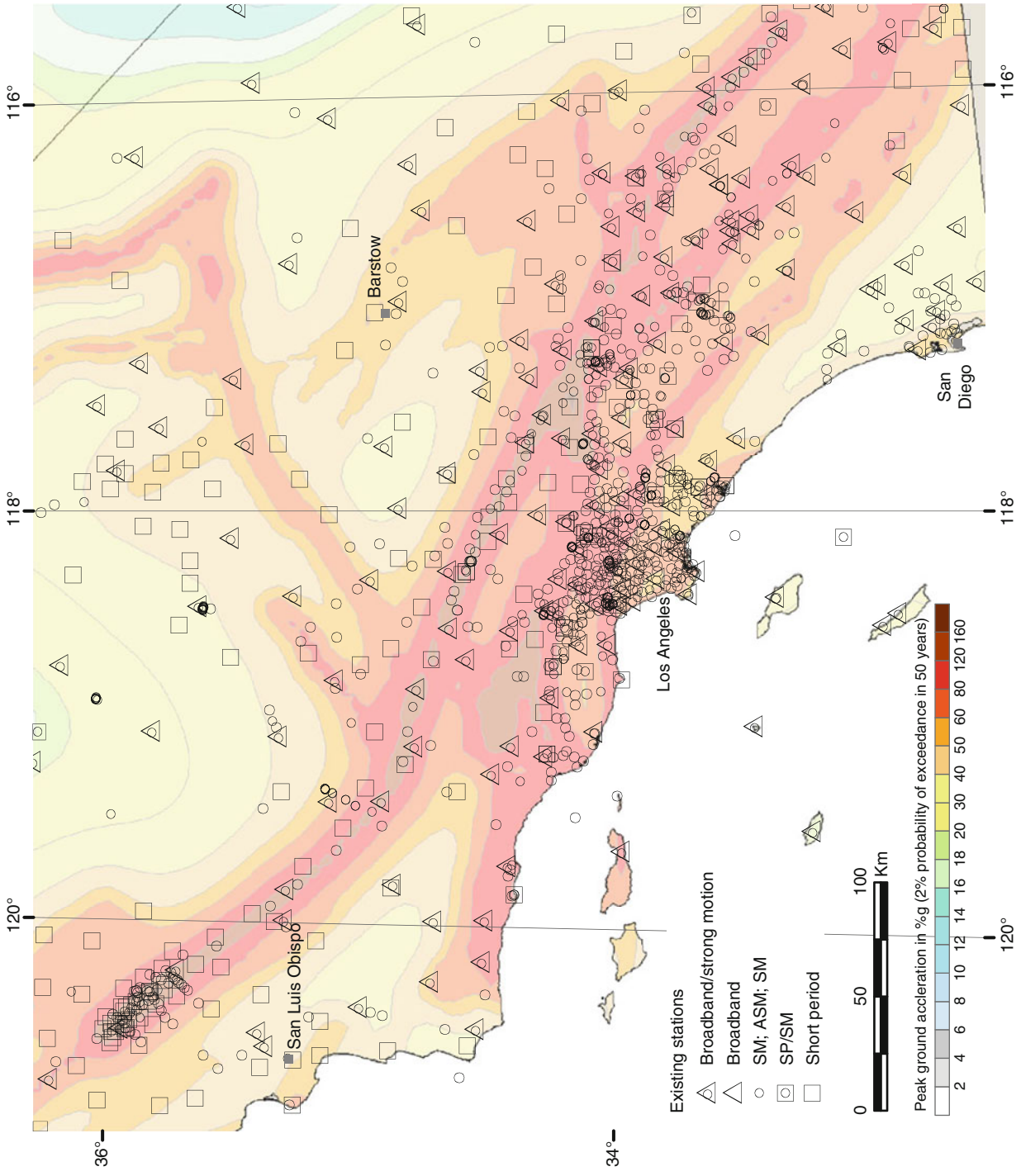
Seismological Networks, Figure 4 Locations of stations in the CTBTO IMS network (Courtesy of Ronan LeBras, Comprehensive Test-Ban-Treaty Organization). Symbols defined as follows: filled symbols – operational; open symbols – not yet operational; *stars* – array stations: *triangles* – three-component stations; *large symbols* – primary stations; and *small symbols* – auxiliary stations.



Seismological Networks, Figure 5 Locations of the contributing stations to the Euro-Med Bulletin 1998–2007 (Courtesy of Stephanie Godey, European-Mediterranean Seismological Centre).



Seismological Networks, Figure 6 Permanent seismic stations in the 48 conterminous states of the United States. (Courtesy of Susan Rhea, U.S. Geological Survey, National Earthquake Information Center.)



Seismological Networks, Figure 7 The Southern California Seismic Network: an example for a very dense regional network. (Courtesy of Susan Rhea, U.S. Geological Survey, National Earthquake Information Center.)

networks typically exchange data with neighboring countries in order to improve the accuracy of the determination of earthquake parameters in border regions.

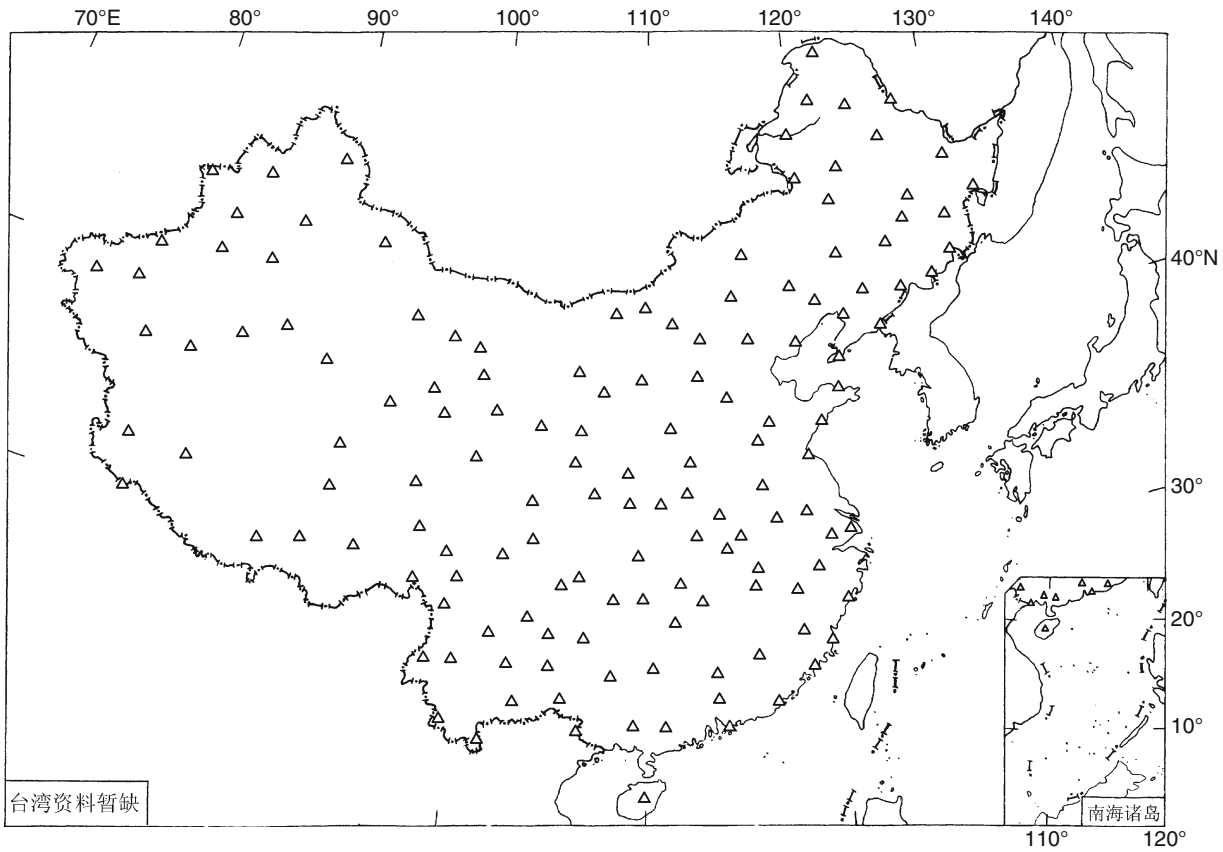
Regional networks play an important role in monitoring the seismicity of the Earth within a well-defined region. Regional networks may be part of a national network (e.g., USA, China, Japan, and Russia) or may be constituted from multinational networks (e.g., Euro-Mediterranean region). The magnitude detection threshold of regional networks is typically lower than that of the global networks, which allows for producing bulletins for a specific region that are more complete than it could be achieved at a global scale.

Local networks focus on a much smaller area than regional networks. Examples of local networks include temporary deployments to capture aftershock sequences of a large earthquake, networks monitoring volcano activities, and dense local networks installed in urban areas with elevated seismic risk.

Euro-Med region

The Euro-Med region encompasses Europe, North Africa, and the Middle East. The European-Mediterranean Seismological Centre (CSEM/EMSC, http://www.emsc-csem.org) produces earthquake bulletins in the region since 1998, using parametric data reported by national and local networks in the region. Its mission is to produce a rapid comprehensive seismological bulletin and issue alerts for potentially damaging earthquakes in the region. The EMSC and the National Earthquake Information Center (NEIC) routinely exchange data, and NEIC bulletin data for the European-Mediterranean region are incorporated in the EMSC bulletin. Earthquake parameters (location, magnitude, phase picks, moment tensors, etc.) provided by the EMSC are incorporated in the global bulletin prepared by the International Seismological Centre (ISC).

Figure 5 shows the seismic station network that currently contributes to the EMSC bulletin (Godey et al., 2006; Godey et al., 2009). All stations are registered at the International Registry of Seismograph Stations jointly maintained by the ISC and the NEIC. The recent years saw a dramatic increase in the number of stations in the region, especially in North Africa and the Middle East. These networks are vital for improving the azimuthal station coverage for the events, and thus improving their location accuracy. Waveform data from most stations contributing to the EMSC can be obtained from Observatories and Research Facilities for European Seismology (ORFEUS)



Seismological Networks, Figure 8 Locations of stations belonging to the China Digital Seismograph Network. (Courtesy of Liu Ruifeng, China Earthquake Network Center.)

the coordinating organization for seismic networks in Europe (http://www.orfeus-eu.org).

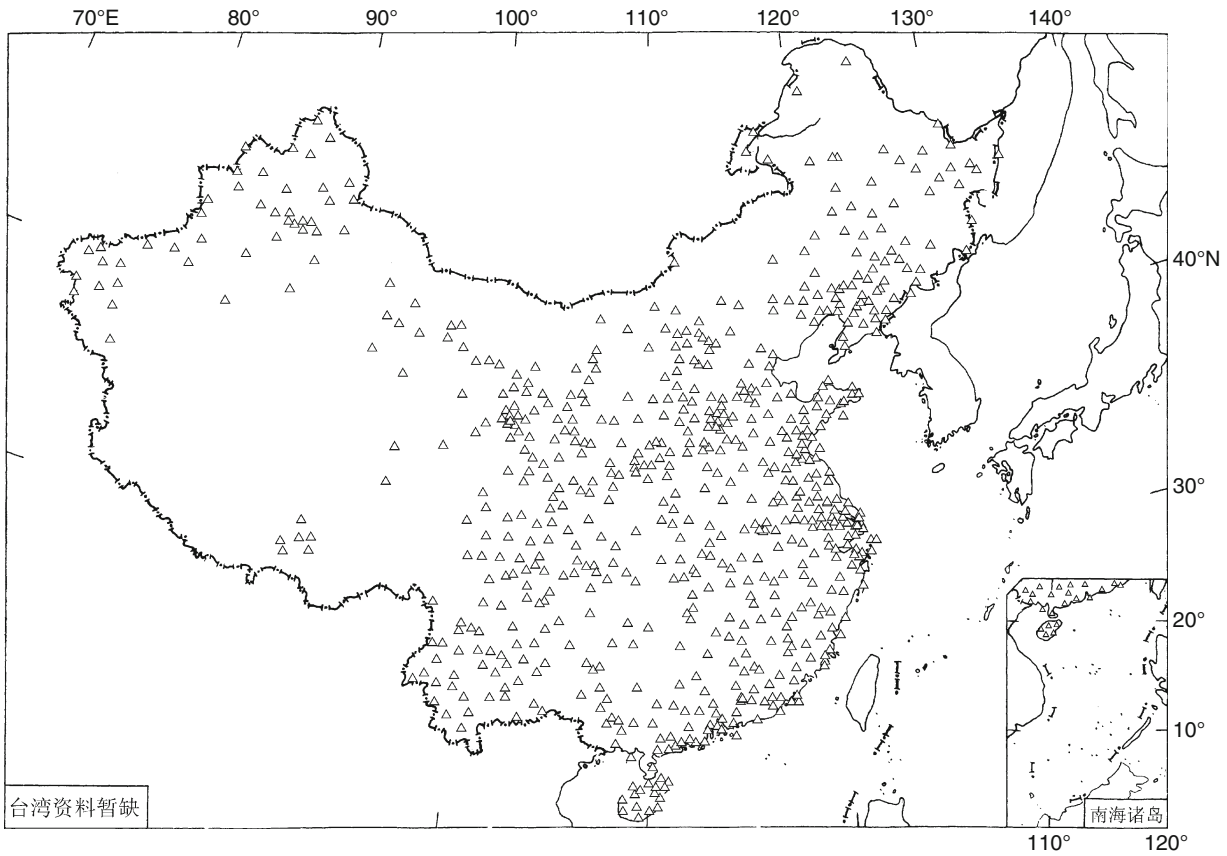
United States

In the United States, the U.S. Geological Survey's National Earthquake Information Center (NEIC, http://earthquake.usgs.gov/regional/neic) is the national data center and archive for earthquake information. Its mission is to determine as rapidly and accurately as possible the location and size of all significant earthquakes that occur worldwide. The NEIC operates 24-h-a-day and issues rapid alerts for earthquakes larger than magnitude 3 in the conterminous states of the United States and for those larger than magnitude 5 globally. The NEIC publishes global earthquake bulletins on a daily, weekly, and monthly basis.

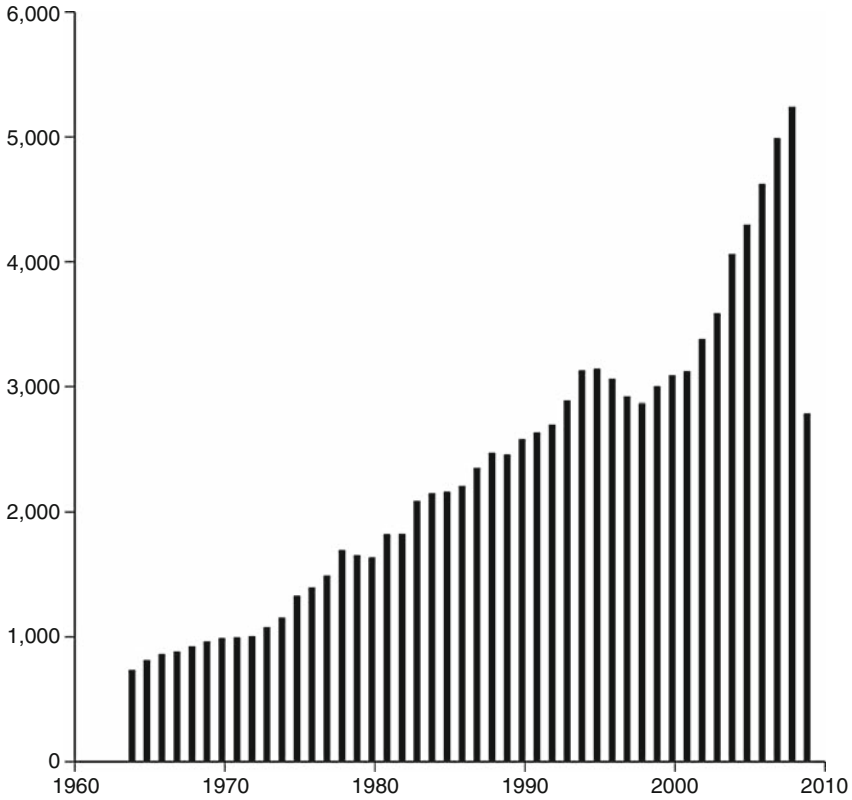
The NEIC serves as the National Operations Center of the Advanced National Seismic System (ANSS), which includes the ANSS backbone network and regional seismic networks across the United States. Figure 6 shows the permanent digital seismic and strong motion stations in the 48 conterminous states of the United States. The strong motion stations (circles) represent an integral part of the ANSS network as first arriving phases are picked on strong motion records and used in the earthquake location procedures. Figure 7 shows the regional Southern California Seismic Network. As in Figure 6, the map is colored by the peak ground acceleration, a measure of seismic hazard. The network is much denser in urban areas and regions with high seismic hazard.

China

The China Digital Seismograph Network (CDSN) has been established in 1986, with nine digital seismic stations. During the past 10 years, the Chinese seismic network, both at the national and regional levels, went through unprecedented improvements. The China Earthquake Administration (CEA) has completed the analog-to-digital transition of the existing analog stations and deployed a large number of new digital stations (Liu et al., 2008). Figure 8 shows the currently operating national network of 145 broadband seismographic stations. Some of these stations also belong to the FDSN. Dense regional networks further support the national network. Each of the 31 provinces, autonomous regions, and municipalities operates regional digital networks. Figure 9 shows the distribution of the 792 regional seismic stations.



Seismological Networks, Figure 9 Digital seismographic stations operated by 31 regional networks in China. (Courtesy of Liu Ruifeng, China Earthquake Network Center.)



Seismological Networks, Figure 10 Number of stations reporting to the ISC between 1964 and 2009. (Courtesy of James Harris, International Seismological Centre.)

The National Earthquake Network Center (NENC) archives the continuous waveform data from the national and regional networks and performs real-time data processing. The NENC issues rapid earthquake alerts for Ms >4.5 events inside China within 10 min and focal mechanism solutions in no more than 30 min after the earthquake occurred.

International Registry of Seismograph Stations

Since the humble beginnings of instrumental global seismology in the late nineteenth century, the number of seismographic stations has steadily increased every year. Figure 10 shows the number of stations reporting to the ISC in each year since its operations began. In order to be able to uniquely identify seismographic stations, the ISC and the NEIC, in its capacity as World Data Center A for Seismology, jointly maintain the International Registry of Seismograph Stations (IRSS). The IRSS represents a global catalog of seismic stations (currently with some 15,000 registered stations) and contains information about the station coordinates, instruments, operating networks, and when a station began/stopped operating. To facilitate data exchange between networks and organizations, it is strongly recommended that network operators register their stations and use the international station codes.

Summary

The high-quality data recorded by seismographic networks would be of limited value without international cooperation by all countries worldwide in the acquisition and exchange of seismic measurements and waveforms. This cooperation is essential for the location of earthquakes, for understanding the physics of earthquakes, and for studies of Earth's internal structure, properties, and processes. One of the latest challenges to acquiring, processing, and distributing data from seismographic networks by data centers globally is the automatic performance of traditional tasks in as close to real time as current technology permits.

Bibliography

Godey, S., Bossu, R., Gilbert, J., and Mazet-Roux, G., 2006. The Euro-Mediterranean Bulletin: a comprehensive seismological bulletin at regional scale. *Seismological Research Letters*, **77**, 460–474.

Godey, S., Mazet-Roux, G., Bossu, R., Merrer, S., and Guilbert, J., 2009. Ten years of seismicity in the Euro-Mediterranean region: Panorama of the EMSC Bulletin 1998–2007 (http://www.emsc-csem.org/docs/publications/ECGS_2009.pdf), Cahiers du Centre Européen de Géodynamique et de Séismologie, Vol. 28, pp. 1–14.

Liu, R.-F., Gao, J-Ch, Chen, Y.-T., Wu, Z.-L., Huang, Z.-B., Xu, Z.-G., and Sun, L., 2008. Construction and development of China digital seismological observation network. *Acta Seismologica Sinica*, **21**, 535–541.

Cross-references

Earthquake, Aftershocks
Earthquake, Location Techniques
Earthquake, Magnitude
Seismic Monitoring of Nuclear Explosions
Seismology, Monitoring of CTBT

SEISMOLOGY, GLOBAL EARTHQUAKE MODEL

Peter Suhadolc
Department of Geosciences, University of Trieste, Trieste, Italy

Definition

Exposure: Elements at risk; an inventory of those people or artifacts (and their characteristics) that are exposed to a hazard.

Seismic Hazard: A potentially damaging physical event, a phenomenon induced by an earthquake that may cause loss of life or injury, property damage, social and economic disruption, or environmental degradation. Seismic hazards can be single, sequential, or combined in their origin and effects. Each hazard is characterized by its location, intensity, frequency, and probability.

Seismic Risk: The probability of harmful consequences, or loss of life or injury, property damage, social and economic disruption or environmental degradation, resulting from interactions between seismic hazards and vulnerable conditions.

Vulnerability: The degree of loss of life or injury, property damage, social and economic disruption, or environmental degradation to a given element (or set of elements) resulting from a given hazard at a given level of intensity.

Introduction

A Global Earthquake Model (GEM) is being developed to establish an independent, uniform standard to calculate and communicate earthquake risk worldwide. By developing much-needed tools and software for reliable seismic risk assessment for basic and expert users in all regions of the world, GEM will provide necessary input for increased earthquake risk awareness and the undertaking of mitigating action.

The GEM initiative is organized as a public–private partnership, while an international community of scientists and professionals drives the development of the global earthquake model.

The work for GEM started in 2009, and a first comprehensive model plus accompanying tools and software, will become available by the end of 2013. Although much of the work is under development at the moment this article is being written, it is felt to be important to report on the initiative in this encyclopedia, because the global earthquake model will constitute an important contribution to the scientific community.

Scientific needs for GEM

The assessment and subsequent mitigation of earthquake risks is among the ultimate goals of both applied seismology and earthquake engineering. In spite of the fact that earthquake occurrence is globally a steady process and that most seismic events occur in uninhabited areas, the explosion in urban development and the enormous growth in the number of megacities in earthquake-prone areas (Figure 1), has considerably increased the seismic risk worldwide. Both cities in developed countries, like Tokyo or Los Angeles, and cities in developing countries, like Kathmandu or Jakarta, can suffer substantial damage due to a large earthquake in the next future. In fact, over 600,000 people died in the last century due to earthquakes (reference to USGS web site). Most of those deaths occurred in developing and emerging countries.

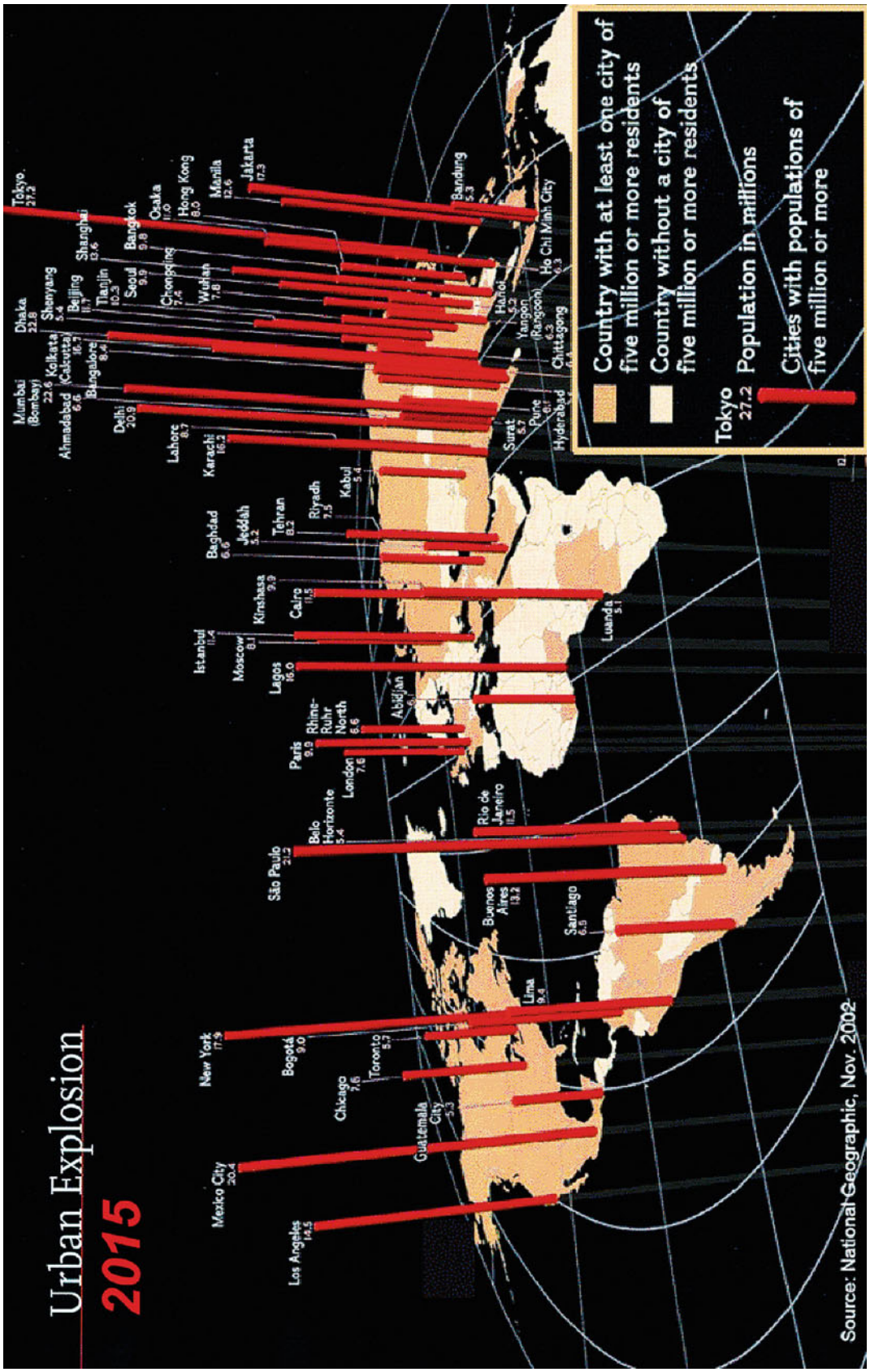
In many earthquake-prone regions, no seismic risk models exist, and even where they exist, they are not widely accessible. Such models are needed for accurate assessment of risks by (local) scientists, engineers, and practitioners, in order to promote mitigating actions, such as improvement of building codes and construction, sustainable land use, improved response, and protection of critical infrastructures. In order to have an effect on society at large, however, such models and the information resulting from it are needed as well by individuals and (international) organizations, to become aware of seismic risk and to undertake mitigating actions such as obtaining insurance, improved project management, and allocation of budgets for hazard mitigation.

The GEM initiative

By functioning as a community effort, the GEM initiative will produce a state-of-the-art dynamic and updatable model for the assessment of seismic risk worldwide; a model that is based on the probabilistic assessment of earthquake occurrence, the resulting ground motions, and the impact these have on structures and populations in terms of damage, social and economic loss. A model with underlying databases, that can be improved and enlarged with future data, and can be openly accessed through user-friendly software for data analysis and production of results, that can be also improved as our knowledge and technical capabilities rise in the future.

The global earthquake model is being designed and built by hundreds of experts and practitioners around the world, to ensure that less-monitored areas are also covered and to establish uniform standard, which allow for risk comparisons between countries and regions and for benchmarking output obtained through other sources. The model will reflect the needs, knowledge, and data of a variety of end users through GEM's extensive partner network. Such partnerships are essential in making sure that the information reaches the people that need it.

Technology transfer on the use of the software together with workshop opportunities will be provided, especially



Seismology, Global Earthquake Model, Figure 1 Urban explosion; it is predicted that by 2015 the projected increase in global population of 3 billion will concentrate itself in urban areas. (From National Geographic, 2002.)

in areas where risk assessment tools and data are currently less available.

A cooperative public–private partnership

The construction of the Global Earthquake Model is a cooperative public–private endeavor, and GEM is, therefore, structured as a partnership among country governments, private organizations, research institutions, international organizations and global risk initiatives, NGOs and individuals. The partnership includes a number of authoritative global institutions, such as the World Bank, the OECD, UNESCO, and UN's International Strategy for Disaster Reduction, but also the two largest international professional associations in the field: IASPEI (International Association of Seismology and Physics of the Earth's Interior) and IAEE (International Association for Earthquake Engineering), countries such as New Zealand, Norway, and Switzerland, and a number of prominent private corporations. Moreover, there are hundreds of institutions, organizations, and individuals involved in GEM that contribute expertise, data, or software, respond to requests for proposals, participate in regional programs, and take part in reviews and public assessments. Participation of individuals and institutions worldwide ensures that the model is owned by the global community and reflects its needs and knowledge.

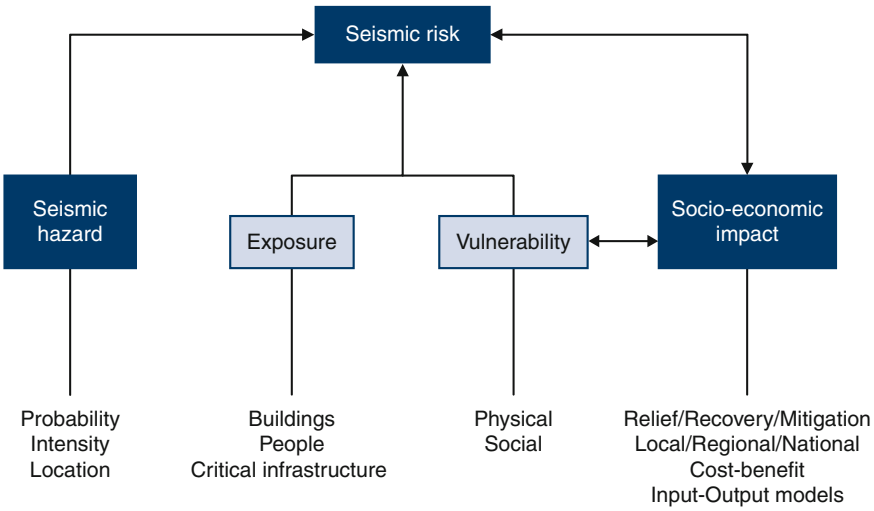
History

The idea for a Global Earthquake Model was born during a workshop called "Earthquake Science and its Contribution to Society" organized by the OECD's Global Science Forum in 2006. It was felt that a Global Earthquake Risk Scenario Map should be created. Various workshops followed in which the idea transformed into the creation of a Global Earthquake Model. Munich Re decided to sponsor the initiative in 2007. In 2008, the business plan for GEM was finalized and more partners were brought in to create a public–private partnership. The Eucentre was awarded the GEM Secretariat at the end of that year. In March 2009, the GEM Foundation was incorporated as a nonprofit foundation and that marked the start of the GEM initiative.

Scientific framework

The GEM scientific framework serves as the underlying basis for constructing the model, and consists of three principal integrated modules (Figure 2: Scientific Framework): Hazard, Risk, and Socioeconomic Impact.

Seismic risk is defined as a product of seismic hazard (the probability of levels of ground shaking, resulting from earthquakes, within a given time span), seismic vulnerability (the probability of loss given a level of ground shaking), and exposure (the elements at risk – mainly buildings, critical infrastructure, and humans). Risk gives an indication of the extent of loss (damage, fatalities, casualties) that can be expected in a given location in the world. Risk can, therefore, be high in an area without significant probabilities of ground shaking, because it has an older, more vulnerable, and densely populated building stock, and lower in an area with high levels of seismicity, but with well-constructed structures that are sparsely inhabited. Earthquakes, however, have an impact that goes beyond physical damage or casualties. Earthquakes can severely damage the economy and influence society and social well-being. Therefore, GEM will include innovative methods for analysis and evaluation of the impacts of earthquakes on the short, medium and long term, on local and global scales. There will also be applications that build upon the model, such as a tool for cost-benefit analysis, allowing users to understand what effect certain mitigation actions, strengthening of the building stock for example, will have on the risk. Insight into earthquake effects over time will directly support decisions on



Seismology, Global Earthquake Model, Figure 2 The current design of the scientific framework of GEM.

short-term needs after an event (relief, shelter), medium-term needs (recovery and reconstruction), and long-term needs related to policies on risk mitigation.

Implementation and current status

In June 2010, the GEM initiative has been able to deliver a proof-of-concept for hazard and risk calculations on a global scale. This will be the basis for development of the final GEM risk engine and the model. The model building process and engine are described in more detail further on.

International consortia, involving hundreds of professionals and institutions, are working on the creation of necessary standards, databases, and methodologies on a global level. These are the global components of the model. The work on Hazard Global Components has started and will be delivered in 2012. The work on Risk Global Components will start in the fall of 2010 and will be delivered in 2012 and 2013 and the work on the Socio-economic Global Components will take off in early 2011, with the goal to be finalized in 2013. The Global Components are further specified below.

Programs are being setup in many regions of the world as independently run, bottom-up projects, and links are established with ongoing regional programs. Both such programs are defined as GEM Regional Programs and involve a great number of local experts who will use GEM software, will generate local data, will validate the data and standards that were created on a global level, and will serve as a starting point for technology transfer in the region. Currently, three GEM Regional Programs are operational: in the regions of Europe and the Middle East, and collaboration is ongoing in Central America. Programs are being prepared in Africa, South-Asia, South-East Asia and Oceania, Central Asia, South America, the Caribbean, North-East Asia.

Global components

Global Components are the scientific modules of GEM that are developed at a global scale to provide standards, models, tools and data. Global components will provide the basic input to the model, but the Regional Programs will deliver detailed feedback and input from a local point of view.

To ensure that Global Components are developed by the international scientific community, GEM releases Requests for Proposals. These RfPs are developed by a group of international experts on the topic, and peer reviewed through an open public commenting system. International consortia, groups of institutions, and individual experts respond to these calls. All proposals are peer-reviewed by at least four external reviewers and extensively discussed by GEM's Scientific Board, who prepares an advice for GEM's Governing Board. Finally, for each component, one consortium is selected to carry out the work.

At the moment of this writing, proposals for the hazard and risk global components have been selected and the work has started or is about to start. The global component for socioeconomic impact is under development.

The hazard module

Five global components on seismic hazard are being developed as input to the global earthquake model:

Global historical catalog and database

The record of past earthquakes is among the most important means to evaluate earthquake hazard, and the distribution of damage associated with past earthquakes is a key to assessment of seismic risk. The instrumental seismic record has only a 100-year span, and yet no plate boundary is so active that this period is sufficient to capture the full range of seismic behavior. Extending the record of large damaging earthquakes several hundred years longer, and in exceptional cases by 1,000 years, is thus extremely valuable. This requires damage descriptions to be converted to numerical intensity scales, and the estimation of magnitude, location, and their uncertainties. Currently, an International consortia is working on the best and most efficient way to make use of the historical earthquake record, honoring at the same time its uncertainties and regional differences in quality and extent.

Global instrumental seismic catalog

An International consortium is building a reliable and uniform global instrumental earthquake database (1900–2009). A uniform location procedure to relocate global earthquakes from 1900 to 2009 shall be used, and standardized methods for computing both associated homogeneous surface-wave magnitude ($M_{S(\mathrm{BB})}$) using amplitude and period data and associated moment magnitude (M_W) from published seismic moments (M_0) and also proxy values, $M_{W[\mathrm{Mx}]}$, converted from other types of primary magnitudes [$\mathrm{M_x}$] using empirical relationships (Scordilis, 2006; Bormann et al., 2009; Bormann and Yadav, 2010). A recent example of such catalog structure and philosophy has been published by Yadav et al. (2009). In addition to producing a uniform global earthquake catalog, the database will also keep track of original input data files and documentation. These materials will allow seismologists to extend earthquake studies to a lower magnitude threshold for a better coverage of seismicity, especially in local and regional areas. The catalog will be the primary tool to be used to characterize the spatial distribution of seismicity, the magnitude–frequency relation, and the maximum magnitude.

Global active fault and seismic source database

Tectonic earthquakes are fault ruptures; seismic hazard assessments should, therefore, incorporate an inventory of active faults. Despite this, many seismic hazard assessments do not consider faults at all, or do so only sparingly because the requisite fault data are absent or inadequate. The need to incorporate active faults in the computation of a seismic hazard map, in order to estimate reliably the future strong ground motion, has been recognized long ago (Wesnousky et al., 1984; Wesnousky, 1986). Within the scope of GEM, a uniform global active fault and seismic source database is built, with a common set of

strategies, standards, and formats. The database will be publicly available. The effort is building upon the 1989–2004 International Lithosphere Program's Major Active Faults of the World. Since some faults cut the earth's surface and others do not, there will be both observational elements (active faults and folds that have slipped in the past 10–100 kyr) and interpretative elements (inferred seismic sources) to the database. In addition to collecting the best fault information presently available, appropriate mechanisms to capture new fault data as it becomes available, to capture the rapid expansion of fault knowledge and the ongoing debate about fault geometry, kinematics, and interaction are also being developed.

Global geodetic strain rate model
In the assessment of seismic hazard, seismic catalogs and active faults database provide complementary means for identifying zones of earthquake hazard. The geodetically measured secular strain rate provides an independent benchmark for crustal deformation and thus the recurrence of large earthquakes. The creation of a comprehensive and uniform model for geodetic strain rates is the scope of this global component. The consortium will critically review all global and regional studies since 1994 and will significantly update the Global Strain Rate Model of 2004. Moreover, they will assess whether the estimated fault slip rates and earthquake activity rates are consistent with the long-term strain accumulation measured from Global Positioning System (GPS) (or derived from interferometric synthetic aperture radar – InSAR).

Global ground-motion prediction equations (GMPEs)
A source of variability in hazard results obtained in small regions has traditionally been the derivation and use of local GMPEs, which are often based on insufficient data. Another factor influencing the variability of hazard results is the large statistical uncertainty of ground motion predictions; in recent years, a number of promising advances have been made – the Next Generation Attenuation (NGA) models (Power et al., 2006) – allowing to correctly account for various effects such as local site conditions, style-of-faulting effect, hanging-wall effect, etc., on ground motion predictions. Within the scope of GEM, by compiling a global reference hazard assessment model, a harmonized suite of GMPEs is developed, built on the most recent advances in the field.

Local soil conditions have a large influence on the seismic ground motions. The International consortium involved in this component is furthermore working on adopting a unified strategy to (a) determine seismically effective soil parameters, preferably based on the average S-wave velocity over the upper 30 m, V_{S30}, or, alternatively, derived from topography data (Wald and Allen, 2007; Allen and Wald, 2009) on a worldwide level, and (b) to represent the spatial distribution of soil classification, as for example proposed by NEHRP (Martin, 1994; BSSC, 2004) or EuroCode8 (CEN, 2004), compatible with the GMPE models.

The risk module

Five Global Components on seismic risk are being developed as input to the global earthquake model:

GEM ontology and taxonomy
Within the fields of hazard and risk assessment, diverse terminology is used and different meanings are sometimes attached to the same words. In order to achieve a shared understanding across the disparate fields and endeavors encompassed by GEM, an international consortium will develop some methodological foundation and terminology. Ontology refers to the entire framework that guides the development of the global earthquake model – the set of concepts and the relationship between these concepts that will allow determination and communication of earthquakes risk. The GEM Ontology will be general and comprehensive enough to be long-lasting, but must be adaptable to future conditions. Taxonomy, a part of ontology, refers to the classification of things in an ordered system, reflecting their relationships. The GEM Taxonomy will be an earthquake-related system of classification, set of terminology, encompassing hazard, risk, and socioeconomic realms. The adopted Ontology and Taxonomy will be evaluated and tested and finally globally promoted and disseminated in a continuous way.

Global earthquake consequences database
The international consortium working on this global component will create a global database of earthquake impacts and consequences. They will assemble and store in a structured and web-accessible way both data (including photos) already acquired and data yet to be acquired following future events. Data (both statistical and observational) typically covers building damage, damage to lifelines and other infrastructures, ground failure, human casualties, social disruption, and financial and economic impacts. All damages will be geographically referenced and viewable on a global mapping system. The database will be equipped with analytical tools enabling data fields to be post-processed across events, globally or within regions.

Global exposure database
Compiling an open database of the global building stock distribution with associated uncertainties containing the spatial, structural, and occupancy-related information for damage, loss, and human casualty estimation models is the focus of this Global Component. The International Consortium working on this will identify, evaluate, and homogenate the various existing databases at country, regional, and city levels throughout the world. Within an appropriate grid cell resolution, the inventory will include the total number of buildings, their floor area, the relative distribution of building types (e.g., timber, masonry, R/C, steel) along with some performance-influencing features, such as construction quality and year of construction. Moreover, the relative distribution of occupancy types (e.g., residential, industrial or agricultural) and the

temporal population (e.g., day- and nighttime) within each cell and for each building type will be considered. Finally, the presence of critical facilities and infrastructures for emergency relief will also be identified for each cell.

Inventory data capture tools

This Global Component aims to provide the tools that will enable the capture and transfer of high-resolution inventory data into either the Global Exposure Database or Global Earthquake Consequences Database. Both data collected using remote sensing or acquired from direct observation will be merged in this process.

Interpretation of low, medium, or high spatial resolution satellite imagery and of aerial images is foreseen for data capturing. This will involve the determination of geo-referenced footprints of buildings, vegetation (important for post-event fire spread), and other infrastructure (e.g., roads, lines, reservoirs). Appropriate software will upload and aggregate such data into a neighborhood or grid level of resolution. Direct observation on the other hand implies street-front direct inspection or interpretation of street-front data from photographs.

Global vulnerability estimation methods

The global vulnerability assessment comprises the estimation of the degree of direct loss due to building damage. Direct loss means both human losses (injuries and fatalities) as well as economic losses (repair costs, downtime, etc.) arising from the damage due to ground shaking at a given level. The goal of this global component is to provide methods and standards for vulnerability assessment that can be applied to a wide taxonomy of structures. The International consortium working on the component will derive a number of initial "default" functions that can be applied on a global scale. The methods will account for varying levels of detail in the input data, include the characterization of uncertainties, and account for the influence of retrofitting on the vulnerability.

The socioeconomic impact module

This module is considered as a single global component, which will allow for assessment of social and economic consequences of earthquakes, further and beyond those direct losses considered in the risk module, with the goal to provide users with a wide array of methods useful to communicate impacts of seismic events. The module is envisioned as a toolbox that shall gather a comprehensive set of models, metrics, data, and tools to be organized following spatial, temporal, and user principles under a standardized, clear, and simple framework.

To achieve this goal, GEM will propitiate the development of a framework for compiling state-of-the-art, widely accepted, useful methods for diagnostics and decision-making. The toolbox will be initially populated with methods generated by GEM in partnership with an international consortium, and after that through a constant process of interactions and consulting with the wider community. The toolbox will permit, to the extent possible, the integration of methods, and include interfaces for end users to input data and parameters, to conduct sensitivity and counterfactual analysis, to evaluate alternative policy interventions, and, to visualize results. Case studies are expected to be generated using the methods included in the toolbox.

All the tasks to be conducted within the social and economic impact module shall be performed following a participatory framework that will encompass GEM, its partners, and the wider community. This must be the case, given the multidimensional character and the complexities involved in linking variables from the natural, social, and economic systems, and considering that no consensus has been perceived regarding a precise definition of social and economic impacts of earthquakes and the methods for measuring or addressing them.

Model building

Currently, GEM risk engine is being developed. The engine will allow for calculations on a global level and, therefore, needs to be able to incorporate the data, standards, models, and methods developed by the international community through the Global Components and Regional Programs. A proof-of-concept of the engine has been delivered in June 2010, demonstrated by first hazard and risk calculations on a global scale, which resulted in preliminary global output. This output, however, has not been validated and will only be used internally to propel further development. The GEM risk engine is characterized by open-source development; hence development and enhancement by a wider community of expert users and programmers. It will be platform-independent, modular (using object-oriented language), flexible (as to allow for future multi-hazard calculations), expandable (in terms of methodologies employed), and scalable.

In order to serve the needs of various users, an understandable user interface is being constructed encompassing tools and software for transparent earthquake risk calculations, and risk communication, incorporating the latest technologies for sharing and contributing data between users.

Outcome and future

GEM is going through a continual user-needs assessment effort, to ensure that the software and tools that are being developed meet the needs of a wide range of possible users. Partnerships and an active user-community are the ingredients that support the initial use of the tools and subsequent adoption of the information that the global earthquake model produces, as a necessary first step toward awareness and risk-mitigating behavior.

The main output of GEM's first 5-year working program will be the OpenGEM platform, which will allow basic and expert users to run applications, access seismic risk information on local, national and regional scale, and visualize the latter in maps, curves, tables and export these in compatible formats. Basic users are likely to want

to view output produced by the global earthquake model, perhaps that related to the location of their own house. Expert users will be able to "plug in" their own data and run their own calculations. Because not everyone will be able to access an internet portal, or would like to run calculations through the internet, a stand-alone OpenGEM software package will be an important derivative.

GEM will, however, produce more than a platform for risk assessment. Global harmonized databases within the fields of earthquake hazard, vulnerability, exposure, and socioeconomic impact will be made available, such as a global earthquake consequences database and a global historical seismic catalog. GEM will also produce best practices and standards related to many aspects of seismic risk assessment, which will help the community to work together under a common framework at a global scale. A community development platform for the computational engine will allow for open-source and object-oriented development of the GEM risk engine by the community. Programmers and other experts will be able to test, use, and further improve GEM's software code. Finally, there will be technical reports for the (scientific) community to use, and technical training programs and workshops for diffusion of the knowledge on GEM software and use.

After completion of the first working program, GEM will continue its activities and work on extending the model, maintenance and improvement of the tools, will support the needs of an ever-growing user community, will work on dissemination of products and results and the development of new applications and partnerships.

The future challenge for the scientists will probably be not so much to invalidate the concept of GEM, but to improve its global components (methods, data and standards), as our scientific knowledge of the earthquake generation, seismic wave propagation, and their impact on the built environment – with all the related consequences – will advance in the years to come.

Current developments of the GEM initiative can be followed on www.globalquakemodel.org.

Acknowledgments

I would like to thank Nicole Keller, the GEM Outreach and Communications Officer, for the help she has given me and for providing Figure 1. Thanks also to David Wald for his constructive criticism on the preliminary version of this contribution.

Bibliography

Allen, T. I., and Wald, D. J., 2009. On the use of high-resolution topographic data as a proxy for seismic site conditions (V_{S30}). *Bulletin of the Seismological Society of America*, **99**, 935–943.

Bormann, P., and Yadav, R. B. S., 2010. Reply to comment on a homogeneous and complete earthquake catalog for Northeast India and the adjoining region by R.B.S. Yadav, P. Bormann, B. K. Rastogi, M.C. Das, and S. Chopra by R. Das and H.R. Wason. *Seismological Research Letters*, **81**(2), 235–240.

Bormann, P., Liu, R. F., Xu, Z. G., Ren, K. X., Zhang, L. W., and Wendt, S., 2009. First application of the new IASPEI teleseismic magnitude standards to data of the China National Seismographic Network. *Bulletin of the Seismological Society of America*, **99**, 1868–1891.

BSSC, 2004. *NEHRP (National Earthquake Hazard Reduction Program) Recommended Provisions for New Buildings and Other Structures (FEMA 450), 2003 Edition*. Washington, DC: Building Seismic Safety Council, National Institute of Building Sciences.

CEN (Comité Européen de Normalisation), 2004. Eurocode 8: Design of Structures for Earthquake Resistance, Part 1: General rules, seismic actions and rules for buildings. European Standard EN 1998-1:2004. CEN: Brussels, Belgium.

Martin, G. M. (ed.), 1994. *Proceedings of the NCEER/SEAOC/BSSC Workshop on Site Response during Earthquakes and Seismic Code Provisions*. University of California, Los Angeles, November 18–20, 1992

Power, M., Chiou, B., Abrahamson, N., and Roblee, C., 2006. The Next Generation of Ground Motion Attenuation Models (NGA) project: an overview. In: *Proceedings of the 8th U.S. National Conference on Earthquake Engineering*, Paper No. 2022.

Scordilis, E. M., 2006. Empirical global relations converting Ms and mb to moment magnitude. *Journal of Seismology*, **10**, 225–236.

Wald, D. J., and Allen, T. I., 2007. Topographic slope as a proxy for seismic site conditions and amplification. *Bulletin of the Seismological Society of America*, **97**, 1379–1395.

Wesnousky, S. G., Scholz, C. H., Shimazaki, K., and Matsuda, T., 1984. Integration of geological and seismological data for the analysis of seismic hazard: a case study in Japan. *Bulletin of the Seismological Society of America*, **74**, 687–708.

Wesnousky, S. G., 1986. Earthquakes, quaternary faults, and seismic hazard in California. *Journal of Geophysical Research*, **91**, 12587–12631.

Yadav, R. B. S., Bormann, P., Rastogi, B. K., Das, M. C., and Chopra, S., 2009. A homogeneous and complete earthquake catalog for Northeast India and the adjoining region. *Seismological Research Letters*, **80**, 609–627.

References to websites

USGS website: http://earthquake.usgs.gov/regional/world/world_deaths.php

Cross-references

Seismic Hazard

SEISMOLOGY, MONITORING OF CTBT

Scientific and Technical Advances in Seismology and their Relevance

Wu Zhongliang[1], Paul G. Richards[2]
[1]Institute of Geophysics, China Earthquake Administration, Beijing, China
[2]Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, USA

Introduction

Monitoring of the *Comprehensive Nuclear-Test-Ban Treaty* (CTBT) entails the detection, identification, and

characterization of atmospheric, underwater, and specifically underground nuclear tests, and the discrimination of nuclear explosions from other artificial and/or natural events such as quarry blasts and earthquakes (CTBTO PrepComm., 2009a). Matching the needs of verification deals with seismic, infrasound, and hydro-acoustic, as well as radionuclide monitoring systems. At present the CTBT verification system specified by the treaty has three major components: the International Monitoring System (IMS) with a global network of 337 monitoring facilities; the International Data Centre (IDC) for the processing of observational data; and an On-Site Inspection (OSI) regime that utilizes a series of high-resolution technologies and has many similarities to the emergency on-site observation of disaster areas such as the meizo-seismal regions of large earthquakes. In this verification system, seismic monitoring has been the leading subject, due to three aspects: (1) Since 1980 up to this time of writing (2010), all nuclear tests have been conducted underground, and seismic monitoring is the most efficient way to monitor tests in that environment. (2) Development of seismic monitoring played the leading role in the development of IMS. Compared to other monitoring techniques, seismic monitoring data, characterized by the real-time transmission of broadband/high-frequency continuous waveforms, plays a major role in the whole monitoring system. (3) Important concepts in CTBT monitoring, including evasion scenarios such as decoupling and mine-masking which are most directly a challenge to seismic monitoring, were directly or indirectly developed from seismic monitoring practice.

Entry *seismic monitoring of nuclear explosions* describes in detail the basic properties of explosion and earthquake signals and seismic technologies for the detection and identification of explosions. The main tools for such a monitoring purpose, local, regional, and global seismic networks, are introduced in the entry *seismological networks*. Sections following focus on recent scientific and technical advances in seismology and their relevance to CTBT monitoring.

Evaluation and design of the seismic monitoring system: a systems engineering perspective

Monitoring and verification require an international observational system composed of the following components: observational stations, communication links from the stations to data center/s, and tools for the analysis of signals recorded for verification purposes. The evaluation and design of the seismic monitoring system from the perspective of systems engineering is one of the significant advances in recent years. It is useful not only for the functioning and sustainability of the CTBT monitoring system but also for the operation of modern seismological networks, which not only record earthquakes but also have missions such as the fast seismological information service and the managing of huge amount of observational data. In 2008, the Project "International Scientific Studies of the Implementation of the CTBT Verification System" (ISS) was organized by the Preparatory Commission of CTBT Organization (CTBTO PrepComm., 2009b), in which *system performance evaluation* provided useful tools for the evaluation and design of the seismic monitoring system.

Evaluation is to assess the readiness and capability of the monitoring and verification system, while design is to plan the road map to ensure the expected readiness and capability. Design and evaluation of a seismological monitoring system includes four perspectives: (1) *physical perspective,* including primary seismic stations and arrays, auxiliary stations, national data centers, international data center, network maintenance centers, communication links, system for the automatic data processing and review of analyst, database, and data sharing devices; (2) *functional perspective,* including data transmission, automatic data processing, automatic seismic phase picking and event location, interactive phase picking and location by analysts, array processing for location, magnitude determination, determination of earthquake parameters such as moment tensor and radiated energy, discrimination tools, and tools for routine evaluation of the state-of-health; (3) *operational perspective,* including the design, implementation, management, and upgrading of the whole system; and (4) *scientific perspective,* dealing with Earth structure model, travel-time database, algorithm for location and phase picking, attenuation model and site correction for magnitude determination, Green functions or attenuation relations for source parameter determination, azimuth correction database for array detection, and event database for discrimination. It may be unnecessary for a seismologist to directly use the concepts and tools of systems engineering. However, methodologies such as specification of the system and its components as well as the relationships among the components, specification of the functions of the system and decomposition of system functional requirements into component functional requirements, and development of measures of performance at system level and component level, respectively, provide a clear concept for the systematic evaluation and design of the seismic monitoring system. This becomes especially important when dealing with a modern seismological monitoring system with many stations, working continuously in a real-time mode, being a technically complex system.

Advances in seismology, potentially relevant to CTBT monitoring

Recent decades have witnessed the fast development of seismology (*Seismological Grand Challenges* Writing Group, 2008). These advances have apparent relevance to CTBT monitoring. Location of earthquake epicenters using conventional methods has the uncertainty of up to 10^1–10^2 km. This uncertainty is significantly reduced by the modern methods of earthquake location using waveform cross-correlation and relative location technique (Richards et al., 2006). Recent investigation reveals that "repeating earthquakes," identified by waveform

cross-correlation, seem to be much more common than expected in seismic activity (e.g., Schaff and Richards, 2004), providing the evaluation of location accuracy with an innovative and efficient tool (e.g., Jiang and Wu, 2006). By waveform comparison and relative location, it becomes possible for seismologists to determine the accurate location of new tests if data from previous nearby tests is available (e.g., Wen and Long, 2010). As archives grow, new seismic events that appear problematic often can best be resolved by comparison of their signals with those of previously recorded events from the same region.

In the evaluation of location accuracy, one of the key concepts is the *ground truth* (GT) *events* (International Seismological Centre, 2009), a database of reference earthquakes and/or explosions for which the epicenter information is known with high confidence (to ±5 km or better) with seismic signals recorded at regional and/or teleseismic distances. GT events are used as "calibrating information" to assess the location accuracy of the monitoring system. Furthermore, a GT database is useful to facilitate better visualization of Earth structure, better modeling of velocities of seismic waves, and more accurate travel-time determinations, which are all essential to the enhancement of monitoring capability. Using and developing the GT event database is thus endorsed by international seismological organizations such as the International Association of Seismology and Physics of the Earth's Interior (IASPEI, 2009a).

Recent developments of seismology kept challenging some of the conventional ideas well-accepted in the CTBT monitoring practice. At present, there are several complementary methods for the identification of an explosion from an earthquake, such as the difference between m_b and M_S, and the ratio of high-frequency (>2 Hz) P to S energy (Richards and Zavales, 1990; Bowers and Selby, 2009). One of the important measures is the minimum size of earthquakes and/or explosions for the discrimination, called the *magnitude threshold,* below which the discrimination tool fails to work. Conventionally, the *magnitude threshold* for discrimination is regarded as different from, often 0.5 magnitude unit larger than, the *magnitude threshold* for detection (Hannon, 1985). However, new results (CTBTO PrepComm., 2009b) showed that these thresholds appear to be very similar if regional data, adequate for measurement of spectral ratios, is available. For purposes of estimating the location and yield of underground explosions, it is always important to have accurate information on the inhomogeneous structure of the Earth within which seismic waves propagate. New approaches to this problem using seismic noise (e.g., Shapiro et al., 2005) provide an innovative tool for imaging the Earth structure. Increase of seismic stations and development of inversion techniques kept improving the Earth structure model (*Seismological Grand Challenges* Writing Group, 2008). Accordingly, traditionally problematic earthquake parameters such as focal depth, and "modern" earthquake parameters such as moment tensor and radiated energy potentially can contribute more to CTBT monitoring, along with improved 3D Earth models, and accurate account of wave propagation in complex 3D media.

With the development of digital seismic waveform analysis techniques, it has also become possible for seismologists to capture the time-lapse behavior of Earth medium with considerable accuracy (e.g., Grêt and Snieder, 2005). Combining with the recent advances in active source technology, this development in turn has direct impact on the OSI technology. Combination of seismic data with remote sensing images shows potential in revealing more details of the explosion source, such as decoupling (Sykes et al., 1993). This combination, enabled by modern information technology such as "Digital Earth" (Wu and Chen, 2000), or more practically Google Earth®, provides an alternative approach that supplements the traditional concepts of OSI and even verification.

"Forensic seismology": evidences and judgments

Seismology is a branch of observational science based on the analysis of signals from different sources, penetrating through the Earth, and at last recorded by seismic networks. The field related to seismological observatory practice (Bormann, 2002) and the operation of seismic networks is sometimes called *seismological observation and interpretation* (IASPEI, 2009b). When seismology is applied to the monitoring of CTBT, more considerations are needed beyond the recording and analysis of seismic signals. Key concepts related to the practical functioning of the seismic monitoring system are *monitoring* and *verification*. The term *monitoring* refers to technical activities and analyses associated with data from observational systems that acquire signals from different sources, either explosions or earthquakes, in a continuous and, in some cases, real-time regime. The term *verification* refers to evaluations that include non-seismologists, to enable authorities and the public to judge whether some detected phenomena had its origin in a nuclear explosion. To this end, *verification* is regarded as a branch of *forensic seismology* (Bowers and Selby, 2009), in which the word "forensic" means answering questions of interest to (international) legal system/s.

In concept, *monitoring* is different from *verification* in that *monitoring* basically provides objective observational evidences, while *verification* has to have some subjective decision-making processes based on the objective (but in some cases limited) evidences provided by monitoring. An example is the announced test of DPRK in 2009. It was declared by DPRK authorities that the nuclear test was successfully conducted, but there were no direct evidences (radionuclide) to prove that the explosion, recorded clearly by seismic networks, was really a successful *nuclear* test. Another example is the 1998, Indian-Pakistan announced tests – even if seismic monitoring data could not provide persuasive evidences that *all* the tests were conducted as successfully as announced. In most cases, however, the monitoring system serves for

the clarification of special events, such as the earthquake near Lop Nor of China on March 13, 2003 (Bowers and Selby, 2009). In fact, an important task of the seismic monitoring system is to *screen* events of interest, to identify events that are definitely not nuclear tests. Note that in the context of forensic science, to verify that a nuclear test has been conducted somewhere and to verify that no nuclear test has been conducted in a certain place for a specific time duration, even when using the same monitoring system, need different sets of evidence from that system.

Concluding remarks: interaction between science and CTBT monitoring

The concept of seismic monitoring of nuclear tests can be traced back to the earliest underground nuclear explosions (Bolt, 1976; Richards and Kim, 2009). In September 1957, during the General Assembly of the International Union of Geodesy and Geophysics (IUGG), Toronto, K. E. Bullen gave the address "Seismology in our Atomic Age." Needs of nuclear test monitoring played an important role in the development of global seismic networks in the 1960s (Richards, 2002). In the meantime, researches have been carried out extensively for the detection and identification of nuclear tests (Husebye and Mykkeltveit, 1981; Richards, 2002), in which one of the remarkable techniques is seismic array for the detection of explosion signals smaller than those monitored by conventional stations (Douglas, 2002).

In recent years, the international CTBT monitoring system has been in operation as a "big science device" that is similar in scale to accelerators in high-energy physics and satellites in space science. It provides not only experiences and lessons in the practice of monitoring, but also unique datasets which are interesting in basic research and the science and technology for sustainability. Advances in science and technology also promote the implementation and development of CTBT monitoring practice, and assessments of the overall capability. In the above sections, these relevant advances are introduced via three components: *engineering*, *science*, and *decision making*. It is worth mentioning that, as indicated by a series of important events, especially the Conference "CTBT: Synergies with Science, 1996–2006 and Beyond" (CTBTO PrepComm., 2006) and the ISS Project (CTBTO PrepComm., 2009b), a new era of cooperation between CTBT monitoring communities and scientific research communities has started, which will in turn contribute both to the CTBT monitoring practice and to the development of geophysical science. In the ISS Project, *science and technology foresight*, a professional practice that emerged in recent years, aiming at identifying today's research and innovation priorities on the basis of scenarios of future developments in science and technology, society and economy, was introduced to the scientific assessment of CTBT monitoring, which indicates that the interaction between science and CTBT monitoring has started to be systematic and comprehensive through a professionally planned road map.

Bibliography

Bolt, B. A., 1976. *Nuclear Explosions and Earthquakes – The Parted Veil*. San Francisco: W. H. Freeman.

Bormann, P. (ed.), 2002. *IASPEI New Manual of Seismological Observatory Practice*. Potsdam: GFZ.

Bowers, D., and Selby, N. D., 2009. Forensic seismology and the comprehensive Nuclear-Test-Ban treaty. *Annual Review of Earth and Planetary Sciences*, **37**, 209–236.

CTBTO PrepComm. 2006. CTBT: Synergies with Science, 1996–2006 and Beyond. http://www.ctbto.org/the-organization/ctbt-synergies-with-science1996–2006-and-beyond/.

CTBTO PrepComm. 2009a. Verification Regime. http://www.ctbto.org/verification-regime/.

CTBTO PrepComm., 2009b. ISS 09 – International Scientific Studies. http://www.ctbto.org/specials/the-international-scientific-studies-project-iss/.

Douglas, A., 2002. Seismometer arrays – their use in earthquake and test ban seismology. In Lee, W. H. K., Kanamori, H., Jennings, P. C., and Kisslinger, C. (eds.), *International Handbook of Earthquake and Engineering Seismology*. Amsterdam: Academic, pp. 357–368.

Grêt, A., and Snieder, R., 2005. Monitoring rapid temporal change in a volcano with coda wave interferometry. *Geophysical Research Letters*, **32**, L06304, doi:10.1029/2004GL021143.

Hannon, W., 1985. Seismic verification of a comprehensive test ban. *Science*, **227**, 251–257.

Husebye, E. S., and Mykkeltveit, S. (eds.), 1981. *Identification of Seismic Sources – Earthquake or Underground Explosion*. Dordrecht: D. Reidel.

IASPEI, 2009a. Resolutions at the 2009 Scientific Assembly, Cape Town, South Africa. http://www.iaspei.org/resolutions/resolutions_2009_cape_town.pdf.

IASPEI, 2009b. Commission on Seismological Observation and Interpretation. http://www.iaspei.org/commissions/CSOI.html.

International Seismological Centre, 2009. Reference Event Bulletin. http://www.isc.ac.uk/GT/index.html.

Jiang, C. S., and Wu, Z. L., 2006. Location accuracy of the China national seismograph network estimated by repeating events. *Earthquake Research in China*, **20**, 67–74.

Richards, P. G., 2002. Seismological methods of monitoring compliance with the comprehensive nuclear test ban treaty. In Lee, W. H. K., Kanamori, H., Jennings, P. C., and Kisslinger, C. (eds.), *International Handbook of Earthquake and Engineering Seismology*. Amsterdam: Academic, pp. 369–382.

Richards, P. G., Waldhauser, F., Schaff, D. P., and Kim, W.-Y., 2006. The applicability of modern methods of earthquake location. *Pure and Applied Geophysics*, **163**, 351–372.

Richards, P. G., and Kim, W.-Y., 2009. Monitoring for nuclear explosions. *Scientific American*, **300**(3), 64–71.

Richards, P. G., and Zavales, J., 1990. Seismic discrimination of nuclear explosions. *Annual Review of Earth and Planetary Sciences*, **18**, 257–286.

Schaff, D. P., and Richards, P. G., 2004. Repeating seismic events in China. *Science*, **303**, 1176–1178.

Seismological Grand Challenges Writing Group, 2008. *Seismological Grand Challenges in Understanding Earth's Dynamic Systems, Report of the Long Range Science Plan for Seismology Workshop, September 18–19, 2008*. Washington: Incorporated Research Institutions for Seismology (IRIS). http://www.iris.edu/hq/lrsps/.

Shapiro, N. M., Campillo, M., Stehly, L., and Ritzwoller, M. H., 2005. High resolution surface wave tomography from ambient seismic noise. *Science*, **307**, 1615–1617.

Sykes, L. R., Deng, J., and Lyubomirskiy, P., 1993. Accurate location of nuclear explosions at Azgir, Kazakhstan, from satellite images and seismic data: implications for monitoring decoupled explosions. *Geophysical Research Letters*, **20**, 1919–1922.
Wen, L., and Long, H., 2010. High-precision location of North Korea's 2009 nuclear test. *Seismological Research Letters*, **81**, 26–29.
Wu, Z. L., and Chen, Y. T., 2000. Potential impact of "Digital Earth" on seismological monitoring. *Seismological Research Letters*, **71**, 548–552.

Cross-references

Seismic Monitoring of Nuclear Explosions
Seismological Networks

SEISMOLOGY, ROTATIONAL

William H. K. Lee
U.S. Geological Survey, Menlo Park, CA, USA

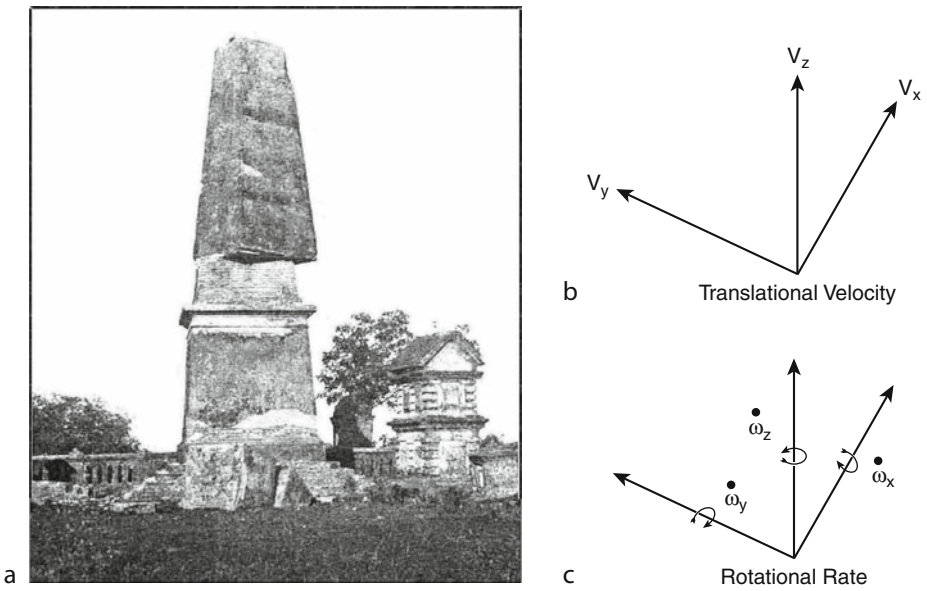
Definition

Rotational seismology is an emerging study of all aspects of rotational motions induced by earthquakes, explosions, and ambient vibrations. The subject is of interest to several disciplines, including seismology, earthquake engineering, geodesy, and earth-based detection of Einstein's gravitation waves.

Introduction

Rotational effects of earthquake waves together with rotations caused by soil-structure interaction have been observed for centuries (e.g., rotated chimneys, monuments, and tombstones relative to their supports). A summary of historical examples of observations on earthquake rotational effects is provided by Kozák (2009), including reproduction of the relevant sections from Mallet (1862) and Reid (1910). Figure 1a shows the rotation of the monument to George Inglis (erected in 1850 at Chatak, India) as observed by Oldham (1899) after the 1897 Great Shillong earthquake. This monument had the form of an obelisk rising over 60 ft. high from a base 12 ft. on each side. During the earthquake, the topmost six-foot section was broken off and fell to the south and the next nine-foot section was thrown to the east. The remnant is about 20 ft. in height and is rotated about 15° relative to the base.

A few early authors proposed rotational waves or at least some "vortical" motions. Many different terms were used for the rotational motion components at this early stage of the field's development. For example, "rocking" is rotation around a horizontal axis, sometimes also referred to as "tilt." Mallet (1862) proposed that rotations of a body on the Earth's surface are due to a sequence of different seismic phases emerging at different angles. Reid (1910) studied this phenomenon, which was observed in the 1906 San Francisco earthquake, and pointed out that the observed rotations are too large to be produced by waves of elastic distortion. Such waves "*produce very small rotations, whose maximum amount, . . . is given by the expression* $2\pi A = \lambda$*, where A is the amplitude and* λ *the wavelength; with a wave as short as 10,000 feet (3 km) and an amplitude as large as 0.2 of a foot (6 cm), the maximum rotation would only be about 0.25 of*



Seismology, Rotational, Figure 1 (**a**) Rotation of the monument to George Inglis (erected in 1850 at Chatak, India) as observed by Oldham (1899) after the 1897 Great Shillong earthquake. (**b**) Coordinate system for translational velocity. (**c**) Coordinate system for body-fixed or instantaneous rotational rate.

a minute of arc [0.0042 degrees, or 7.3 micro-radians (μrad)], *a quantity far too small to be noticeable*" (Reid, 1910, p. 44). A modern analysis of such rotational effects is presented in Todorovska and Trifunac (1990).

Observational seismology is based mainly on measuring *translational* motions because of a widespread belief that *rotational* motions are insignificant. For example, Richter (1958, footnote on p. 213) states that "Theory indicates, and observation confirms, that such rotations are negligible." Richter provided no references, and there were no instruments at that time sensitive enough to either confirm or refute this claim. Recent advances in rotational seismology become possible because sensitive rotational sensors have been developed in aeronautical and astronomical instrumentation. In this article we provide some background information and selected highlights of recent advances in rotational seismology. An extensive list of references to original papers is given because few "synthesis" articles exist.

Early attempts in studying rotational motions

Ferrari (2006) summarized two models of an electrical seismograph with sliding smoked paper, developed by P. Filippo Cecchi in 1876 to record three-component *translation* motions and also the *torsion* movements from earthquakes. Although these instruments operated for several years, no rotational motion could be recorded because of low transducer sensitivity. Pioneers in several countries attempted to measure rotational motions induced by earthquakes. Nearly a century ago, Galitzin (1912) suggested using two identical pendulums installed on different sides of the same axis of rotation for separate measurement of rotational and translational motion. This was later implemented, for example, by Kharin and Simonov (1969) in an instrument designed to record strong ground motion. Using an azimuthal array of seismographs, Droste and Teisseyre (1976) derived rotational seismograms for rock bursts in a nearby mine. Inspired by Walter Munk, Farrell (1969) constructed a gyroscopic seismometer, and obtained a static displacement of <1 cm and a tilt of <0.5 μrad at La Jolla, California, during the Borrego Mountain earthquake of April 9, 1968 (magnitude 6.5), at an epicentral distance of 115 km.

Early efforts also included studies of explosions. For example, Graizer (1991) recorded tilts and translational motions in the near field of two nuclear explosions using seismological observatory sensors to measure point rotations directly. Nigbor (1994) measured rotational and translational point ground motions directly with a commercial rotational MEMS sensor and found significant near-field rotational motions (660 μrad at 1 km distance) from a one-kiloton explosion.

Rotations and strains of the ground and of response of structures have been deduced indirectly from accelerometer arrays using methods valid for seismic waves having wavelengths that are long compared to the distances between sensors (e.g., Trifunac, 1979, 1982; Oliveira and Bolt, 1989; Spudich et al., 1995; Bodin et al., 1997; Huang, 2003, Suryanto et al., 2006; Wassermann et al., 2009). The rotational components of ground motion have also been estimated theoretically, using kinematic source models (Bouchon and Aki, 1982; Wang et al., 2009) and linear elastodynamic theory of wave propagation in elastic solids (Lee and Trifunac, 1985, 1987).

Measuring rotational motions

The general motion of the particles or a small volume in a solid body can be divided into three parts: translation (along the X-, Y-, and Z-axes), rotation (about the X- Y-, and Z-axes), and strain (six components). Figure 1b shows the axes in a Cartesian coordinate system for *translational velocity* measured by the usual seismometers in seismology, and Figure 1c shows the corresponding axes of *rotation rate* measured by rotational sensors (Evans, 2009). These are "body attached" coordinates, those that a seismic instrument would measure at a given instant as the sensors move and rotate through space. Converting an extended record of these body-fixed motions to recover motions in an Earth-fixed, quasi-inertial coordinate system has been performed for decades in "strapped down" inertial navigation systems such as those attached to a moving airplane. Lin et al. (2010) introduce these equations into seismology and earthquake engineering for recovering inertial-frame ground and structure motions.

Rotational ground motions can be measured directly by gyroscopic sensors or inferred indirectly from an array of translational sensors. According to Cochard et al. (2006), in a linear elastic medium the displacement $\mathbf{u}$ of a point $\mathbf{x}$ is related to a neighboring point $\mathbf{x} + \delta\mathbf{x}$ by

$$\mathbf{u}(\mathbf{x} + \delta\mathbf{x}) = \mathbf{u}(\mathbf{x}) + \boldsymbol{\varepsilon}\,\delta\mathbf{x} + \omega \times \delta\mathbf{x} \tag{1}$$

where $\boldsymbol{\varepsilon}$ is the strain tensor and

$$\omega = {}^1\!/_2 \nabla \times \mathbf{u}(\mathbf{x}) \tag{2}$$

is a pseudo-vector representing the infinitesimal angle of rigid rotation generated by the disturbance. The three components of rotation about the X-axis, Y-axis, and Z-axis are given by the following equations for such infinitesimal motions:

$$\begin{aligned} \omega_x &= {}^1\!/_2\left(\partial u_z/\partial y - \partial u_y/\partial z\right), \\ \omega_y &= {}^1\!/_2\left(\partial u_x/\partial z - \partial u_z/\partial x\right), \\ \omega_z &= {}^1\!/_2\left(\partial u_y/\partial x - \partial u_x/\partial y\right) \end{aligned} \tag{3}$$

Therefore, rigid rotations can be observed by: (1) an array of translational seismometers indirectly for "cord" rotations associated with long wave lengths by assuming that contamination of translational signals by rotational motions is small, and that the linear elasticity theory is valid (e.g., Spudich and Fletcher, 2008), or (2) rotational sensors directly for "point" body-fixed rotations (e.g., Lee et al., 2009b; Lin et al., 2010).

In the past decade, rotational motions from small local earthquakes to large teleseisms were successfully recorded by sensitive rotational sensors in several countries (e.g., Takeo, 1998; McLeod et al., 1998; Igel et al., 2005, 2007; Suryanto et al., 2006; Cochard et al., 2006). In particular, the application of Sagnac interferometry in large ring laser gyros provided greatly improved sensitivity to rotations at teleseismic distances and showed that they are a good match to those estimated from linear elastic wave theory. Such motions provide additional observations that – combined with translations – lead to new approaches to the seismic inverse problem (Bernauer et al., 2009; Fichtner and Igel, 2009). Recently Kurrle et al. (2010b) reported the first observations of Earth's free oscillations using ring laser recordings, and opened up potential applications of rotational seismology at long periods.

In contrast, strong-motion observations near the source in both Japan and Taiwan showed that the amplitudes of these rotations can be one to two orders of magnitude greater than that expected from linear elasticity theory (e.g., Takeo, 1998; Lee et al., 2009b). Theoretical work suggests that in shallow granular or cracked continua (e.g., weathered rock at or near Earth's surface), asymmetries of the stress and strain fields can create rotations separate from and larger than those predicted by classical elastodynamic theory (e.g., Teisseyre et al., 2006, 2008). Unlike the traditional fault-slip model, Knopoff and Chen (2009) consider the case for faulting that takes place on a fault of finite thickness. They show that there is an additional single-couple term in the body-force equivalence and additional terms in the far-field displacement. They also show that the single-couple equivalent does not violate the principles of Newtonian mechanics because the torque imbalance in the single-couple is counterbalanced by rotations within the fault zone, with torque waves being radiated.

Large ring laser gyros

An unexpected advance in studying rotational ground motions came from a different field of geophysics. Recent developments of highly sensitive ring laser gyroscopes to monitor the Earth's rotation also yield valuable data on rotational motions from large teleseismic events. The most important property that makes such rotation sensors useful for seismology is its very low noise floor and high sensitivity to rotational motions and its insensitivity to translational and cross-rotational motions. The rotation rates expected and observed in seismology range from the order of 10^{-1} rad/s (e.g., Nigbor, 1994; Trifunac, 2009) near seismic sources down to order 10^{-11} rad/s for large earthquakes at teleseismic distances (e.g., Igel et al., 2005, 2007). This range spans at least 10 orders of magnitude, 200 dB, much as for translational motions, and it is unlikely that one instrument or one instrumental technology will be capable of providing accurate measurements over such a large range of amplitudes. Ring laser technology is currently the most promising approach to recording the small rotational motions induced by teleseisms, but the primary drawback is its very high cost.

Ring lasers detect the Sagnac beat frequency of two counter-propagating laser beams (Stedman, 1997; and Figure 2c). These active interferometers generally form triangular or square closed loops several meters across and are evacuated. If this instrument is rotating on a platform with respect to inertial space, the effective cavity length between co-rotating and counter-rotating laser cavity differs and one observes frequency splitting resulting in a beat frequency. This beat frequency δf is directly proportional to the rotation rate $\boldsymbol{\Omega}$ around the surface normal $\mathbf{n}$ of the ring laser system, as given by the Sagnac equation:

$$\delta f = \frac{4A}{\lambda P}\mathbf{n}\cdot\boldsymbol{\Omega}\,, \tag{4}$$

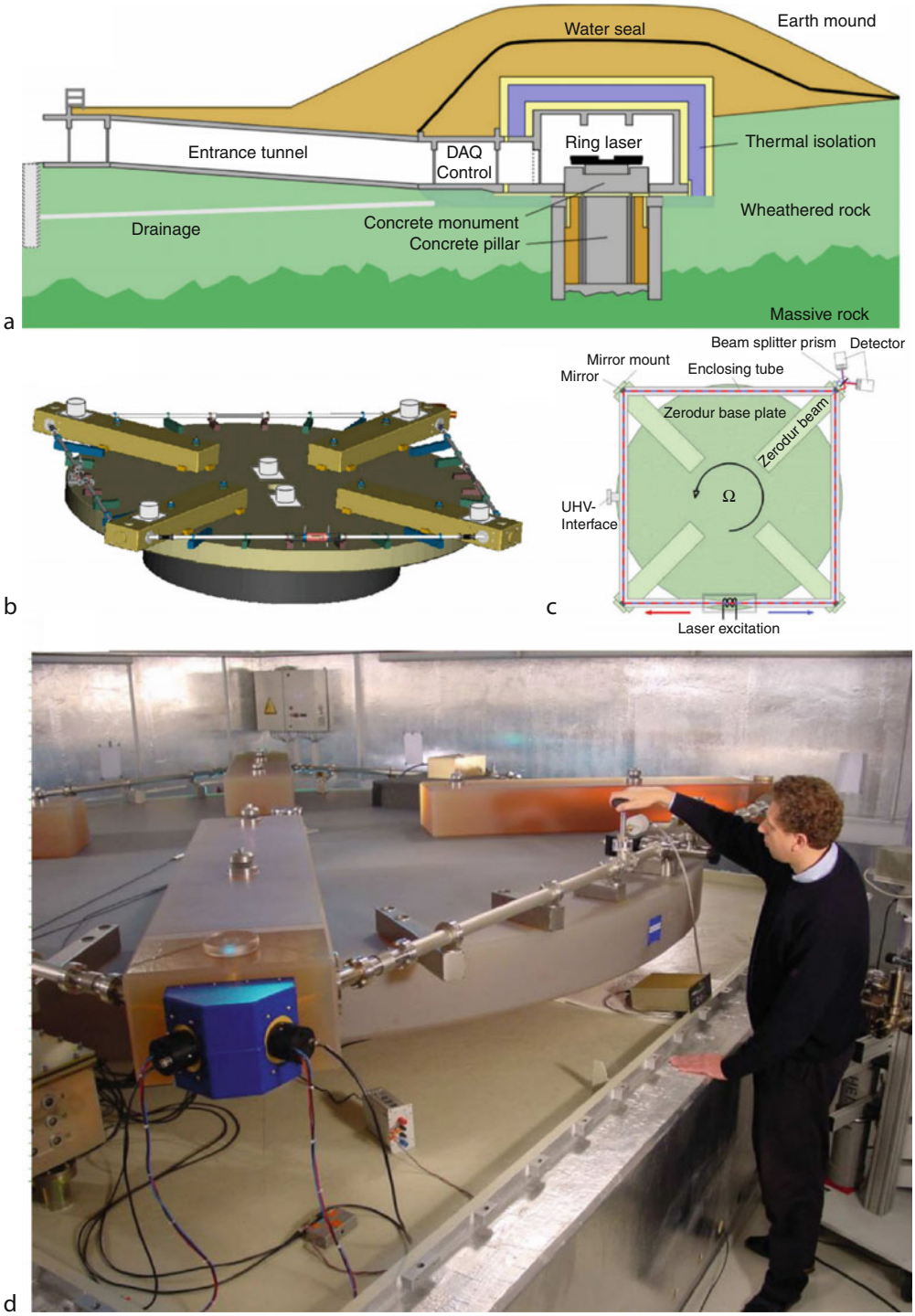
where P is the perimeter of the instrument, A its area, and λ the laser wavelength. This equation has three contributions that influence the beat frequency δf: (1) variations in the scale factor $(4A/\lambda P)$ have to be avoided by making the instrument mechanically as rigid and stable as possible, (2) changes in orientation $\mathbf{n}$ (tilting relative to Earth's rotation axis) enter the beat frequency via the inner product, and (3) variations in $\boldsymbol{\Omega}$ (e.g., changes in Earth's rotation rate and seismically induced rotations). Thankfully, the dominant contribution to δf is $\boldsymbol{\Omega}$. Note that translations do not contribute to the Sagnac frequency unless they affect P or A in some indirect manner.

Ring lasers are sensitive to rotations only, assuming stable ring geometry and lasing. However, for co-seismic observations at the Earth's surface the horizontal components of rotation (i.e., tilts) will contribute to the vertical component of rotation rate. As shown by Pham et al. (2009), the tilt-coupling effect is several orders of magnitude below the level of the earthquake-induced rotational signal unless one is very close to the source (where sensitive ring lasers would not be the appropriate technology).

At present, there are ring laser gyros capable of measuring rotation (induced by small local earthquakes or distant large teleseisms) at four sites: (1) Cashmere cavern, Christchurch, New Zealand (McLeod et al., 1998); (2) Wettzell, Germany (Schreiber et al., 2005); (3) Conway, Arkansas (Dunn et al., 2009); and (4) Piñon Flat, California (Schreiber et al., 2009a).

G Ring laser and recording teleseisms

Since 2001, the "G Ring" laser (capable of measuring rotation rate of about 10^{-10} rad/s) has been operating at the primary geodetic station (Fundamentalstation) at Wettzell, in Bavaria, (http://www.fs.wettzell.de/). A cross-sectional view of the site of the G Ring laser is shown in Figure 2a. The instrument is resting on a polished granite table (Figure 2b) embedded in a 90-t concrete monument. As shown in Figure 2a, the monument is



Seismology, Rotational, Figure 2 G Ring laser gyro at the Wettzell Superstation, Germany. (**a**) Cross-sectional view of the instrument site. (**b**) Instrument resting on a granite table. (**c**) Schematic drawing. (**d**) Photo of G Ring laser gyro with its designer, Ulli Schreiber.

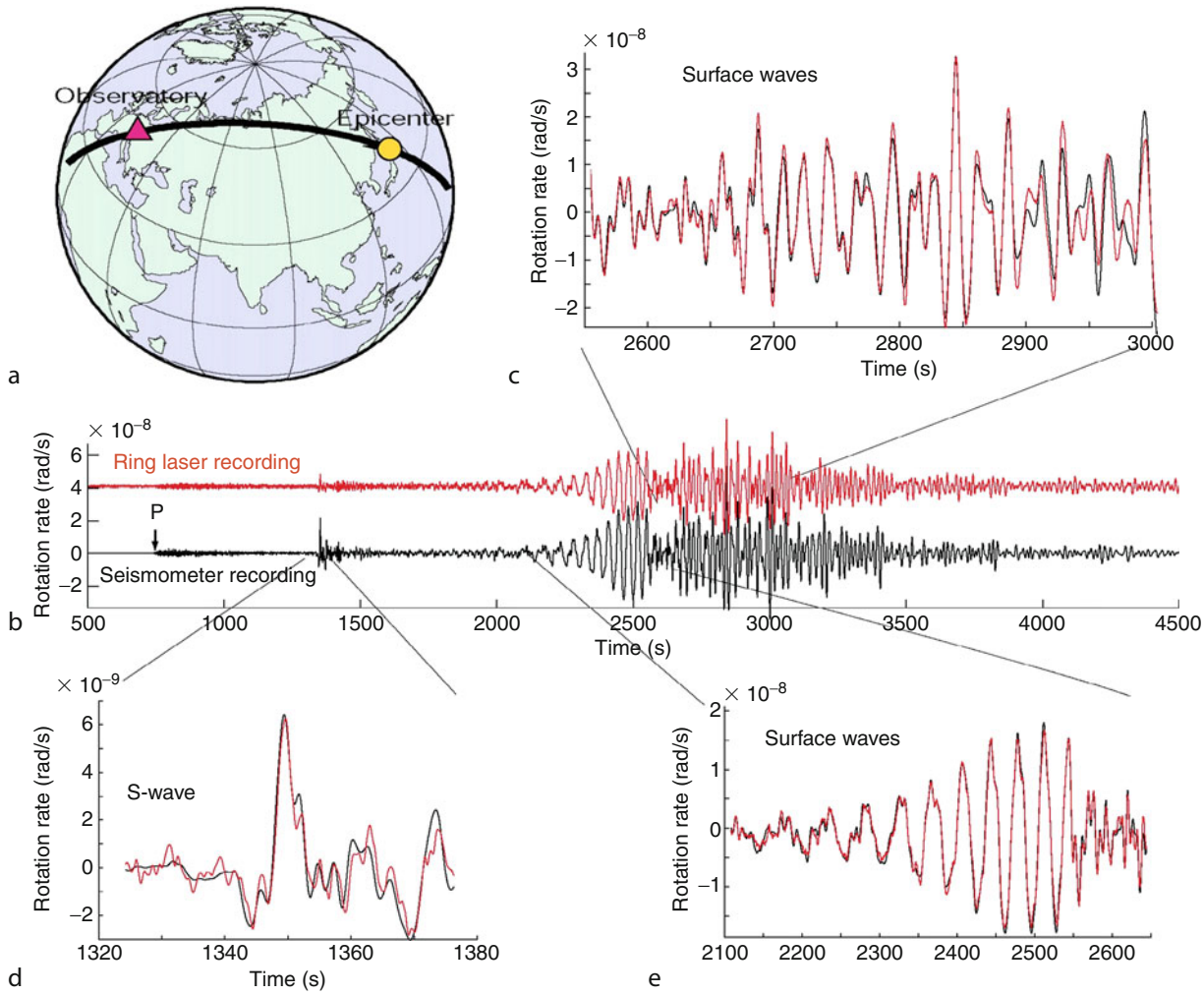
attached to a massive 2.7-m diameter concrete pillar and this is founded on crystalline bedrock 10 m below. A system of concrete rings and isolation material shields the monument and pillar from adjacent weathered rock to eliminate its deformation and heat-flow contributions. The G Ring laser is protected against external influences by a subsurface installation with passive thermal stability provided by a 2-m layer alternating between Styrofoam

and wet clay, this beneath a 4-m soil mound. A lateral entrance tunnel with five isolating doors and a separate control room minimize thermal perturbations during maintenance. After 2 years of thermal adaptation, the average temperature reached 12.2°C with seasonal variations of less than 0.6°C. Figure 2c shows the schematic drawing of instrument, and Figure 2d is a photo of the G Ring laser with its designer, Ulli Schreiber.

Figure 3 is a comparison of direct point measurements of ground rotations around a vertical axis (red lines) to transverse accelerations (black lines, converted to rotation rate for each time window) for the M8.1 Tokachi-oki earthquake, September 25, 2003 (Igel et al., 2005). Figure 3a is a schematic view of the great-circle-path through the epicenter in Hokkaido, Japan, and the observatory in Wettzell, Germany. 3b–e show the superposition of the rotation rate derived from transverse translations (black) and measured directly (red) for various time windows: Figure 3b the complete signal, Figure 3c the latter part of the surface wave train, Figure 3d the direct *S*-wave arrival, and Figure 3e for the initial part of the surface wave train. These results confirm the expectation from linear elasticity that the waveforms of transverse acceleration and rotation rate (around the vertical axis) should be identical assuming plane harmonic waves. Information on subsurface structure is contained in the ratio between the corresponding motion amplitudes.

Strong-motion inertial angular sensors

In aerospace, automotive, and mechanical engineering, smaller rotational-motion sensors are common and generically known as gyroscopic or inertial angular sensors. Nigbor (1994) used a MEMS-based Coriolis rotation-rate sensor to measure the rotational components of strong



Seismology, Rotational, Figure 3 Comparison of direct measurements of ground rotational motions around a vertical axis (*red lines*) with transverse accelerations (*black lines*, converted to rotation rate for each time window) for the M8.1 Tokachi-oki earthquake, September 25, 2003.

ground motions close to a large chemical explosion. Similar sensors were used by Takeo (1998) to measure rotational motions from an earthquake swarm ~3 km away. However, such sensors do not have the sensitivity to record rotations from small local earthquakes (magnitude ~4) at distances of tens of kilometers.

The eentec™ model R-1™ rotational seismometer is the first modestly priced sensor capable of recording small earthquakes at distances up to several tens of kilometers. It uses electrochemical technology in which the motion of an electrolytic fluid inside a torus is sensed electronically, yielding a voltage signal proportional to rotational velocity. Nigbor et al. (2009) carried out extensive tests of commercial rotational sensors and concluded that the R-1 sensor generally meets the specifications given by the manufacturer but that clip level and frequency response vary from those specifications and between individual channels enough that more detailed calibrations are warranted for each unit. A typical transfer function for the R-1 can be found at the manufacturer's website (http://www.eentec.com/). The instrument response is roughly "flat" from 0.1 to 20 Hz, and its self noise (rms) is <10 μrad/s over the same frequency band.

The R-1 rotational seismometers successfully recorded several hundred local earthquakes and two explosions in Taiwan (Lee et al., 2009b). Figure 4a shows the instruments deployed at station HGSD in eastern Taiwan (Liu et al., 2009). The top frame is a schematic drawing of the various seismic, geodetic, and strain instruments there. The bottom frame shows the subset of the instruments deployed in the shallow vault at the left hand side of the upper drawing; these include a datalogger (Quanterra Q330), an accelerometer (Kinemetrics Episensor), a six-channel digital accelerograph (Kinemetrics K2 with an external rotational seismometer, R-1 by eentec), and a short-period seismometer (Mark Products L-4A). The K2+R-1 instrument is at the left hand side, and the yellow-color box is the R-1 rotational seismometer.

The largest peak rotational rate recorded at the HGSD station (to early 2008) is from a M_w 5.1 earthquake at a hypocentral distance of 51 km at 13:40 UTC, 23 July 2007. Figure 4c shows the amplitudes and spectra of translational accelerations recorded by the K2's accelerometer. The peak ground acceleration was 0.47 m/s^2, and the two horizontal components have much higher amplitude than the vertical. Figure 4d shows the amplitudes and spectra for rotational rates recorded from its external R-1 seismometer. The peak rotational rate was 0.63×10^{-3} rad/s for the vertical component, much more than that for the horizontal components. The spectra in Figure 4c show that the dominant frequency band in ground acceleration is about 2–5 Hz (horizontal components) while those in Figure 4d show that the dominant frequency band in ground rotation rate is about 2.5–5.5 Hz for the vertical component. Other studies report observations with the R-1 sensor and compare their point measurements of rotation to array-derived area rotations (e.g., Wassermann et al., 2009).

Discussions

Many authors have emphasized the benefits of studying rotational motions (e.g., Twiss et al., 1993; Spudich et al., 1995; Takeo and Ito, 1997; Teisseyre et al., 2006; Trifunac, 2006, 2009; Igel et al., 2007; and Fichtner and Igel, 2009). We discuss some basic issues briefly.

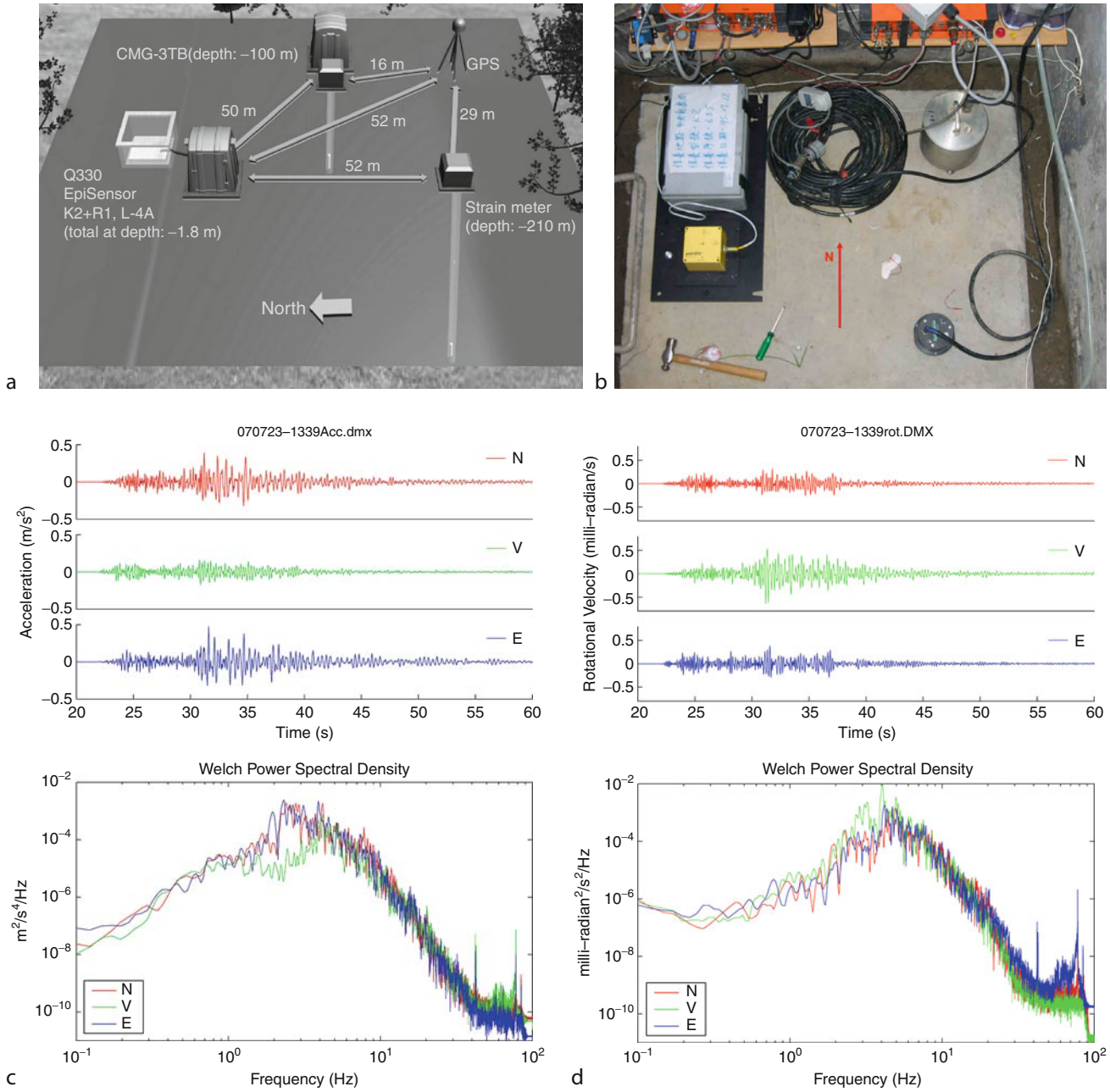
Linear and nonlinear elasticity

Real materials of the Earth are heterogeneous, anisotropic, and nonlinear, especially in the damage zone surrounding faults and in poorly consolidated sediments, soil, and weathered, fractured rock just beneath seismic instruments, particularly typically installed strong-motion instruments. In the presence of significant nonlinearity we are forced to consider the mechanics of chaos (Trifunac, 2009) and to interpret such complexities must record both the rotational and translational components of strong motion.

Seismology is primarily based on the linear elasticity theory, which is applicable to simple homogeneous materials under infinitesimal strain. "Cord" rotation is defined as the curl of the displacement field in Equation 2, and in the classical elasticity theory, the rotational components of motion are contained in the S waves. Meanwhile, continuum mechanics has advanced far beyond the classical theory. In particular, the elasticity theory of the Cosserat brothers (Cosserat and Cosserat, 1909) incorporates (1) a local rotation of continuum particles as well as the translational motion assumed in classical theory, and (2) a couple stress (a torque per unit area) as well as the force stress (force per unit area). In the constitutive equation of classical elasticity theory there are two independent elastic constants while in Cosserat elastic theory there are six or more elastic constants. Pujol (2009) provides a tutorial on rotations in the theories of finite deformation and micropolar (Cosserat) elasticity. Twiss (2009) derives an objective asymmetric micropolar moment tensor from a discrete-block model for a deforming granular material. He also investigates seismogenic deformation associated with volumes of distributed seismicity in three different geographic areas, and finds support in the micropolar model for the effects of a granular substructure on the characteristics of seismic focal mechanisms.

Near-field seismology

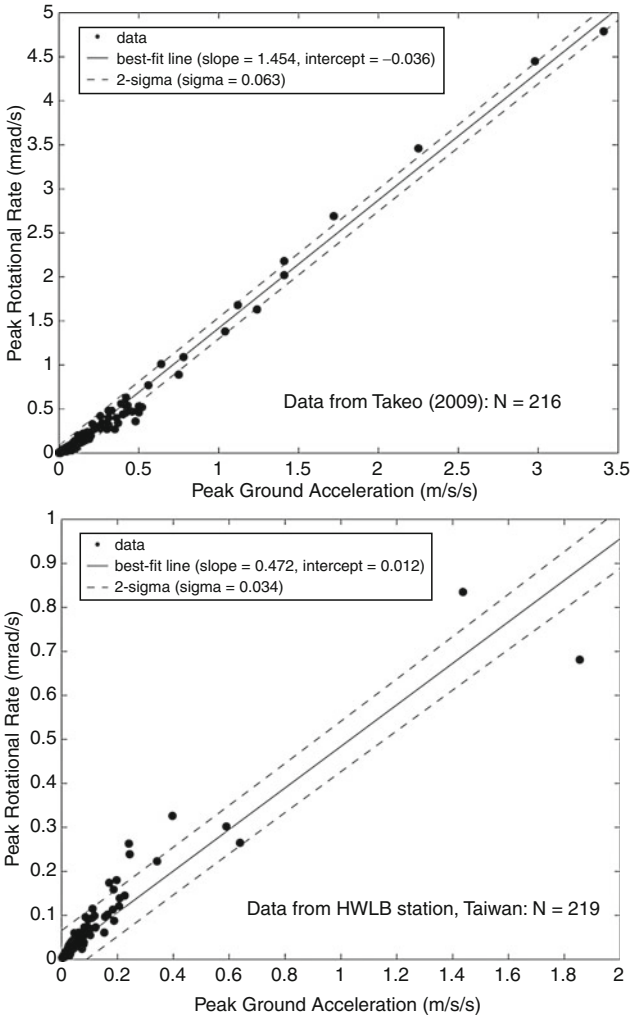
Although the observed rotational motions agree well with the classical elasticity theory for teleseisms, it is not the case for local earthquakes. As first noted by Takeo (1998) and confirmed by Lee et al. (2009b), observed rotational rates from local earthquakes are much larger than those predicted from the classical elasticity theory. For example, Bouchon and Aki (1982) obtained theoretically a maximum rotational rate of 1.5×10^{-3} rad/s for a magnitude 6.5 earthquake, whereas a maximum rotational rate of $>1 \times 10^{-3}$ rad/s for several magnitude 4.5–5.5 earthquakes had been observed in Japan and Taiwan. Takeo (1998) reported the largest rotational rate



Seismology, Rotational, Figure 4 (**a**) Schematic of instrument layout at station HGSD in Taiwan. (**b**) Photo showing the instrument vault containing a K-2 accelerograph, an R-1 rotational seismometer (*yellow*), an EpiSensor accelerometer, and an L-4A velocity seismometer. (**c**) Recorded translational accelerations (*top*) and their spectra (*bottom*) from an M_w 5.1 earthquake 51 km from this site. (**d**) Recorded rotational rates (top) and their spectra (*bottom*) from the same earthquake.

was 2.6×10^{-2} rad/s around the north-south axis from the second largest earthquake (magnitude 5.2) during the 1997 swarm (at 14:09 UTC, 3 March), east of Cape Kawana, offshore Ito, Japan (about 3 km away). As of the end of 2009, the largest rotational rate recorded at the HWLB station (Hualien, Taiwan) was 2.58×10^{-3} rad/s around the east–west axis from a M_w 6.4 earthquake offshore (at UTC 13:02, 19 December 2009) at a hypocentral distance of about 49 km. The peak rotational rate is 1.57×10^{-3} rad/s around the north-south axis, and 0.68×10^{-3} rad/s around the vertical axis. The corresponding peak ground acceleration is 1.16, 1.85, and 0.50 m/s^2 for the east–west, north–south, and vertical component, respectively.

Figure 5 shows the *vertical* peak rotational rate versus *horizontal* peak ground acceleration for the earthquake data set from Takeo (2009) (top frame), and for the earthquake data set (22 August, 2008 to 25 December, 2009) recorded at the HWLB station, Taiwan (bottom frame), where N = number of data points. As noted by Takeo (1998, 2009), there is a reasonable linear relationship in such plots. The data scatter is larger in the Taiwan data than the Takeo (2009) data. This may due to the fact that the earthquake sources for the Takeo (2009) data were from a nearby offshore swarm, whereas the Taiwan data were recorded from many different earthquake sources. Several authors noted similar linear relationships before (see e.g., Spudich and Fletcher, 2008; Stupazzini et al., 2009; Takeo, 2009; and Wang et al., 2009). In particular,



Seismology, Rotational, Figure 5 Vertical peak rotational rate (PRR) versus horizontal peak ground acceleration (PGA) for the earthquake data set from Takeo (2009) (top frame), and for the earthquake data set (22 August, 2008 to 25 December, 2009) recorded at the HWLB station, Taiwan (bottom frame). *N* number of data points.

Takeo (2009) showed "*a linear correlation between the maximum rotational displacements around vertical axis and the maximum* [ground] *velocities*." The two plots in Figure 5 are equivalent to Takeo's linear relationship, without performing the integration for the measured rotational rate and ground acceleration to obtain rotational displacements and ground velocities. The linear slope in Figure 5 has the dimension unit of s/km, or unit for slowness. Spudich and Fletcher (2008) interpreted this "slowness" as the inverse of an "apparent velocity," characterizing the seismic wavefield beneath the recording station.

Spudich and Fletcher (2008) computed peak values of ground strain, torsions, and tilts for the 2004 Parkfield earthquake (M_w 6.0) and four aftershocks (M_w 4.7–5.1) using the data recorded by an array of accelerators. Takeo (2009) noted that his observations of peak rotation values are about 100 times larger for earthquakes with similar magnitudes and distances, and proposed the following explanations: the different spatial scale of rotational motion by a single-point gyro measurement and by an array observation, the effect of topography, and the difference of the degree of maturation between the San Andreas fault and the swarm volume of offshore Ito. Two arrays of both rotational and translational sensors have been deployed by Wu et al. (2009) in Taiwan to study this discrepancy problem, and hopeful a resolution may be found after recording sufficient numbers of earthquakes.

Using explosions to study rotational motions

Since a large earthquake occurring near a station is rare, explosions have been used to study rotational motions by several pioneers (e.g., Graizer, 1991; Nigbor, 1994). Lin et al. (2009) deployed an array of 8 triaxial rotational sensors, 13 triaxial accelerometers, and 12 six-channel, 24-bit dataloggers with GPS time receivers to record two explosions in northeastern Taiwan. These instruments were installed at about 250 m (1 station), 500 m (11 stations), and 600 m (1 station) from the explosions. The 11 stations form a "Center Array" with station spacing of about 5 m. The code name for the first shot with 3,000 kg explosives is "N3P," and that for the second shot with 750 kg explosives is "N3." Although the N3P shot used four times larger amounts of explosives than that used for the N3 shot, the peak ground translational acceleration and rotational velocity at the 13 station sites from the N3P shot are only about 1.5 times larger than that for the N3 shot. Large variations (by tens of %) of translational accelerations and rotational velocities were observed at the very small Center Array. The largest peak rotational rate was observed for the horizontal transverse component: 2.74 $\times$ 10^{-3} and 1.75 $\times$ 10^{-3} rad/s at a distance of 254 m from the N3P and N3 shots, respectively.

The acceleration data from these two explosions were used by Langston et al. (2009) to compute acceleration spatial gradients, horizontal strains and horizontal rotation, and to perform a gradiometric analysis of the strong ground motion wave train. The analysis yields a complex, frequency-dependent view of the nature of seismic wave

propagation over short propagation distances that imply significant lateral velocity changes in the near-surface crustal structure. Areal strain and rotation about the vertical have equal amplitudes and suggest significant wave scattering within the confines of the river valley where the experiment was performed and/or significant departure from an axisymmetric explosion source. Gradiometry shows that the P wave arrives at the array 35° off-azimuth clockwise from the straight-line path and appears to have been refracted from the northern side of the valley. Chi et al. (2011) successfully recovered the first order features of vertical rotational rate ground motions from the translational velocity waveforms in the 0.5–20 Hz bandwidth using the software of Spudich and Fletcher (2009), and strain as large as 10^{-4} was deduced. To fulfill the uniform rotation assumption in the linear elasticity theory, it is necessary to use a small-aperture array. However, inverting data from an array of small spatial dimension requires accurate waveforms of high signal to noise ratio and high sampling rates; since waveforms from adjacent stations are very similar, small noises have strong influence on displacement gradients due to the small station spacing. The Lin et al. (2009) recordings were limited to 200 samples per second, and much higher sampling rate will be necessary to record explosions in the near field.

Processing collocated measurements of translations and rotations

Processing collocated observations of rotation and translation is routinely performed in the inertial navigation units of aircraft and other vehicles. A similar analysis is possible for various combinations of strain components, rotations, and translations. With the exception of velocity–strain combinations (e.g., Gomberg and Agnew, 1996) this terrain was largely unexplored until the work of Lin et al. (2010), who have demonstrated an appropriate set of these equations for earthquake engineering and seismology to recover inertial-frame displacements and rotations. Further, it is already apparent that rotational motions provide useful additional analysis opportunities, simply put, that more data at a site yield more results.

Phase Velocities and propagation directions. A simple calculation for non-dispersive linear-elastic plane waves with transverse polarization shows that the ratio of transverse acceleration to vertical-axis rotation rate is proportional to local phase velocity. This result implies that information on subsurface velocity structure (otherwise only accessible through seismic array measurements and combined analyses) is contained in any single-point measurement that includes rotational sensors. It has been shown that such ratio-derived phase velocities agree with velocities predicted by theory (Igel et al., 2005, Kurrle et al., 2010a). In a recent theoretical study based on full ray theory for Love waves (normal mode summation), Ferreira and Igel (2009) demonstrated that the Love wave dispersion relation also can be obtained by taking the spectral ratio of transverse acceleration to vertical-axis rotation rate. This result implies that seismic shear wave tomography is possible without requiring sub-arrays to determine local mean phase velocities. Information on the direction of propagation also is contained in the azimuth-dependent phase fit between rotations and translations; this fit is optimal in the direction of propagation, from which back azimuths can be estimated to within a few degrees (Igel et al., 2007). Linking observational translations, strains, and rotations together also is advocated by Langston (2007) to yield a snapshot of the wavefield including wave direction, slownesses, and radial/azimuthal amplitude gradients independently at each such station.

Toward a new kind of tomography. The possibility of deriving local dispersion relations from single-station records leads to the question of what subsurface volume one resolves and to what depth velocity perturbations can be recovered. The method of choice to answer this type of question is the adjoint method (Fichtner and Igel, 2009), with which sensitivity kernels (first Fresnel zones) can be calculated to indicate the volume in which the observable (typically travel times) is sensitive to structural perturbations. Fichtner and Igel (2009) introduced a new observable quantity – apparent shear wave velocity – which is a time-windowed ratio of the moduli of translational velocity and rotation angle. It turns out that the sensitivity near the source vanishes, leading to a new type of kernel that shows high sensitivity only in the vicinity of the receiver and in a somewhat smaller portion of that volume than the kernels of translational motions alone. This result implies that a tomographic inversion for near-receiver structures based on rotations and translations is possible and further highlights the potential of rotation measurements. Synthetic tomographic inversions are given in Bernauer et al. (2009).

Scattering properties of the crust: Partitioning of P and S waves. The partitioning of *P* and *S* energy and stabilizing the ratio between the two is an important constraint on the scattering properties of a medium. Igel et al. (2007) discovered surprisingly great rotational energy in a time window prior to teleseismic *S*, containing the *P*-coda. Detailed analysis of the signals and modeling of wave propagation through three-dimensional random media demonstrate that these signals can be explained with *P–SH* scattering in the crust with scatterers of very roughly 5-km correlation length and rms perturbation amplitude of 5% (better constrained). This result further illustrates the efficacy of rotation measurements in their own right, for example, as a filter for SH type motion, as noted by Takeo and Ito (1997).

Similar processing steps will be possible for the horizontal components of rotation and the corresponding components of translation, both to reduce the effects of tilt on the horizontal translational sensors and to extract new information. It is conceivable that the combination of these various components might lead to tight constraints on near-receiver structure, results otherwise only available from array measurements.

Conclusion

Seismology has been very successful in the *far field* because large (magnitude >6) earthquakes occur every week somewhere on Earth, and because classical elasticity

theory works very well for interpreting the recorded *translational* motions at large distances. Because of this success and limited instrumentation options, most funding for earthquake monitoring historically has gone into global and regional seismic networks using only translational seismometers. However, to improve our understandings of damaging earthquakes we must also deploy rotational and translational instruments in the *near field* of active faults where potentially damaging earthquakes (magnitude > 6.5) occur *infrequently*. For strong-motion seismology and engineering, this is a risky business because a damaging earthquake on any given fault may not take place for hundreds of years. Recording ground motions in the near field would require extensive seismic instrumentation along some well-chosen active faults and luck.

Ring laser observations at Wettzell, Germany, and at Piñon Flat, California, demonstrated consistent measurements of rotational ground motions in the *far field*. So far this success can only be demonstrated with one component of rotation. The high cost of present high-precision ring laser gyros makes widespread deployment unlikely. Less expensive and/or less sensitive alternatives are now being pursued by five academic groups (Cowsik et al., 2009; Dunn et al., 2009; Jedlička et al., 2009; Schreiber et al., 2009b; and Takamori et al., 2009). As of the end of 2009, only Taiwan has a modest program (Lee et al., 2009b) to monitor both translational and rotational ground motions from local and regional earthquakes at several permanent seismic stations, as well as by two arrays in a building and a nearby free-field site. These two arrays are designed to "capture" a repeat of the 1906 Meishan earthquake (magnitude 7.1) in the *near field* with both translational and rotational instruments (Wu et al., 2009).

Based on the developments described in the BSSA Special Issue on rotational seismology and engineering (Lee et al., 2009a), observation, analysis, and interpretations of both rotational and translational ground motions will soon play a significant role in seismology and earthquake engineering. An international working group on rotational seismology (IWGoRS) was organized in 2006 to promote investigations of rotational motions and their implications, and for sharing experience, data, software, and results in an open Web-based environment (Todorovska et al., 2008). Anyone can join IWGoRS at http://www.rotational-seismology.org, subscribe to the mailing list, and contribute to the content (publications, data, links, etc.). Rotational seismology is also of interest to physicists using Earth-based observatories for detecting Einstein's gravitational waves (e.g., Lantz et al., 2009) because they must correct for the underlying Earth motion.

Bibliography

Bernauer, M., Fichtner, A., and Igel, H., 2009. Inferring Earth structure from combined measurements of rotational and translational ground motions. *Geophysics*, **74**(6), WCD41–WCD47, doi:10.1190/1.3211110.

Bodin, P., Gomberg, J., Singh, S. K., and Santoyo, M., 1997. Dynamic deformations of shallow sediments in the valley of Mexico, part I: three dimensional strains and rotations recorded on a seismic array. *Bulletin of the Seismological Society of America*, **87**, 528–539.

Bouchon, M., and Aki, K., 1982. Strain, tilt, and rotation associated with strong ground motion in the vicinity of earthquake faults. *Bulletin of the Seismological Society of America*, **72**, 1717–1738.

Chi, W. C., Lee, W. H. K., Lin, C. J., and Liu, C. C., 2011. Inverting horizontal translational for vertical rotational ground motions: a test using data from explosions. [Submitted to the *Bulletin of the Seismological Society of America*, July, 2010.]

Cochard, A., Igel, H., Schuberth, B., Suryanto, W., Velikoseltsev, A., Schreiber, U., Wassermann, J., Scherbaum, F., and Vollmer, D., 2006. Rotational motions in seismology: theory, observation, simulation. In Teisseyre, R., Takeo, M., and Majewski, E. (eds.), *Earthquake Source Asymmetry, Structural Media and Rotation Effects*. Heidelberg: Springer, pp. 391–411.

Cosserat, E., and Cosserat, F., 1909. *Théorie des Corps Déformables*. Paris: Hermann (available from the Cornell University Library Digital Collections) (in French).

Cowsik, R., Madziwa-Nussinov, T., Wagoner, K., Wiens, D., and Wysession, M., 2009. Performance characteristics of a rotational seismometer for near-field and engineering applications. *Bulletin of the Seismological Society of America*, **99**(2B), 1181–1189.

Droste, Z., and Teisseyre, R., 1976. Rotational and displacemental components of ground motion as deduced from data of the azimuth system of seismograph. Publication of the Institute of Geophysics. *Polish Academy of Sciences*, **97**, 157–167.

Dunn, R. W., Mahdi, H. H., and Al-Shukri, H. J., 2009. Design of a relatively inexpensive ring laser seismic detector. *Bulletin of the Seismological Society of America*, **99**(2B), 1437–1442.

Evans, J. R., 2009. Suggested notation conventions for rotational seismology. *Bulletin of the Seismological Society of America*, **99**(2B), 1429–1436.

Farrell, W. E., 1969. A gyroscope seismometer: measurements during the Borrego earthquake. *Bulletin of the Seismological Society of America*, **59**, 1239–1245.

Ferrari, G., 2006. Note on the historical rotation seismographs. In Teisseyre, R., Takeo, M., and Majewski, E. (eds.), *Earthquake Source Asymmetry, Structural Media and Rotation Effects*. Heidelberg: Springer, pp. 367–376.

Ferreira, A., and Igel, H., 2009. Rotational motions of seismic surface waves in a laterally heterogeneous Earth. *Bulletin of the Seismological Society of America*, **99**(2B), 1073–1075.

Fichtner, A., and Igel, H., 2009. Sensitivity densities for rotational ground-motion measurements. *Bulletin of the Seismological Society of America*, **99**(2B), 1302–1314.

Galitzin, B. B., 1912. *Lectures on Seismometry*. St. Petersburg: Russian Academy of Sciences. in Russian.

Gomberg, J., and Agnew, D., 1996. The accuracy of seismic estimates of dynamic strains: an evaluation using strainmeter and seismometer data from Piñon Flat Observatory, California. *Bulletin of the Seismological Society of America*, **86**, 212–220.

Graizer, V. M., 1991. Inertial seismometry methods. Izvestiya of the USSR Academy of Sciences. *Physics of the Solid Earth*, **27**(1), 51–61.

Huang, B. S., 2003. Ground rotational motions of the 1991 Chi-Chi, Taiwan, earthquake as inferred from dense array observations. *Geophysical Research Letters*, **30**(6), 1307–1310.

Igel, H., Schreiber, U., Flaws, A., Schuberth, B., Velikoseltsev, A., and Cochard, A., 2005. Rotational motions induced by the M8.1 Tokachi-oki earthquake, September 25, 2003. *Geophysical Research Letters*, **32**, L08309, doi:10.1029/2004GL022336.

Igel, H., Cochard, A., Wassermann, J., Schreiber, U., Velikoseltsev, A., and Pham, N. D., 2007. Broadband observations of rotational ground motions. *Geophysical Journal International*, **168**, 182–197.

Jedlička, P., Buben, J., and Kozák, J., 2009. Strong-motion fluid rotation seismograph. *Bulletin of the Seismological Society of America*, **99**(2B), 1443–1448.

Kharin, D. A., and Simonov, L. I., 1969. VBPP seismometer for separate registration of translational motion and rotations. *Seismic Instruments*, **5**, 51–66 (in Russian).

Knopoff, L., and Chen, Y. T., 2009. Single-couple component of far-field radiation from dynamical fractures. *Bulletin of the Seismological Society of America*, **99**(2B), 1091–1102.

Kozák, J. T., 2009. Tutorial on earthquake rotational effects: historical examples. *Bulletin of the Seismological Society of America*, **99**(2B), 998–1010.

Kurrle, D., Igel, H., Ferreira, A. M. G., Wassermann, J., and Schreiber, U., 2010a. Can we estimate local Love wave dispersion properties from collocated amplitude measurements of translations and rotations? *Geophysical Research Letters*, **37**, L04307, doi:10.1029/2009GL042215.

Kurrle, D., Igel, H., Ferreira, A. M. G., Wassermann, J., and Schreiber, U., 2010b. First observation of rotational motions from Earth's free oscillations, *Geophysical Research Abstracts*, **12**, EGU2010-3281.

Langston, C. A., 2007. Wave gradiometry in two dimensions. *Bulletin of the Seismological Society of America*, **97**, 401–416.

Langston, C. A., Lee, W. H. K., Lin, C. J., and Liu, C. C., 2009. Seismic-wave strain, rotation, and gradiometry for the 4 March 2008 TAIGER explosions. *Bulletin of the Seismological Society of America*, **99**(2B), 1287–1301.

Lantz, B., Schofield, R., O'Reilly, B., Clark, D. E., and DeBra, D., 2009. Review: requirements for a ground rotation sensor to improve advanced LIGO. *Bulletin of the Seismological Society of America*, **99**(2B), 980–989.

Lee, V. W., and Trifunac, M. D., 1985. Torsional accelerograms. *Soil Dynamics and Earthquake Engineering*, **4**(3), 132–139.

Lee, V. W., and Trifunac, M. D., 1987. Rocking strong earthquake accelerations. *Soil Dynamics and Earthquake Engineering*, **6**(2), 75–89.

Lee, W. H. K., Celebi, M., Igel, H., and Todorovska, M. I., 2009a. Introduction to the special issue on rotational seismology and engineering applications. *Bulletin of the Seismological Society of America*, **99**(2B), 945–957.

Lee, W. H. K., Huang, B. S., Langston, C. A., Lin, C. J., Liu, C. C., Shin, T. C., Teng, T. L., and Wu, C. F., 2009b. Review: progress in rotational ground-motion observations from explosions and local earthquakes in Taiwan. *Bulletin of the Seismological Society of America*, **99**(2B), 958–967.

Lin, C. J., Liu, C. C., and Lee, W. H. K., 2009. Recording rotational and translational ground motions of two TAIGER explosions in northeastern Taiwan on 4 March 2008. *Bulletin of the Seismological Society of America*, **99**(2B), 1237–1250.

Lin, C. J., Huang, H. P., Liu, C. C., and Chiu, H. C., 2010. Application of rotational sensors to correcting rotation-induced effects on accelerometers. *Bulletin of the Seismological Society of America*, **100**, 585–597.

Liu, C. C., Huang, B. S., Lee, W. H. K., and Lin, C. J., 2009. Observing rotational and translational ground motions at the HGSD station in Taiwan from 2004 to 2008. *Bulletin of the Seismological Society of America*, **99**(2B), 1228–1236.

Mallet, R., 1862. *Great Neapolitan Earthquake of 1857*. London: Chapman and Hall, Vol. I and II.

McLeod, D. P., Stedman, G. E., Webb, T. H., and Schreiber, U., 1998. Comparison of standard and ring laser rotational seismograms. *Bulletin of the Seismological Society of America*, **88**, 1495–1503.

Nigbor, R. L., 1994. Six-degree-of-freedom ground motion measurement. *Bulletin of the Seismological Society of America*, **84**, 1665–1669.

Nigbor, R. L., Evans, J. R., and Hutt, C. R., 2009. Laboratory and field testing of commercial rotational seismometers. *Bulletin of the Seismological Society of America*, **99**(2B), 1215–1227.

Oldham, R. D., 1899. Report on the Great Earthquake of 12th June 1897, *Memoir of the Geological Survey of India*, **29**, 379.

Oliveira, C. S., and Bolt, B. A., 1989. Rotational components of surface strong ground motion. *Earthquake Engineering and Structural Dynamics*, **18**, 517–526.

Pham, D. N., Igel, H., Wassermann, J., Cochard, A., and Schreiber, U., 2009. The effects of tilt on interferometric rotation sensors. *Bulletin of the Seismological Society of America*, **99**(2B), 1352–1365.

Pujol, J., 2009. Tutorial on rotations in the theories of finite deformation and micropolar (Cosserat) elasticity. *Bulletin of the Seismological Society of America*, **99**(2B), 1011–1027.

Reid, H. F., 1910. *The Mechanics of the Earthquake*. The California Earthquake of April 18, 1906, Report of the State Earthquake Investigation Commission, Vol. 2. Washington: Carnegie Institution of Washington, pp. 43–47.

Richter, C. F., 1958. *Elementary Seismology*. San Francisco: W. H. Freeman.

Schreiber, K. U., Igel, H., Cochard, A., Velikoseltsev, A., Flaws, A., Schuberth, B., Drewitz, W., and Müller, F., 2005. The GEOsensor project: rotations – a new observable for seismology. In Flury, J., Rummel, R., Reigber, C., and Rothacher, M. (eds.), *Observation of the Earth System from Space*. Heidelberg: Springer, pp. 1–19.

Schreiber, K. U., Hautmann, J. N., Velikoseltsev, A., Wassermann, J., Igel, H., Otero, J., Vernon, F., and Wells, J.-P. R., 2009a. Ring laser measurements of ground rotations for seismology. *Bulletin of the Seismological Society of America*, **99**(2B), 1190–1198.

Schreiber, K. U., Velikoseltsev, A., Carr, A. J., and Franco-Anaya, R., 2009b. The application of fiber optic gyroscopes for the measurement of rotations in structural engineering. *Bulletin of the Seismological Society of America*, **99**(2B), 1207–1214.

Spudich, P., and Fletcher, J. B., 2008. Observation and prediction of dynamic ground strains, tilts, and torsions caused by the M_w 6.0 2004 Parkfield, California, earthquake and aftershocks, derived from UPSAR Array observations. *Bulletin of the Seismological Society of America*, **98**, 1898–1914.

Spudich, P., and Fletcher, J. B., 2009. Software for inference of dynamic ground strains and rotations and their errors from short baseline array observations of ground motions. *Bulletin of the Seismological Society of America*, **99**(2B), 1480–1482.

Spudich, P., Steck, L. K., Hellweg, M., Fletcher, J. B., and Baker, L. M., 1995. Transient stresses at Parkfield, California, produced by the M 7.4 Landers earthquake of June 28, 1992: Observations from the UPSAR dense seismograph array. *Journal of Geophysical Research*, **100**(B1), 675–690.

Stedman, G. E., 1997. Ring laser tests of fundamental physics and geophysics. *Reports on Progress in Physics*, **60**, 615–688.

Stupazzini, M., De La Puente, J., Smerzini, C., Kaser, M., Igel, H., and Castellani, A., 2009. Study of rotational ground motion in the near field region. *Bulletin of the Seismological Society of America*, **99**(2B), 1271–1286.

Suryanto, W., Igel, H., Wassermann, J., Cochard, A., Schuberth, B., Vollmer, D., Scherbaum, F., Schreiber, U., and Velikoseltsev, A., 2006. First comparison of array-derived rotational ground motions with direct ring laser measurements. *Bulletin of the Seismological Society of America*, **96**, 2059–2071.

Takamori, A., Araya, A., Otake, Y., Ishidoshiro, K., and Ando, M., 2009. Research and development status of a new rotational seismometer based on the flux pinning effect of a superconductor. *Bulletin of the Seismological Society of America*, **99**(2B), 1174–1180.

Takeo, M., 1998. Ground rotational motions recorded in near-source region. *Geophysical Research Letters*, **25**(6), 789–792.

Takeo, M., 2009. Rotational motions observed during an earthquake swarm in April, 1998, at offshore Ito, Japan. *Bulletin of the Seismological Society of America*, **99**(2B), 1457–1467.
Takeo, M., and Ito, H. M., 1997. What can be learned from rotational motions excited by earthquakes? *Geophysical Journal International*, **129**, 319–329.
Teisseyre, R., Takeo, M., and Majewski, E. (eds.), 2006. *Earthquake Source Asymmetry, Structural Media and Rotation Effects*. Berlin/Heidelberg: Springer.
Teisseyre, R., Nagahama, H., and Majewski, E. (eds.), 2008. *Physics of Asymmetric Continua: Extreme and Fracture Processes: Earthquake Rotation and Soliton Waves*. Berlin/Heidelberg: Springer.
Todorovska, M. I., and Trifunac, M. D., 1990. Note on excitation of long structures by ground waves. *Journal of Engineering Mechanics-ASCE*, **116**(4), 952–964.
Todorovska, M. I., Igel, H., Trifunac, M. D., and Lee, W. H. K., 2008. Rotational earthquake motions – international working group and its activities. In *Proceedings of the 14th World Conference on Earthquake Engineering*, Beijing, China, October 12–17, Paper ID: S03-02-0031.
Trifunac, M. D., 1979. A note on surface strains associated with incident body waves. *Bulletin of the European Association of Earthquake Engineering*, **5**, 85–95.
Trifunac, M. D., 1982. A note on rotational components of earthquake motions on ground surface for incident body waves. *Soil Dynamics and Earthquake Engineering*, **1**, 11–19.
Trifunac, M. D., 2006. Effects of torsional and rocking excitations on the response of structures. In Teisseyre, R., Takeo, M., and Majewski, E. (eds.), *Earthquake Source Asymmetry, Structural Media and Rotation Effects*. Heidelberg: Springer, pp. 569–582.
Trifunac, M. D., 2009. Earthquake engineering, nonlinear problems in. In Meyers, R. A. (ed.), *Encyclopedia of Complexity and Systems Science*. New York: Springer, pp. 2421–2437.
Twiss, R. J., 2009. An asymmetric micropolar moment tensor derived from a discrete-block model for a rotating granular substructure. *Bulletin of the Seismological Society of America*, **99** (2B), 1103–1131.
Twiss, R., Souter, B., and Unruh, J., 1993. The effect of block rotations on the global seismic moment tensor and patterns of seismic P and T axes. *Journal of Geophysical Research*, **98**, 645–674.
Wang, H., Igel, H., Gallovic, F., and Cochard, A., 2009. Source and basin effects of rotations: comparison with translations. *Bulletin of the Seismological Society of America*, **99**(2B), 1162–1173.
Wassermann, J., Lehndorfer, S., Igel, H., and Schreiber, U., 2009. Performance test of a commercial rotational motions sensor. *Bulletin of the Seismological Society of America*, **99**(2B), 1449–1456.
Wu, C. F., Lee, W. H. K., and Huang, H. C., 2009. Array deployment to observe rotational and translational ground motions along the Meishan fault, Taiwan: a progress report. *Bulletin of the Seismological Society of America*, **99**(2B), 1468–1474.

Cross-references

Body Waves
Earthquake Rupture: Inverse Problem
Earthquake, Magnitude
Earthquakes, Strong-Ground Motion
Great Earthquakes
Propagation of Elastic Waves: Fundamentals
Seismic Hazard
Seismic Instrumentation
Seismological Networks
Surface Waves

SHEAR-WAVE SPLITTING: NEW GEOPHYSICS AND EARTHQUAKE STRESS-FORECASTING

Stuart Crampin
British Geological Survey, Murchison House, Edinburgh, Scotland, UK

Synonyms

Birefringence; Double refraction

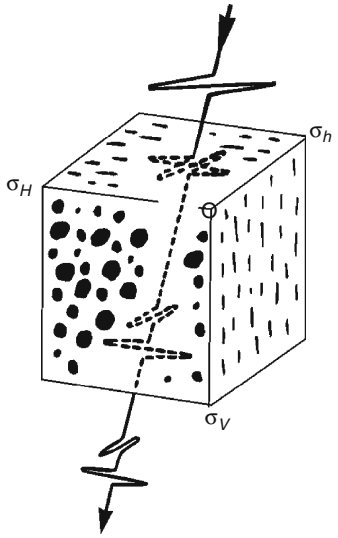
Definition

Observations of shear-wave splitting in the Earth lead to a new understanding of fluid-rock deformation: a New Geophysics, where earthquakes can be stress-forecast.
Shear-wave splitting: Seismic shear-waves propagating through effectively anisotropic solids, such as aligned crystals or aligned microcracks, split into two phases with different velocities and approximately orthogonal polarizations.
New Geophysics: A new understanding of fluid-rock deformation in distributions of stress-aligned fluid-saturated microcracks where microcracks in the crust are so closely spaced that they verge on fracturing and hence are critical-systems (also known as complex-systems). Critical-systems impose fundamentally new properties on conventional subcritical geophysics and hence are a New Geophysics.
Earthquake stress-forecasting: Shear-wave splitting above swarms of small earthquakes can be used as stress-measuring stations to monitor stress-accumulation and stress-relaxation (crack-coalescence) before earthquakes, and hence *stress-forecast* the time, magnitude, and, in some cases, fault-break of impending large earthquakes.

Introduction

Transversely polarized seismic shear-waves propagating through in situ rocks with some form of elastic anisotropy, such as aligned microcracks, split into two orthogonal polarizations which propagate with different velocities, and hence lead to shear-wave splitting. The polarizations are strictly orthogonal for phase-velocity propagation, and approximately orthogonal for group-velocity propagation. Figure 1 is a schematic illustration of shear-wave splitting in the parallel vertical stress-aligned microcracks characteristic of most rocks in the Earth's crust, once below near-surface weathering and stress-release anomalies. Originally referred to as birefringence or double refraction, shear-wave splitting was known to exist from the properties of anisotropic elastic solids (Love, 1927; Nye, 1957).

Double refraction has been observed experimentally as two orthogonally polarized shear-waves with different velocities in a stressed rock sample (Nur and Simmons, 1969). The phenomenon of shear-wave splitting was first



Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting, Figure 1 Schematic illustration of stress-aligned shear-wave splitting through parallel vertical microcracks (after Crampin,1994). σ_V, σ_H, and σ_h are principal axes of vertical stress, and maximum and minimum horizontal stress, respectively.

recognized (and named) in synthetic seismograms in an anisotropic crystalline Earth model (Keith and Crampin, 1977) and in distributions of parallel vertical microcracks (Crampin, 1978). Shear-wave splitting was first positively identified in the field (above swarms of small earthquakes) by Crampin et al. (1980) and in seismic reflection surveys in hydrocarbon reservoirs by Alford (1986). An earlier attempt by Gupta (1973) to demonstrate double refraction in shear-waves some 100 km from small earthquakes failed because of misunderstanding the behavior of shear-wave splitting and the shear-wave window (Crampin et al., 1981).

Although variation of velocity with direction is certainly characteristic of seismic propagation in anisotropic media, shear-wave splitting (or the generalized three-dimensional coupling of surface-wave motion), easily identifiable in particle-motion diagrams (hodograms), is highly diagnostic of some form of effective anisotropy. One of the major difficulties is that observations of shear-wave splitting are path-integration phenomena that give no indication of where the anisotropy is located along the ray path, nor its extent or strength. The only exception is in sedimentary basins where the processing seismic surveys in exploration seismology may locate the anisotropy more accurately (for example, in Angerer et al., 2002).

Note that shear-wave splitting, as in Figure 1, seems quite simple in concept which may tempt authors, such as Gupta (1973), to make unfounded assumptions which lead to incorrect and misleading conclusions. Shear-wave splitting *is* comparatively straightforward, but requires a jump in comprehension from previous isotropic experience. Two early comprehensive reviews (Crampin, 1977, 1981) outlined seismic-wave propagation in anisotropic and cracked elastic media. Crampin and Peacock (2008) recently reviewed the current understanding of shear-wave splitting and identified 17 commonly assumed fallacies in understanding shear-wave splitting that can lead to unjustified conclusions.

It is suggested that the behaviour of shear-wave splitting is now comparatively well-understood. The major new advance is that observations of shear-wave splitting indicate that the stress-aligned fluid-saturated microcracks in almost all rocks are so closely spaced that they verge on failure by fracturing (and earthquakes). Such verging on failure indicates that the distributions of microcracks are critical-systems. Critical systems are a *New Physics* (Davies, 1989), hence a *New Geophysics* that imposes a range of fundamentally new properties on the previous subcritical geophysics. This is a major advance in understanding fluid-rock deformation that has applications to many branches of solid-earth geophysics (Crampin and Peacock, 2005, 2008; Crampin, 2006).

This entry will not discuss the theory of seismic anisotropy and shear-wave splitting which is adequately outlined in Crampin (1977, 1981), Helbig (1994), Maupin and Park (2007), and elsewhere. Here, we merely outline the behaviour of shear-wave splitting in the crust that leads to the New Geophysics and stress-forecasting earthquakes.

Terminology: *references to terminology in the text will be indicated by* (qv).

APE: anisotropic poro-elasticity is a model for deformation/evolution of critical-systems of distributions of stress-aligned fluid-saturated microcracks under changing conditions (Zatsepin and Crampin, 1997; Crampin and Zatsepin, 1997).

Aspect-ratio: Crack thickness over crack diameter.

Band-1 directions: Band-1 ray-path directions are the solid angle between 15° and 45° either side of the crack planes in a distribution of vertical parallel microcracks (Crampin, 1999). APE shows that the effect of increasing (or decreasing) stress is to increase (or decrease) the average time-delays in Band-1 directions.

Band-2 directions: Band-2 ray-path directions are in the solid angle ±15° to the crack planes (Crampin, 1999). Band-2 directions are sensitive to crack density, but APE shows that crack density does not vary consistently for small changes of stress.

Complex-systems: see critical-systems.

Crack density: Dimensionless crack density, *CD,* is specified by $CD = N a^3$ where N is number of cracks of radius a per unit cube.

Critical-systems: Critical-systems (also known as complex-systems, Davies, 1989) are complex heterogeneous interactive systems which verge on fracture-criticality and failure. Extremely common in a huge range of different phenomena, critical-systems impose fundamentally new properties on the previously

subcritical physics and subcritical solid-earth geophysics, and hence are a New Geophysics (Crampin, 2006).

Earthquake prediction: The prediction of time, magnitude, and location of earthquakes by conventional subcritical techniques, but not using shear-wave splitting which implies critical techniques.

Earthquake stress-forecasting: Using shear-wave splitting to monitor stress-accumulation and stress-relaxation in the surrounding rock mass that allows the time and magnitude of earthquakes to be stress-forecast, where other precursory phenomena may indicate the fault-break.

EDA: Extensive-Dilatancy Anisotropy is the name given to the distributions of fluid-saturated microcracks pervading most rocks in the crust (Crampin et al., 1984). EDA-cracks, in ostensibly unfractured rocks, are observed to have crack densities between ~0.015 and ~0.045: that is shear-wave velocity anisotropy of ~1.5 – ~4.5% (Crampin, 1994). This comparatively small percentage of velocity anisotropy has massive implications for a huge range of phenomena and leads to the New Geophysics.

Fracture-criticality: The value of crack density, ~0.055, at which stress-aligned microcracks are so closely spaced they verge on fracturing, and hence are critical-systems (qv) (Crampin, 1994, 1999; Crampin and Zatsepin, 1997).

New Geophysics: Shear-wave splitting indicates that microcracks in the crust are so closely spaced they verge on fracture-criticality and fracturing, and hence are critical-systems. Critical-systems are a *New Physics* (Davies, 1989), hence a *New Geophysics*, that imposes a range of fundamentally new properties on conventional subcritical geophysics (Crampin, 2006). Most of these properties have been observed in the field (Crampin and Peacock, 2005, 2008).

PTL-anisotropy: Distributions of horizontal Periodic Thin Layers lead to a form of anisotropy (hexagonal symmetry, commonly known as transverse isotropy, with a vertical axis of symmetry) that is common in sedimentary basins, including many hydrocarbon reservoirs (Wild and Crampin, 1991). Shear-wave velocity anisotropy in PTL-anisotropy may exceed 30% and would be likely to generate cusps in shear-wave velocity sheets. Such cusps have been observed (and modeled) in hydrocarbon reservoirs in the field (Slater et al., 1993).

Seismic anisotropy: Seismic anisotropy describes an elastic solid; these properties vary with direction as in aligned crystals or solids pervaded by aligned microcracks.

Seismic isotropy: Isotropy describes an elastic solid which has the same elastic properties in all directions.

Stress-forecasting earthquakes: see Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting.

TIH- (or HTI-) anisotropy: Transverse isotropy with a horizontal axis of cylindrical symmetry: the typical symmetry of EDA-cracks. Note that although transverse isotropy is specified by five elastic constants, distributions of vertical parallel EDA-cracks are specified by only three: crack density, crack aspect-ratio, and crack strike.

TIV- (or VTI-) anisotropy: Transverse isotropy with a vertical axis of cylindrical symmetry: the typical symmetry of PTL-anisotropy. In TIV-anisotropy, propagation in the horizontal plane is isotropic with no variation with azimuth.

Time-delay: The time-delay between the arrivals of the two split shear-wave phases in seismic anisotropy. Time-delays above earthquakes are frequently normalized by path length to ms/km.

Transverse isotropy: Transverse isotropy (strictly hexagonal symmetry) is the anisotropic symmetry system with five elastic constants, and is isotropic in directions perpendicular to an axis of cylindrical rotational symmetry.

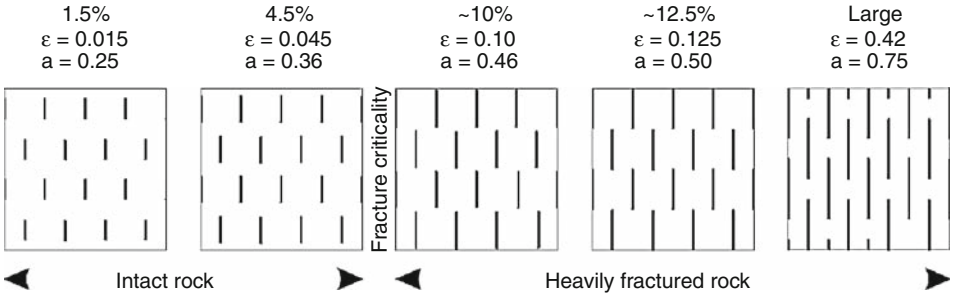
90°-flips in shear-wave polarisations: Abrupt 90° changes in shear-wave polarisations (90°-flips) have been observed and modeled along shear-wave ray paths passing near to point-singularities (see below) (Bush and Crampin, 1991; Crampin, 1991). 90°-flips can also be caused by critically high pore-fluid pressures locally re-orienting stress-aligned fluid-saturated EDA-cracks, both in fluid-injection in critically pressurized hydrocarbon reservoirs (Angerer et al., 2002) and adjacent to seismically-active (critically-pressurized) faults (Crampin et al., 2002).

Fundamental features of shear-wave splitting in the crust

Observations of shear-wave splitting at a horizontal free surface typically display stress-aligned parallel polarisations. Since only transverse isotropy with a horizontal axis of symmetry (TIH-anisotropy) has such parallelism, and only parallel vertical cracks have such symmetry, the splitting is necessarily caused by stress-aligned EDA-microcracks. Microcracks is conformed as temporal changes in time-delays are observed, in fluid-injections (Angerer et al., 2002) and before earthquakes (Crampin and Peacock, 2008). Only microcracks have sufficient compliance to allow such changes.

Observations of azimuthally varying stress-aligned shear-wave splitting, in almost all igneous, metamorphic, and sedimentary rocks in the crust, indicate a minimum of ~1.5% shear-wave velocity anisotropy and a maximum in ostensibly unfractured rock of ~4.5% (Crampin, 1994, 1999; Crampin and Peacock, 2008). Since crack density can be written as a dimensionless quantity $CD = N\,a^3$, where N is the number of cracks of radius a per unit cube and if Poisson's ratio is 0.25, $CD \approx 1/100$ of the percentage of shear-wave velocity anisotropy, observed shear-wave splitting can be imaged as uniform distributions of parallel vertical dimensionless cracks in Figure 2 (Crampin, 1994; Crampin and Peacock, 2008).

The evolution of an initially random distribution of stress-aligned fluid-saturated microcracks under changing



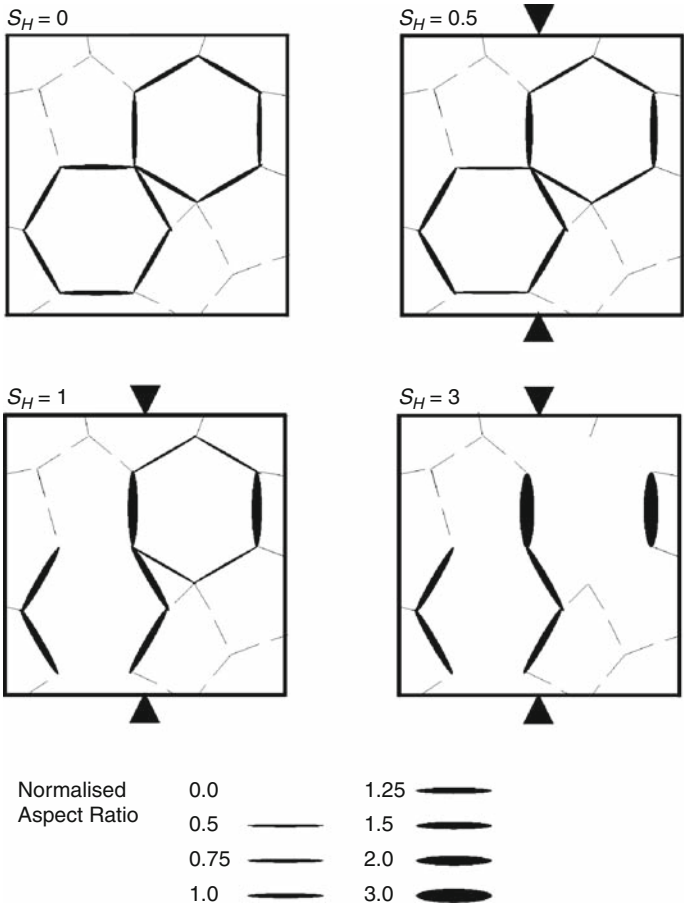
Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting, Figure 2 Schematic illustration of dimensionless three-dimensional crack distributions in in situ rocks for a range of percentages shear-wave velocity anisotropy, where ε is crack density and *a* is normalized crack radius (after Crampin, 1994). Fracture-criticality is at ~5.5% shear-wave velocity anisotropy (Crampin and Zatsepin, 1997).

conditions can be modeled by the equations of anisotropic poro-elasticity, APE, (Crampin and Zatsepin, 1997). Figure 3 gives a schematic illustration of APE-deformation. The effect of increasing (or decreasing) stress is to increase (or decrease) the average aspect-ratio for cracks parallel to the increasing stress. APE-deformation is almost without parameters and yet matches a huge range of phenomena in exploration and earthquake seismology (Crampin and Peacock, 2008), where the general applicability is because microcracks are so closely spaced that they are critical-systems and lead to the New Geophysics. Note that APE-modeling, because of the universality of critical-systems, is almost without parameters, and yet is widely applicable.

Classes of anisotropic symmetry and shear-wave splitting

Shear-wave splitting has characteristic patterns of polarizations and velocities in three-dimensional propagation depending on the class of anisotropic symmetry (Nye, 1957; Crampin, 1977; Crampin and Kirkwood, 1981). Elastic constants allow eight classes of anisotropic symmetry (including isotropic symmetry which describes a material with two elastic constants, usually written λ and μ, and has identical elastic properties in all directions of propagation). It is the characteristic 3D patterns of planes of mirror symmetry which define the classes of anisotropic symmetry. Four types of anisotropic symmetry are commonly found in the Earth (where the horizontal plane is typically a plane of mirror symmetry):

1. *Orthorhombic symmetry* with nine independent elastic constants has three mutually-orthogonal planes of mirror symmetry. Crystalline olivine, a possible upper mantle constituent has orthorhombic symmetry, where shear-waves split in 3D patterns of polarizations. In addition, simple combinations of EDA- and PTL-anisotropy also have orthorhombic symmetry (Wild and Crampin, 1991). The effects of combinations of EDA- and PTL-anisotropies have been observed in hydrocarbon reservoirs (Bush and Crampin, 1991).
2. *Transverse isotropy with a vertical axis of cylindrical symmetry* commonly referred to as TIV-anisotropy. In TIV, shear-waves split strictly into phases oscillating in the vertical plane (*SV* waves) and phases oscillating horizontally (*SH* waves). There is no azimuthal variation. TIV is the effective anisotropic symmetry of finely divided horizontal layers found in many hydrocarbon reservoirs. It is also the symmetry of small horizontal platelets in shales. Thomsen (1986) derived three highly important approximations (the Thomsen parameters) for the behaviour of seismic waves in TIV media which are directly analogous to parameters measured in reflection surveys. Thomsen parameters are widely used and highly effective in interpreting exploration surveys. Thomsen parameters cannot be used to describe *azimuthally varying* anisotropy, except in isolated directions of sagittal symmetry.
3. *Transverse isotropy with a horizontal axis of cylindrical symmetry* commonly referred to as TIH-anisotropy. TIH is commonly observed above small earthquakes where the faster split shear-wave is polarized parallel to the direction of maximum horizontal stress. Only TIH-anisotropy can produce such parallelism. Only fluid-saturated microcracks are common to the huge variety of geological materials where TIH is observed, and since temporal changes before earthquakes are observed in the field, only fluid-saturated microcracks have sufficient compliance to display temporal changes for small changes in parameters. Consequently, the observed parallelism of fast polarizations directly confirms stress-aligned microcracks throughout almost all rocks in the crust (Crampin and Peacock, 2008).
4. *Monoclinic symmetry* has 13 elastic constants and is the symmetry of two (or more) sets of intersecting non-orthogonal vertical parallel EDA-cracks. Shear-wave splitting in such monoclinic symmetry has been observed and modeled in the field (Liu et al., 1993a). If two intersecting sets of vertical EDA-cracks are orthogonal, the symmetry is orthorhombic. Liu et al. (1993b) display the theoretical effects of intersecting sets of parallel vertical EDA-cracks for a range of parameters and a range of angles of intersection.



Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting, Figure 3 Schematic illustration of anisotropic poro-elastic (APE) modelling of the effects of increasing stress on aspect-ratios of (6% porosity) microcracks, where aspect-ratios are given below (Crampin and Zatsepin, 1995; Crampin, 2006).

Cusps

In strong shear-wave velocity anisotropy, the direction of group-velocity (energy) propagation of the faster of the two split shear-wave may deviate so strongly that they overlap themselves and cause large amplitude cusps. Cusps have been observed and modeled in hydrocarbon reservoirs (Slater et al., 1993). However, the commonest form of cusps that occur in all anisotropic symmetry systems are the cuspoidal lids, fins, and ridges associated with point- and line-singularities in group-velocity shear-wave surfaces (Crampin, 1991; Wild and Crampin, 1991).

Shear-wave window

Shear-waves with *SV* orientations incident on the free surface of an isotropic medium are identical to the incoming shear-wave (but double the amplitude) *only within* the shear-wave window. Outside the window, *S*-wave energy is lost to *SV*-to-*P* conversions so that *SV* waves are heavily distorted and even the arrival of the direct *S* wave may be unreadable. The edge of the shear-wave window is outlined by the critical angle of incidence, $i_c = \sin^{-1}$ (Vs/Vp), where Vp and Vs are the isotropic *P*-wave and *S*-wave velocities, and i_c is the critical angle for *S*-to-*P* conversions (Evans, 1984; Booth and Crampin, 1985). The conical solid angle of the shear-wave window may be considered as *above* small earthquakes or *below* surface recorders.

Other features of the shear-wave window:

1. The critical angle of incidence i_c is $\sim 35^\circ$ for a uniform isotropic medium with a Poisson's ratio of 0.25, where Lamé parameters $\lambda = \mu$.
2. Because of near-surface low-velocity layers, ray paths tend to curve upwards as they approach the free surface, so that the effective shear-wave window is often out to straight-line source-to-receiver incidence of 45° or greater.
3. Topographic irregularities may seriously distort the shear-wave window and may cause *S*-to-*P* conversions (the Local *SP*-waves) both inside and outside the theoretical shear-wave window. Local *SP*-waves are frequently observed in the field as large single-sided pulses on the *SV*-component, and may be mistakenly interpreted as anomalous shear-wave splitting (Crampin, 1990).

4. The shear-wave window only refers to *SV*-propagation. No energy is lost by *SH* waves at any angle of incidence on a plane horizontal free surface. Consequently, only the *SV*-component of any incident-split shear-wave is affected by the shear-wave window.
5. In anisotropic propagation, the shear-wave window may severely disturb shear-wave splitting.

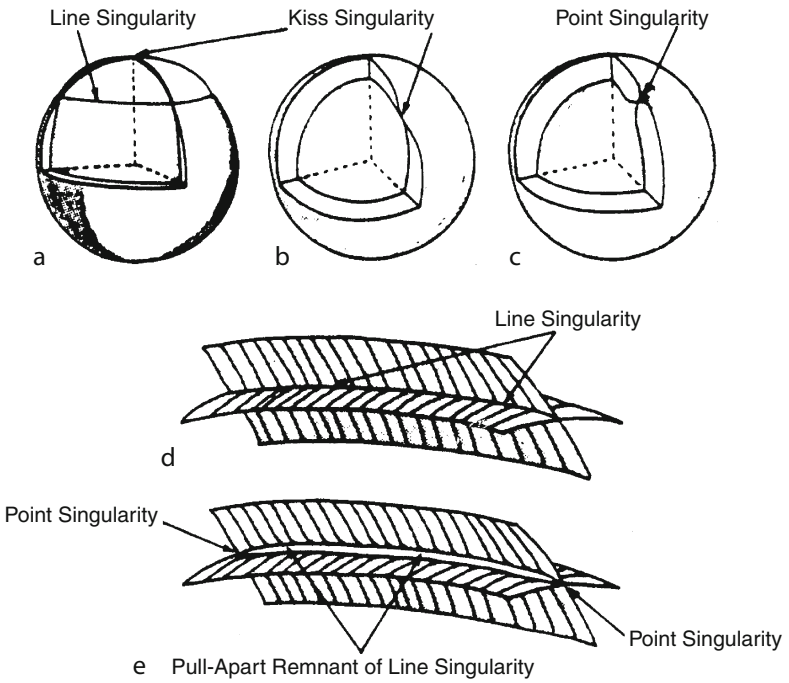
These various anomalies mean that shear-wave splitting recorded at a free surface needs to be interpreted with caution (Crampin and Peacock, 2008).

Shear-wave singularities

The three body waves in anisotropic solids trace out three velocity surfaces for propagation in three-dimensional propagation: a *P* wave; and two orthogonally-polarized shear-wave surfaces. The phase-velocity shear-wave surfaces from a point of propagation necessarily touch and are continuous in a variety of singularities in three-dimensional patterns of direction characteristic of the anisotropic symmetry class (Crampin and Yedlin, 1981; Crampin and Kirkwood, 1981; Wild and Crampin, 1991). The only way for energy of a split shear-wave arrival to pass from one shear-wave surface to the other is along directions passing directly through a singularity.

Figure 4 shows the three types of singularity in *phase-velocity surfaces*:

1. *Kiss-singularities* where two shear-wave surfaces touch tangentially (Figure 4a, 4b). Kiss-singularities occur in directions of principal crystallographic axes and are restricted to cubic, tetragonal anisotropic symmetry, and along the cylindrical symmetry direction of transversely isotropic symmetry. Kiss-singularities can also occur in orthorhombic and monoclinic anisotropic symmetry when two shear-wave surfaces happen to touch serendipitously.
2. *Point-singularities* (sometimes called conical points) where the two shear-wave surfaces touch at the vertices of convex and concave cones (Figure 4c). Shear-waves in all classes of anisotropic symmetry possess point-singularities except transverse isotropy (hexagonal symmetry), which only has kiss- and line-singularities (Figure 4a). Point-singularities are particularly numerous in combinations of PTL- and EDA-anisotropy, where thin layers are pervaded by vertical microcracks (Wild and Crampin, 1991).
3. *Line- or intersection-singularities* occur only in transverse isotropy where the two shear-wave phase-velocity surfaces intersect each other in circular line-(intersection-) singularities about the symmetry axis (Figure 4a, 4d). In even marginally disturbed transverse isotropy, there are dramatic changes. Since both PTL- and EDA-anisotropy have transversely isotropy symmetry with line-singularities, Wild and Crampin



Shear-Wave Splitting: New Geophysics and Earthquake Stress-Forecasting, Figure 4 Schematic illustration of the intersection of two shear-wave phase-velocity surfaces showing topography near: (**a**) line-singularity or intersection-singularity (with kiss-singularity); (**b**) kiss-singularity; and (**c**) point-singularity (after Crampin and Yedlin, 1981). (**d**) A segment of a representation of a line-singularity, and (**e**) a *pull-apart remnant of a line-singularity* when the transverse isotropy symmetry is disturbed (after Wild and Crampin, 1991).

(1991) show that the phase-velocity representations of combinations of PTL and EDA have *pull-apart remnants of line-singularities*. The intersections of the inner and outer shear-wave surfaces separate, and 'pinch' together, in varying degrees of tightness, between a string of point-singularities (Figure 4e).

Singularities in *group-velocity surfaces* have a variety of behaviour:

1. *Kiss-singularities* have simple group-velocity representation, the shear-wave surfaces again touch tangentially.
2. *Point-singularities* in group-velocity surfaces: the faster shear-wave surfaces transform into open almost circular "holes", whereas the slower surfaces transform into flat cuspoidal "lids" which exactly fit into the open holes. The effects of such point-singularities have been observed and modeled in a hydrocarbon reservoir the field (Bush and Crampin, 1991). Crampin (1991) modeled synthetic seismograms for ray path directions passing close to point-singularities. The effects can be dramatic with possibly great amplification or attenuation of amplitudes and abrupt 90° changes in shear-wave polarisations (90°-flips) either side of point-singularities.
3. *Line- or intersection-singularities*. In undisturbed transverse isotropy, the group-velocity representations of line-singularities merely intersect with each sheet showing no evidence of intersection, as in the phase-velocity representation in Figure 4d. However, group-velocity representations of pull-apart remnants of line-singularities in disturbed transverse isotropy typically display thin cuspoidal "fins", "ridges", and "lids" of extraordinary complexity (and beauty) (Wild and Crampin, 1991).

There are two conclusions. The effects of pull-apart remnants of line-singularities are sensitive to very small differences in direction of propagation or microcrack geometry, say. The subtleties of shear-wave propagation near the extensive pull-apart remnants of line-singularities are likely to be the reason for many of the complications frequently observed in field records of shear-waves in sedimentary basins. If correctly interpreted, such features can provide crucial information particularly about the relative proportions of PTL- and EDA-anisotropy (Bush and Crampin, 1991).

Sources of shear-wave splitting

There are three common areas of investigation into shear-wave splitting (and seismic anisotropy).

1. *Shear-wave splitting in the upper mantle*, which is typically assumed to be caused by strain-aligned crystals such as olivine, and is used to demonstrate directions of stress and flow in plate tectonics (reviewed by Savage, 1999).
2. *Shear-wave splitting in hydrocarbon reservoirs* in seismic exploration in industrial seismology, where initially "anisotropy is a nuisance" (Helbig, 1994) because many well-established processing procedures assume isotropic propagation and are no longer valid if the rocks are anisotropic. More recently, however, azimuthally varying shear-wave splitting in TIH-anisotropy is being actively used to determine fracture orientations and improve fluid-flow for hydrocarbon recovery (reviewed by Helbig and Thomsen, 2005).
3. *Shear-wave splitting above small earthquakes* initially stimulated development of the theory (reviewed by Crampin, 1981) and observation (Crampin et al., 1980; Alford, 1986), and is now demonstrating the *New Geophysics* of a crack-critical crust (Crampin, 2006), where time-delays directly monitor low-level fluid-rock deformation at levels of deformation well-below those at which fracturing takes place (Crampin and Zatsepin, 1997).

Remarkably, there has been very little interaction between these three areas of investigation. Each has developed largely independently where papers in one area seldom cite other areas of research. Shear-wave splitting in these three areas of investigation will be discussed separately.

Shear-wave splitting in the upper mantle

Seismic anisotropy in the mantle was first recognized from the incompatibility of Rayleigh and Love surface-wave inversions (Anderson, 1961) and from the velocity anisotropy of horizontally propagating *Pn*-waves in oceanic basins (Hess, 1964; Raitt et al., 1969). Another anisotropic surface-wave phenomenon directly analogous to shear-wave splitting of body waves is the coupling of surface-wave motion of Raleigh and Love modes into Generalized modes varying in three-dimensions (Crampin and King, 1977). Just as shear-wave splitting is highly diagnostic of some form of seismic anisotropy, coupled surface-wave motion is highly diagnostic of anisotropy in the mantle. The pronounced coupling between Second Mode Rayleigh and Second Mode Love waves across Eurasia could be caused by as little as 4 km of 4% shear-wave velocity anisotropy at the top of the upper mantle (Crampin and King, 1977).

Shear-wave splitting of body waves in the mantle only became observable with advances in digital instrumentation. The first reports of such shear-wave splitting were by Ando et al. (1980) who reported time-delays of ~1s on 2s-period shear-waves propagating vertically. This was attributed to 2–4% shear-wave velocity anisotropy in propagation through a magma reservoir, 100–200 km in depth beneath a volcanic area in Japan above the subducting Pacific Plate.

Since that time, shear-wave splitting has been observed extensively in the uppermost 200 km of the mantle and in the D region (Silver and Chan, 1988, 1991), reviewed by Savage (1999), and Long and Silver (2009) and others. These observations are principally of shear-wave splitting in *SKS*-phases, which are frequently large amplitude,

where the *P*-S conversion on exiting from the core provides a known *SV* orientation at the start the final shear-wave path to the surface. Frequencies are typically 0.5–0.25 Hz leading to 2–4% shear-wave velocity anisotropy, similar to the crack-induced shear-wave velocity anisotropy in the crust.

The cause of the shear-wave splitting in the mantle is not wholly resolved. The most frequently cited cause is referred as lattice-preferred orientation (LPO) of the crystallographic axes of elastically anisotropic crystals, such as olivine (Karato and Wu, 1993) or orthopyroxene (Ben-Ismail et al., 2001). These crystals are assumed to be aligned by convective flow where the presence of water, induced by bending of subducting slabs (Faccenda et al., 2009), say, could have major effects on crystal orientations and shear-wave splitting (Karato and Jung, 1998; Katayama et al., 2005).

However, the similarities in the degree of shear-wave splitting and the polarization to splitting in the crust which is certainly caused by EDA-cracks suggests the alternative source of anisotropy of fluid-saturated stress-aligned cracks. There is water in much of the upper mantle. Water lowers the melting point of rock, and when crystalline rock first begins to melt, it melts along grain boundaries. Thus the anisotropy in the mantle could be caused by aligned cracks of films of liquid melt (Crampin, 2003).

Shear-wave splitting in hydrocarbon reservoirs

The effective TIV-anisotropy of finely stratified rocks was demonstrated by Postma (1955) and others, where double refraction (shear-wave splitting) only has *SV*- and *SH*-polarisations. Shear-wave splitting of TIV-anisotropy was not much investigated at that time since the predominant single-component vertical seismometers only recorded *SV* waves.

Azimuthally varying shear-wave splitting in hydrocarbon reservoirs burst onto the seismic exploration industry in a series of some dozen papers in two sessions at the 1986 SEG symposium in Houston. Alford (1986) rotated ("Alford rotations") the polarisations of reflection surveys recorded on two-component horizontal geophones from two-component horizontal vibrators, and showed that seismic shear-wave reflection sections displayed coherent behavior only when both source and receivers were rotated into the preferred (anisotropic) polarizations. These papers stimulated great activity in exploring seismics in order to obtain preferred directions of fluid-flow by locating fracture orientations (reviewed by Helbig and Thomsen, 2005). This interest in anisotropy still continues today but is frequently investigated by wide-angle and other *P*-wave surveys because of the expense of three-component shear-wave surveys including instrumentation as well as the tripling of the quantity of data to be recorded and processed.

Note that azimuthally varying shear-wave splitting in hydrocarbon reservoirs has been typically interpreted by the hydrocarbon industry as being caused by *aligned fractures*, where dimensions of the fractures are unspecified but are expected to be large enough to lead to improved hydrocarbon recovery. In contrast, Crampin and Peacock (2008) suggest that the remarkable compliance of shear-wave splitting to low-level deformation necessarily means that the splitting is caused by stress-aligned fluid-saturated *microcracks*. Since in many circumstances microcracks are parallel to macro-fractures, the difference in interpretation is often immaterial. However, the recent demonstration, both in theory (Chapman, 2003) and observations (Maultzsch et al., 2003), that anisotropy is frequency-dependent and varies with the dimensions of microcracks may change this interpretation.

Shear-wave splitting above small earthquakes

Shear-wave splitting was first positively identified in the crust above a swarm of small earthquakes near the North Anatolian Fault in Turkey in the Turkish Dilatancy Project (TDP1) experiment, (Crampin et al., 1980). This experiment was designed to search for the microcrack-induced shear-wave splitting suggested by Crampin (1978). Similar observations have now been observed above earthquakes worldwide (reviewed by Crampin and Peacock, 2008). However observations above earthquakes need to be recorded by three-component seismic networks within the shear-wave window of the earthquakes. This is a severe restriction as swarms of small earthquakes are scarce and intermittent, and seldom observed by seismic networks. Nevertheless, such swarms of small earthquakes have been used as 'stress-measuring stations' to monitor the effects of stress changes before impending earthquakes (Crampin and Peacock, 2008).

Iceland is an optimum location for studying shear-wave splitting. It is an offset of the Mid-Atlantic Ridge where, uniquely, (two) transform faults run onshore. These transform faults have persistent swarm activity monitored by a state-of-the-art seismic network (Stefánsson et al., 1993) available over the Internet (Volti and Crampin, 2003).

The APE model of microcrack deformation (Crampin and Zatsepin, 1997) shows that increases of stress increase crack aspect-ratios (crack thicknesses), and increases of aspect-ratio can be recognized by increases of average time-delays in Band-1 directions of the shear-wave window (Crampin, 1999), and have been used to stress-forecast earthquakes (Crampin et al., 1999, 2008).

The new geophysics

The shear-wave velocity anisotropy observed in the crust of the Earth indicates that the distributions of stress-aligned fluid-saturated microcracks (EDA-cracks) are so closely spaced they verge on fracturing and hence are critical-systems (Figure 2). Such critical microcracks are the most compliant elements of in situ rock, and indicate profound differences from conventional subcritical solid-earth geophysics. Critical systems are a *New Physics* (Davies, 1989), hence a *New Geophysics*, which imposes

a range of fundamentally new properties on conventional subcritical physics (Crampin and Peacock, 2005; Crampin, 2006). All complex heterogeneous interactive systems are critical-systems, and they are extremely common: the weather, clustering of traffic on roads, the life-cycle of fruit flies, and a huge range of physical phenomena from stellar radiation to quantum mechanics. Since the Earth is an archetypal complex heterogeneous interactive system, it is necessarily expected to be a critical system.

The new geophysical properties include:

1. *Monitorability* – effects of criticality (on the geometry of EDA-cracks) can be monitored within the interior of the crust by shear-wave splitting (Crampin, 1994, 2006).
2. *Calculability* – the evolution of microcrack geometry can be calculated by APE (Crampin and Zatsepin, 1997; Crampin, 1999; Angerer et al., 2002).
3. *Predictability* – if changing conditions are known, the effects on crack geometry can be predicted (as in Angerer et al., 2002).
4. *Controllability* – if the intended effects of some operation (fluid-injection, say) are desired (the opening of specific fractures for fluid-flow, say), the effects can be calculated by APE, and the effects can be controlled by feedback by analyzing shear-wave splitting. Crack aspect-ratios, as in Angerer et al. (2002) where the effects of fluid-injections were calculated, could be controlled by feedback.
5. *Universality* – the changes in critical-systems are pervasive over all available space where appropriate conditions are maintained. The coincidence of observations of 1.5–4.5% shear-wave velocity anisotropy with similar stress-oriented polarizations in all types of rock regardless of porosity, rock types, or geology (Crampin, 1994; Crampin and Peacock, 2005, 2008) is an example of universality in the crack-critical Earth, that is difficult to explain in a conventional subcritical Earth.
6. *Sensitivity* – extreme (butterfly wings) sensitivity to initial conditions. A prime example of such sensitivity is the SMSITES experiment in Iceland where distinctive variations in *P*- and shear-wave travel-time variations were observed at 70 km from swarm activity with an energy equivalent to a $M = 3.5$ earthquake at 70 km distance. A $M = 3.5$ earthquake is small, so these effects are observed at hundreds of times the likely source diameter in a conventional subcritical earth and are a clear demonstration of critical-system sensitivity (Crampin et al., 2003).

Stress-forecasting earthquakes

Shear-wave splitting and the New Geophysics lead to a new understanding of tectonic deformation before earthquakes. Initially, tectonic stress accumulates by interactions at plate boundaries leading to deformation of EDA-cracks which can be monitored by shear-wave splitting. Such stress-accumulations are not initially associated with any particular source-zone and may be observable at great distances from any potential source. As the stress-field approaches, fracture-criticality weaknesses are identified and microcracks begin to coalesce onto the potential slip-plane. There is stress-relaxation, and eventually the impending earthquake occurs (Gao and Crampin, 2004). Logarithms of the durations of increases (stress-accumulation) and decreases (crack-coalescence) are each separately proportional (self-similar) to impending earthquake magnitudes (Crampin et al., 2008), similar to the linearity of the Gutenberg–Richter relationship. Consequently, shear-waves above swarms of small earthquakes can be used as "stress-measuring stations" to monitor such stress-accumulation and stress-relaxation before impending earthquakes.

Peacock et al. (1988) were the first to recognize such changes in shear-wave time-delays (now identified with stress-accumulation), using seismic station KNW as a stress- measuring station, before a $M = 6$ earthquake near the San Andreas Fault in California. A few other examples were found worldwide (reviewed by Crampin and Peacock, 2008), but the significant advances came when persistent seismicity in Iceland become available for analysis online. Volti and Crampin (2003) displayed increases of Band-1 time-delays before five earthquakes in Iceland monitoring stress-accumulation. The time, magnitude, and fault-plane of a fifth earthquake were successfully stress-forecast 3 days before it occurred in a comparatively tight time/magnitude window (Crampin et al., 1999, 2008).

Crack-coalescence (stress-relaxation), before the impending earthquake occurs when the increasing stress-accumulation abruptly begin to decrease, was first recognized by Gao and Crampin (2004). Since then, stress-accumulation has been recognized before 15 earthquakes worldwide, of which nine had sufficient seismicity beneath stress-measuring stations to also display crack-coalescence (Crampin and Peacock, 2008).

Unfortunately, swarms of small earthquakes are far too scarce and irregular to be used as stress-measuring stations for reliable routine stress-forecasting. Reliable stress-forecasting requires three 1–1.5 km-deep borehole Stress-Monitoring Sites (SMSs), where a Downhole-Orbital Vibrator source (DOV) (Leary and Walter, 2005) radiate shear-waves to borehole geophones in Band-1 directions. A prototype SMS in non-optimal source-to-geophone geometry between existing boreholes in Iceland recorded exceptional sensitivity to low-level seismicity equivalent to a $M = 3.5$ earthquake at 70 km-distance (Crampin et al., 2003) demonstrating that SMSs have sufficient sensitivity to stress-forecast times and magnitudes of $M = 5$ earthquakes up to 1,000 km from impending source zones (Crampin and Gao, 2010).

Summary

Shear-wave splitting in the Earth reveals a fundamentally new understanding of fluid-rock deformation, where

APE-modeling shows that modifications to shear-wave splitting directly monitor stress-induced modifications to microcrack geometry. Since if the changing conditions can be specified, the deformation can be calculated and predicted by APE; this is a substantial advance, a New Geophysics, on conventional subcritical solid-earth geophysics. Changes in microcrack geometry can be monitored, calculated, predicted, and in some circumstances potentially controlled. There is extreme sensitivity to initial conditions, and universality, so that effects are widely and uniformly distributed.

The New Geophysics allows stress-accumulation and stress-relaxation (crack-coalescence) before earthquakes to be recognized and impending large earthquakes stress-forecast. There are many important implications and applications. In particular, any solid-earth application that cannot accommodate compliant stress-aligned fluid-saturated EDA-microcracks pervading most in situ rocks is in error and may lead to serious misunderstandings. New Geophysics leading to monitorability; calculability; predictability; potential controllability; universality; extreme sensitivity; and earthquake stress-forecasting, is likely to be the most fundamental advance in solid-earth geoscience for many decades.

Bibliography

Alford, R. M., 1986. Shear data in the presence of azimuthal anisotropy: Dilley, Texas. *56th Annual International Society of Exploration Geophysicists Meeting, Houston, Expanded Abstracts,* 476–379.

Anderson, D. L., 1961. Elastic wave propagation in layered anisotropic media. *Journal of Geophysical Research*, **66**, 2953–2963.

Ando, M., Ishikawa, Y., and Wada, H., 1980. *S*-wave anisotropy in the upper mantle under a volcanic area in Japan. *Nature*, **286**, 43–46.

Angerer, E., Crampin, S., Li, X.-Y., and Davis, T. L., 2002. Processing, modelling, and predicting time-lapse effects of overpressured fluid-injection in a fractured reservoir. *Geophysical Journal International*, **149**, 267–280. Available at file http://www.geos.ed.ac.uk/homes/scrampin/opinion/.

Ben-Ismail, W., Barruol, G., and Mainprice, D., 2001. The Kaapvaal craton seismic anisotropy: petrophysical analyses of upper mantle kimberlite nodules. *Geophysical Research Letters*, **28**, 2497–2500.

Booth, D. C., and Crampin, S., 1985. Shear-wave polarizations on a curved wavefront at an isotropic free-surface. *Geophysical Journal of the Royal Astronomical Society*, **83**, 31–45.

Bush, I., and Crampin, S., 1991. Paris Basin VSPs: case history establishing combinations of matrix- and crack-anisotropy from modelling shear wavefields near point singularities. *Geophysical Journal International*, **107**, 433–447.

Chapman, M., 2003. Frequency dependent anisotropy due to meso-scale fractures in the presence of equant porosity. *Geophysical Prospecting*, **51**, 369–379.

Crampin, S., 1977. A review of the effects of anisotropic layering on the propagation of seismic waves. *Geophysical Journal of the Royal Astronomical Society*, **49**, 9–27.

Crampin, S., 1978. Seismic wave propagation through a cracked solid: polarization as a possible dilatancy diagnostic. *Geophysical Journal of the Royal Astronomical Society*, **53**, 467–496.

Crampin, S., 1981. A review of wave motion in anisotropic and cracked elastic-media. *Wave Motion*, **3**, 343–391.

Crampin, S., 1990. The scattering of shear waves in the crust. *Pure and Applied Geophysics*, **132**, 67–91.

Crampin, S., 1991. Effects of point singularities on shear-wave propagation in sedimentary basins. *Geophysical Journal International*, **107**, 531–543.

Crampin, S., 1994. The fracture criticality of crustal rocks. *Geophysical Journal of the Royal Astronomical Society*, **118**, 428–438. Available at file http://www.geos.ed.ac.uk/homes/scrampin/opinion/.

Crampin, S., 1999. Calculable fluid-rock interactions. *Journal of the Geological Society*, **156**, 501–514. Available at file http://www.geos.ed.ac.uk/homes/scrampin/opinion/.

Crampin, S., 2003. Aligned cracks not LPO as the cause of mantle anisotropy, *EGS-AGU-EUG* joint assembly, nice. *Geophysical Research Abstracts*, **5**, 00205. Available at file http://www.geos.ed.ac.uk/homes/scrampin/opinion/.

Crampin, S., 2006. The new geophysics: a new understanding of fluid-rock deformation. In Van Cotthem, A., et al. (eds.), *Eurock 2006: Multiphysics Coupling and Long Term Behaviour in Rock Mechanics*. London: Taylor and Francis, pp. 539–544. Available at file http://www.geos.ed.ac.uk/homes/scrampin/opinion/.

Crampin, S., and Gao, Y., 2010. A review of earthquake stress-forecasting. *Proceedings of the International Symposium on Earthquake Seismology and Earthquake Prediction, Beijing, 2009, Pure and Applied Geophysics,* in press.

Crampin, S., and King, D. W., 1977. Evidence for anisotropy in the upper mantle beneath Eurasia from generalized higher mode surface waves. *Geophysical Journal of the Royal Astronomical Society*, **49**, 59–85.

Crampin, S., and Kirkwood, S. C., 1981. Velocity variations in systems of anisotropic symmetry. *Journal of Geophysics*, **49**, 35–42.

Crampin, S., and Peacock, S., 2005. A review of shear-wave splitting in the compliant crack-critical anisotropic Earth. *Wave Motion*, **41**, 59–77. Available at file http://www.geos.ed.ac.uk/homes/scrampin/opinion/.

Crampin, S., and Peacock, S., 2008. A review of the current understanding of shear-wave splitting and common fallacies in interpretation. *Wave Motion*, **45**, 675–722. Available at file http://www.geos.ed.ac.uk/homes/scrampin/opinion/.

Crampin, S., and Yedlin, M., 1981. Shear-wave singularities of wave propagation in anisotropic media. *Journal of Geophysics*, **49**, 43–46.

Crampin, S., and Zatsepin, S. V., 1995. Production seismology: the use of shear waves to monitor and model production in a poro-reactive and interactive reservoir. *65th Annual International Society of Exploration Geophysicists Meeting, Houston, Expanded Abstracts,* 199–202.

Crampin, S., and Zatsepin, S. V., 1997. Modelling the compliance of crustal rock: II - response to temporal changes before earthquakes. *Geophysical Journal International*, **129**, 495–506. Available at file http://www.geos.ed.ac.uk/homes/scrampin/opinion/.

Crampin, S., Evans, R., Üçer, B., Doyle, M., Davis, J. P., Yegorkina, G. V., and Miller, A., 1980. Observations of dilatancy-induced polarization anomalies and earthquake prediction. *Nature*, **286**, 874–877.

Crampin, S., Evans, R., Doyle, M., and Davis, J. P., 1981. Comments on papers about shear-wave splitting in dilatancy-induced anisotropy by Gupta, I. N., and by Ryall, A., and Savage, W. U. *Bulletin. Seismological Society of America*, **71**, 375–377.

Crampin, S., Evans, R., and Atkinson, B. K., 1984. Earthquake prediction: a new physical basis. *Geophysical Journal of the Royal Astronomical Society*, **76**, 147–156.

Crampin, S., Volti, T., and Stefánsson, R., 1999. A successfully stress-forecast earthquake. *Geophysical Journal International*, **138**, F1–F5.

Crampin, S., Volti, T., Chastin, S., Gudmundsson, A., and Stefánsson, R., 2002. Indication of high pore-fluid pressures in a seismically-active fault zone. *Geophysical Journal International*, **151**, F1–F5. Available at file http://www.geos.ed.ac.uk/homes/scrampin/opinion/.

Crampin, S., Chastin, S., and Gao, Y., 2003. Shear-wave splitting in a critical crust: III - preliminary report of multi-variable measurements in active tectonics. *Journal of Applied Geophysics*, **54**, 265–277. Special Issue. Available at file http://www.geos.ed.ac.uk/homes/scrampin/opinion/.

Crampin, S., Gao, Y., and Peacock, S., 2008. Stress-forecasting (not predicting) earthquakes: A paradigm shift? *Geology*, **36**, 427–430. Available at file http://www.geos.ed.ac.uk/homes/scrampin/opinion/.

Davies, P., 1989. The new physics: a synthesis. In Davies, P. (ed.), *The New Physics*. Cambridge: Cambridge University Press, pp. 1–6.

Evans, R., 1984. Effects of the free surface on shear wavetrains. *Geophysical Journal of the Royal Astronomical Society*, **76**, 165–172.

Faccenda, M., Gerya, T. V., and Burlini, L., 2009. Deep slab hydration induced by bending-related variations in tectonic pressure. *Nature Geoscience*, **2**, 790–793.

Fouch, M. J., and Rondenary, S., 2006. Seismic anisotropy beneath stable continental interiors. *Physics of the Earth and Planetary Interiors*, **258**, 292–320.

Gao, Y., and Crampin, S., 2004. Observations of stress relaxation before earthquakes. *Geophysical Journal International*, **157**, 578–582. Available at file http://www.geos.ed.ac.uk/homes/scrampin/opinion/.

Gupta, I. N., 1973. Premonitory variations in S-wave velocity anisotropy before earthquakes in Nevada. *Science*, **182**, 1129–1132.

Helbig, K., 1994. Foundations of anisotropy for exploration seismics. In *Handbook of Geophysical Exploration, Section 1, Seismic Exploration*. Oxford: Elsevier Science, p. 248.

Helbig, K., and Thomsen, L., 2005. 75-plus years of anisotropy in exploration and reservoir seismics: A historical review of concepts and methods. *Geophysics*, **70**, 9ND–23ND.

Hess, H., 1964. Seismic anisotropy of the uppermost mantle under oceans. *Nature*, **203**, 629–631.

Karato, S.-I., and Jung, H., 1998. Water, partial melting and the origin of the seismic low velocity and high attenuation zone in the upper mantle. *Earth and Planetary Science Letters*, **157**, 193–207.

Karato, S.-I., and Wu, P., 1993. Rheology of the upper mantle: a synthesis. *Science*, **260**, 771–778.

Katayama, I., Karata, S.-I., and Brandon, M., 2005. Evidence of high water content in the deep upper mantle inferred from deformation microstructures. *Geology*, **33**, 613–616.

Keith, C. M., and Crampin, S., 1977. Seismic body waves in anisotropic media: synthetic seismograms. *Geophysical Journal of the Royal Astronomical Society*, **49**, 225–243.

Leary, P. C., and Walter, L. A., 2005. Physical model for the downhole orbital vibrator (DOV) – I. Acoustic and borehole seismic radiation. *Geophysical Journal International*, **163**, 647–662.

Liu, E., Crampin, S., Queen, J. H., and Rizer, W. D., 1993a. Velocity and attenuation anisotropy caused by microcracks and macrofractures in multiazimuthal reverse VSPs. *Canadian Journal of Exploration Geophysics*, **29**, 177–188.

Liu, E., Crampin, S., Queen, J. H., and Rizer, W. D., 1993b. Behaviour of shear waves in rocks with two sets of parallel cracks. *Geophysical Journal International*, **113**, 509–517.

Long, M. D., and Silver, P. G., 2009. Shear wave splitting and mantle anisotropy: measurements, interpretations, and new directions. *Surveys in Geophysics*, **30**, 407–461.

Love, A. E. H., 1927. *A Treatise of the Mathematical Theory of Elasticity*. New York: Dover. 1944 reprinted.

Maultzsch, S., Chapman, M., Liu, E., and Li, X.-Y., 2003. Modelling frequency dependent seismic anisotropy in fluid-saturated rock with aligned fractures: implication of fracture size estimation from anisotropic measurements. *Geophysical Prospecting*, **51**, 381–392.

Maupin, V., and Park, J., 2007. Theory and observations - wave propagation in anisotropic media. In Schubert, G. (ed.), *Treatise on Geophysics*. Amsterdam: Elsevier, pp. 289–321, doi:10.1016/B978-044452748-6/00007-9.

Nur, A., and Simmons, G., 1969. Stress-induced anisotropy in rock: an experimental study. *Journal of Geophysical Research*, **74**, 6667–6674.

Nye, J. F., 1957. *Physical Properties of Crystals*. Oxford: Clarendon.

Peacock, S., Crampin, S., Booth, D. C., and Fletcher, J. B., 1988. Shear-wave splitting in the Anza seismic gap, southern California: temporal variations as possible precursors. *Journal of Geophysical Research*, **93**, 3339–3356.

Postma, G. W., 1955. Wave propagation in a stratified medium. *Geophysics*, **20**, 780–806.

Raitt, R. W., Shor, G. G., Francis, T. J. G., and Morris, G. B., 1969. Anisotropy of the Pacific upper mantle. *Journal of Geophysical Research*, **74**, 3095–3109.

Savage, M. K., 1999. Seismic anisotropy and mantle deformation: What have we learned from shear wave splitting? *Reviews of Geophysics*, **37**, 65–106.

Silver, P. G., and Chan, W. W., 1988. Implications for continental structure and evolution from seismic anisotropy. *Nature*, **335**, 34–39.

Silver, P. G., and Chan, W. W., 1991. Shear wave splitting and subcontinental mantle deformation. *Journal of Geophysical Research*, **96**, 16,429–16,454.

Slater, C., Crampin, S., Brodov, L. Y., and Kuznetsov, V. M., 1993. Observations of anisotropic cusps in transversely isotropic clay. *Canadian Journal of Exploration Geophysics*, **29**, 216–226.

Stefánsson, R., et al., 1993. Earthquake prediction research in the South Iceland Seismic Zone and the SIL Project. *Bulletin Seismological Society of America*, **83**, 696–716.

Thomsen, L., 1986. Weak elastic anisotropy. *Geophysics*, **51**, 1954–1966.

Volti, T., and Crampin, S., 2003. A four-year study of shear-wave splitting in Iceland: 2 Temporal changes before earthquakes and volcanic eruptions. *Geological Society, London, Special Publication*, **212**, 135–149. Available at file http://www.geos.ed.ac.uk/homes/scrampin/opinion/.

Wild, P., and Crampin, S., 1991. The range of effects of azimuthal isotropy and EDA-anisotropy in sedimentary basins. *Geophysical Journal International*, **107**, 513–529.

Zatsepin, S. V., and Crampin, S., 1997. Modelling the compliance of crustal rock: I - response of shear-wave splitting to differential stress. *Geophysical Journal International*, **129**, 477–494. Available at file http://www.geos.ed.ac.uk/homes/scrampin/opinion/.

Cross-references

Seismic Anisotropy
Seismic Tomography
Seismic Wave Propagation in Real Media: Numerical Modeling Approaches
Seismic Waves, Scattering
Seismic, Receiver Function Technique
Seismic, Viscoelastic Attenuation

SINGLE AND MULTICHANNEL SEISMICS

Tamás Tóth
Geomega Limited, Budapest, Hungary

Definition

Seismics or Seismic survey. The method of imaging the subsurface with reflected and refracted seismic waves, which are generated by controlled sources and sensed by seismic detectors.

Single-channel seismics. Seismic survey using one detector (channel) for sensing returning seismic waves.

Multichannel seismics. Seismic survey using multiple detectors (channels) at different locations for sensing returning seismic waves generated by each source points.

Introduction

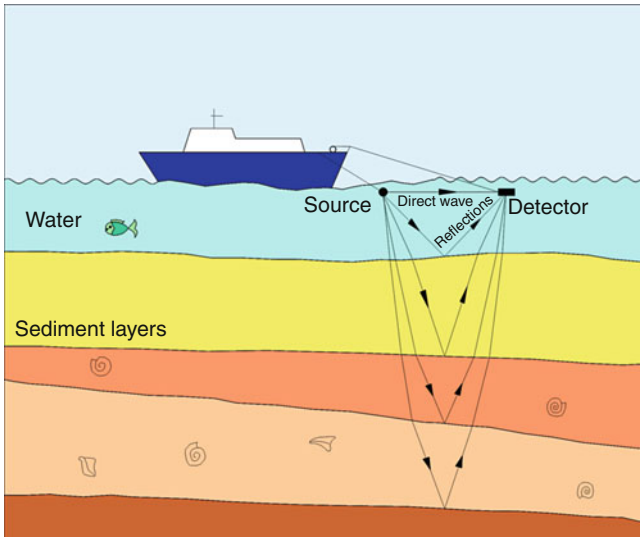
Reflection and refraction seismology has been applied for imaging the subsurface since the 1920s (Karcher, 1987 and Keppner, 1991) and is still one of the most powerful methods of exploration geophysics. Applied technology has evolved considerably over the decades, but the principles have not changed. Imaging is performed by elastic waves generated by controlled sources, such as explosive, vibroseis, air gun, sparker, boomer, or similar sources. Generated waves propagate through the subsurface layers and part of them get reflected or refracted back from layer boundaries. These returning waves are sensed on the surface by seismic detectors, geophones on land, and hydrophones in water.

In case of single-channel seismic recording, elastic waves generated by each shot are recorded by one single detector or a group of detectors. On the other hand, multichannel seismic recording utilizes multiple detectors or groups of detectors located at different positions for recording the wavefield generated by each shot. Multichannel seismics provide several advantages over single-channel seismics; however, simplicity of the recording equipment and high resolution of the recorded seismic profiles make single-channel seismics still a powerful and cost-effective tool for shallow investigations.

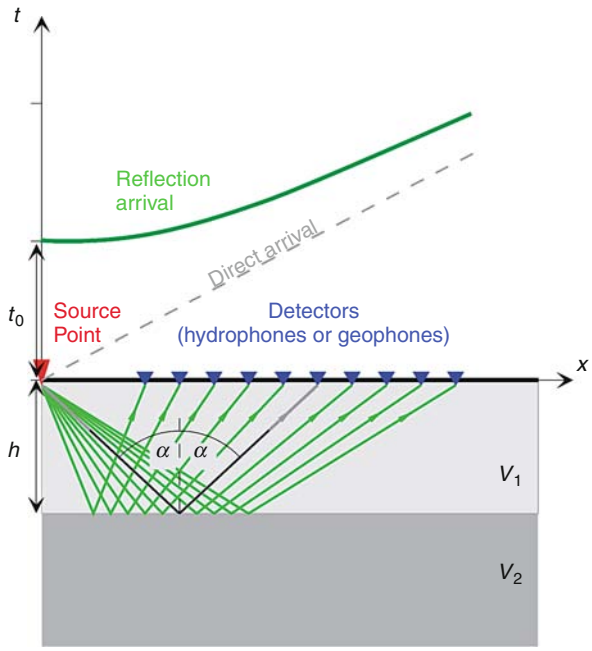
Single-channel seismics

Simplest possible configuration for a seismic reflection survey is single-channel recording. Reflected energy from each shot is sensed by only one receiver or one group of receivers, normally positioned at a constant distance from the source (see Figure 1).

The source–receiver pair is moved, and each shot generates one seismic trace for the profile. In case of single-channel recording, each shot provides information from one subsurface location only, therefore it is critical to have a low cost and high repetition rate source. Shot generation is more labor intensive on land; hence, single-channel recording is hardly used for onshore surveys. Situation is



Single and Multichannel Seismics, Figure 1 During single-channel recording, seismic energy generated by each source point images one subsurface point of each layer boundary, generating one seismic trace.



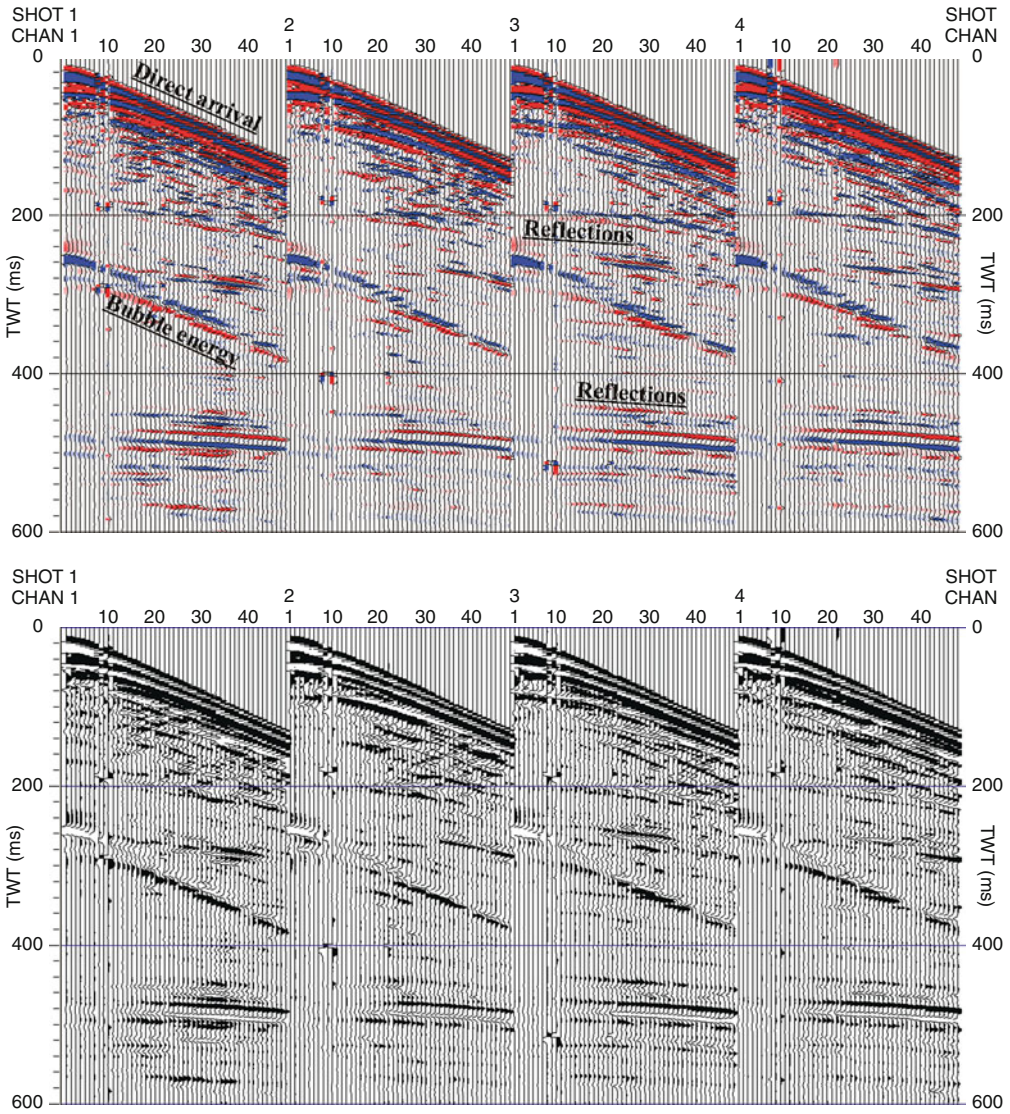
Single and Multichannel Seismics, Figure 2 Reflected energy generated by each source point is recorded by multiple detectors. Each detector images a different subsurface point. Reflection at each subsurface point follows Snell's law. Angles of incidence and reflection (marked α) are equal; hence, distance between the subsurface reflection points is half of the distance between the detectors. Upper part of the Figure shows the theoretical hyperbola of the reflection arrivals and the direct arrival between source point and detectors.

more favorable for offshore surveys, as easy-to-operate sources with high repeatability are available and make single-channel recording an attractive alternative. This is especially true, when relatively shallow "penetration" is sufficient, but *high resolution* is needed.

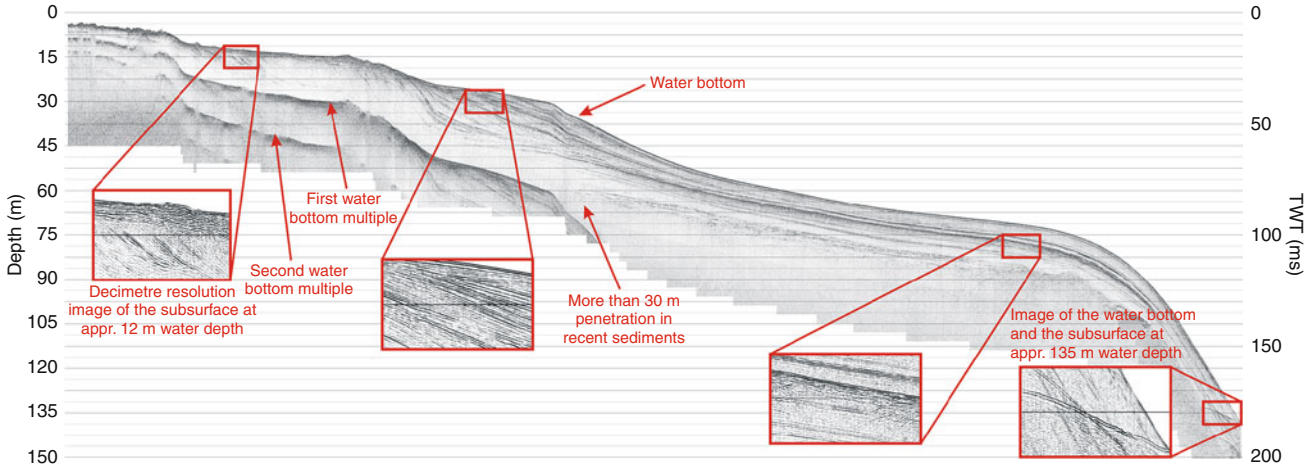
Penetration, on the one hand, strongly depends on the seismic source used; on the other hand, it is also determined by the physical properties of the subsurface strata. Seismic *resolution* is a function of the frequency content of the source used. It is important to note that penetration and resolution are not independent parameters. Higher frequency sources ensure maximum resolution, but as high frequencies attenuate quicker, they provide less penetration. Highest frequency sources are Compressed High Intensity Radar Pulse *(CHIRP)* and boomers, while sparkers and small volume air gun sources with their higher energy provide deeper penetration, but generally lower resolution.

Multichannel seismics

Several advantages are gained by recording the reflected energy of each shot by multiple detectors, which is called multichannel seismic recording (see sketch in Figure 2 and sample shot gathers in Figure 3). The most important of



Single and Multichannel Seismics, Figure 3 The figure shows four subsequent shot gathers recorded by a 48-channel recording system. Upper series displays the shot gathers in red-blue color display with wiggle trace overlay, while the lower series shows the same traces in wiggle trace – varying area (WT/VA) display. Seismic energy was generated by an air gun and the direct arrival, bubble energy and reflections are clearly visible on all of four shot gathers.



Single and Multichannel Seismics, Figure 4 Single-channel boomer profile recorded in the Bay of Napoli showing decimeter resolution and more than 30 m penetration. Comparing the left and right inserts, the attenuation of high frequencies and decreasing of seismic resolution with depth are also visible. High resolution of the seismic combined with gravity coring enabled exceptional dating along the recorded profiles. Details of the research can be found in Sacchi et al., 2009.

these advantages are multifold imaging of the same location, called the common depth point (CDP) and the possibility to estimate seismic velocity from the recorded data itself. Multifold imaging enhances the signal/noise ratio of the final seismic section through stacking. The best possible estimate of velocity field of the subsurface layers is necessary for normal move out *(NMO) correction* and "*migration*" before and after stacking. This information is derived from the parameters of the hyperbola fitted on the reflections in the CDP gathers, a process called "velocity analysis."

Commonly used high-resolution single- and multichannel seismic sources

The CHIRP systems use sweep signals for generating pressure waves. Electromagnetic sweep signals are commonly used in commercial and military radar systems, while acoustic sweeps are used in sonar systems for marine applications, and for *vibroseis* on land. CHIRP systems (e.g., Schock and LeBlanc, 1990; Gutowski et al., 2002) became very popular in marine surveys, while vibroseis is the most commonly used seismic source on land. Marine CHIRP systems typically operate in the 400 Hz–25 kHz frequency range, providing decimeter resolution of the sub-bottom layers down to a few meters or maximum few tens of meters depth in unconsolidated sediments. Advantage of the CHIRP systems is the high repeatability and good S/N (signal-to-noise) ratio. Disadvantage is the loss of phase information of the signal due to the cross-correlation during signal recording. CHIRP sources are used mainly for single-channel seismic recording; however, high-resolution 3D Chirp system has also been developed and used for decimeter-scale object detection (Vardy et al., 2008; Plets et al., 2009).

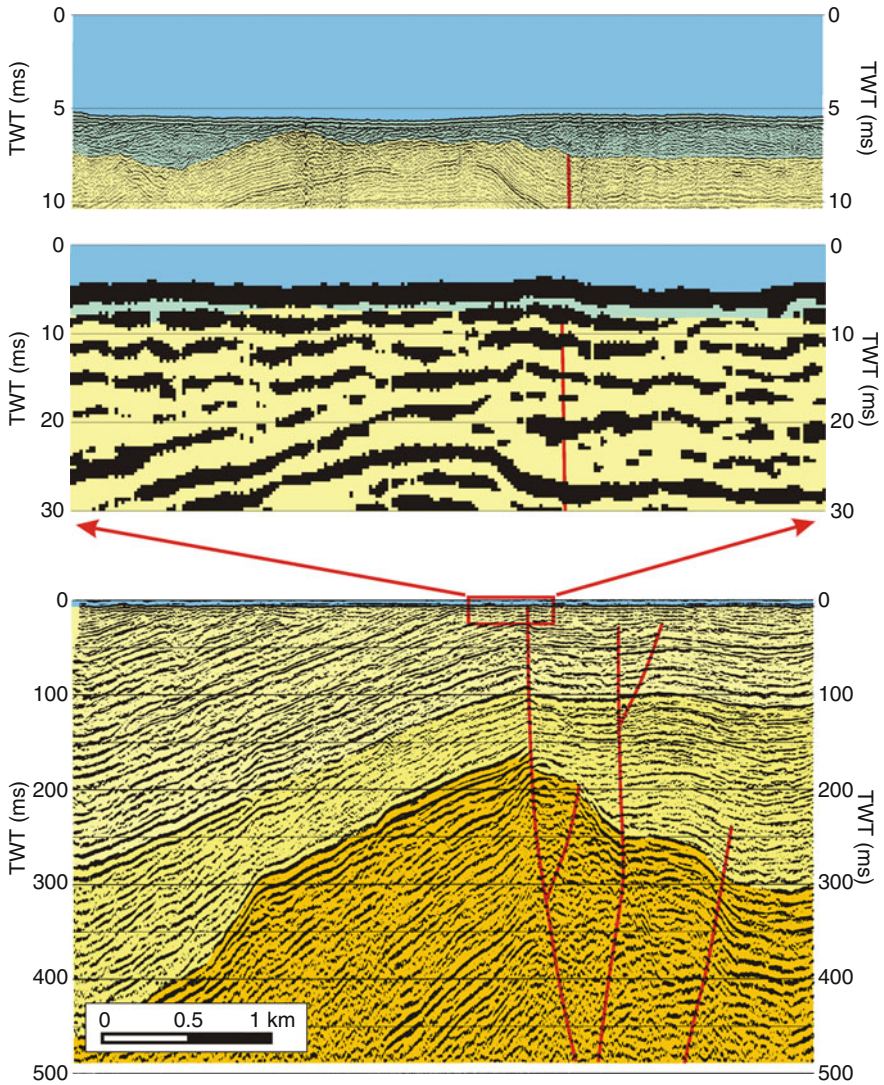
*Boomer*s are electromagnetically driven sources consisting of a flat coil and a metal plate below the coil (Edgerton and Hayward, 1964). Discharge of a high-voltage capacitor bank through the coil generates Eddy-current in the metal plate. The generated Eddy-current violently repels the plate from the coil. Pulling back the plate by strong springs or rubber bands creates a cavitation in the water acting as a sound source. Energy of the source depends on the capacitor bank, which is typically ranging between few 100 and 1,000 J, but several kJ systems have also been used. Boomers generate highly repeatable source signatures in the frequency range of 300 Hz–20 kHz with decimeter resolution and several tens of meters penetration. Innovative system design, such as the IKB-Seistec system utilizing a boomer source and a hydrophone group shielded by a focusing cone (Simpkin and Davis, 1993), can further improve S/N ratio of boomer-based systems providing exceptional combination of resolution and penetration, as shown in Figure 4. Boomers used to be typical single-channel seismic sources; however, multichannel systems developed for high-resolution 3D data acquisition are also built around boomer sources (Müller et al., 2009).

In *sparker* sources, similar to the sparking plug of an engine, discharge of a capacitor bank creates a spark between the positive and negative electrodes of the sparker. This spark vaporizes water between the electrodes and generates a pressure impulse. Energy and shape of the sparker wavelet are influenced by the physical design of the sparker, but is mainly controlled by the capacitance and the voltage of the high-voltage capacitor bank, commonly called "shot box." Typical energy values per shot range between few 100 J and several thousand or even tens of thousands of Joules (Allen, 1972). Large energy sparkers are rarely used nowadays as air gun sources mainly replaced them. Small-energy sparkers are still in use for engineering and research applications as they provide an alternative source fitting between boomers and small air guns. Sparkers

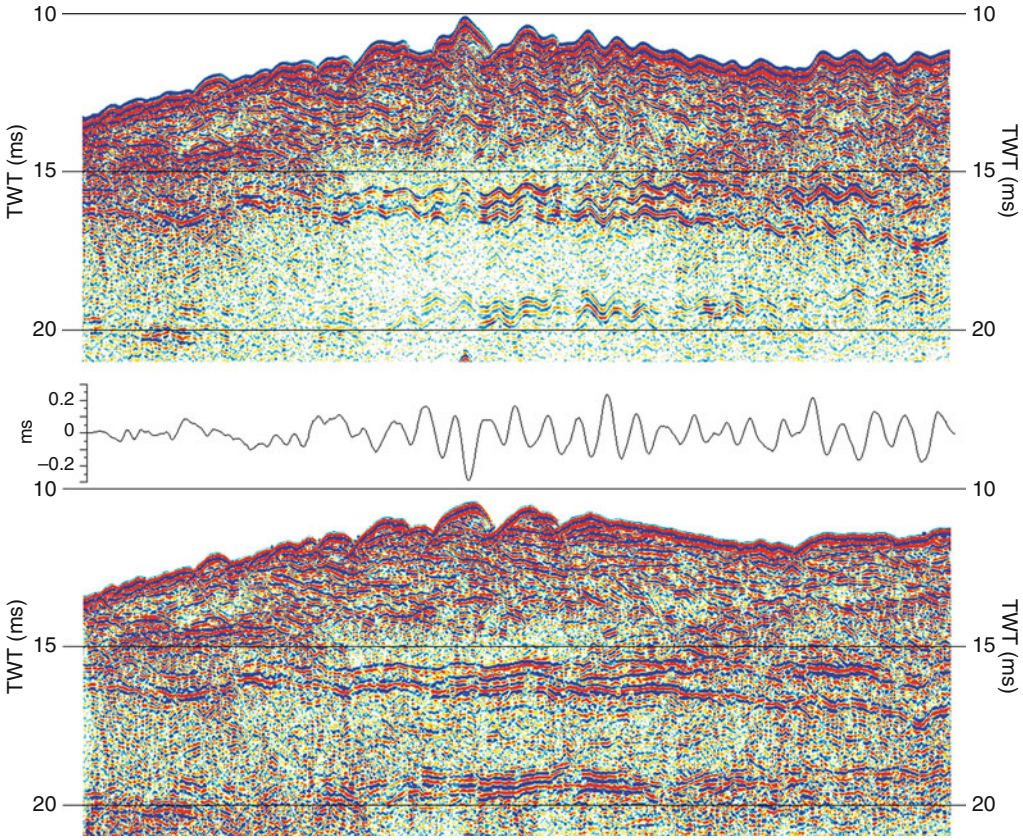
are cost-effective to run, but shape of the wavelet, directivity, and repeatability are better controlled for other sources. Use of the sparker is also limited by the conductivity of the water. It is normally used in saltwater environment as in fresh water, extra care has to be taken to increase the conductivity of the fluid between the electrodes. This can be achieved, for example, by wrapping the sparker in a plastic bag and adding salt inside.

Air gun is a pneumatic seismic energy source generating pressure waves by the sudden release of high-pressure (70–200 bar) air into the water (Giles, 1968). The gun is charged with high-pressure air from a compressor or air container and explosively releases the air from its chamber. The release of the high-pressure air produces a shock wave followed by bubble forming and several oscillations resulting from the expansion and collapse of the air bubble. Good example of bubble energy from the first oscillation is shown in Figure 3.

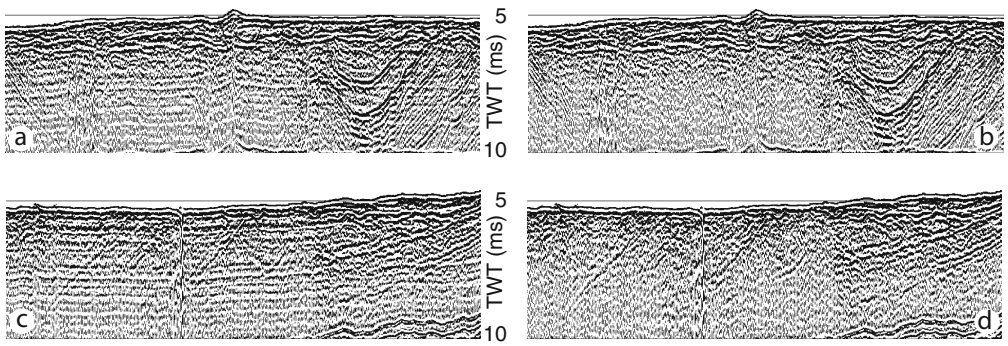
Energy and frequency content of the air gun are determined by the volume of the gun, the pressure of the air and tow depth of the gun. High-resolution surveys require the smallest volume guns, typically 5–20 in^3 (1 in^3 = 16.39 cm^3); however, gun clusters of 1,000–10,000 in^3 are also commonly used during surveys with deeper targets. High-frequency content of small air guns can exceed 1,000 Hz, and large volume guns or gun clusters may



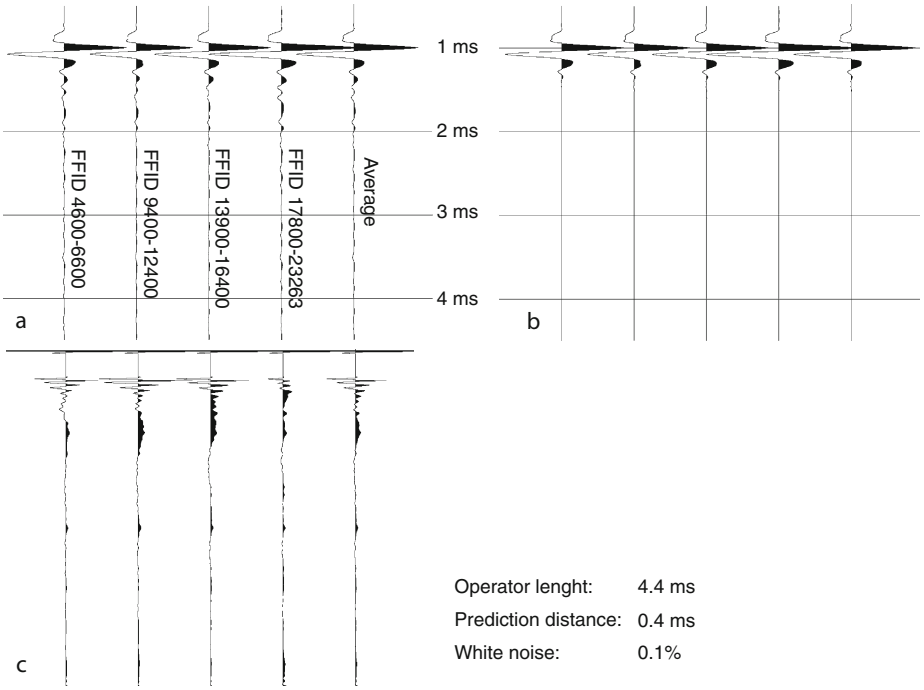
Single and Multichannel Seismics, Figure 5 Bottom image shows a multichannel seismic profile recorded on river Danube, Hungary. Middle image is an optical zoom of the area in the red rectangle. Top image shows the corresponding single-channel boomer profile recorded along the same section of the river. Blue color indicates water, green, recent river sediments, while shades of yellow, Pliocene-Pleistocene sediments. Note the order of magnitude difference in resolution but also in penetration between the two profiles.



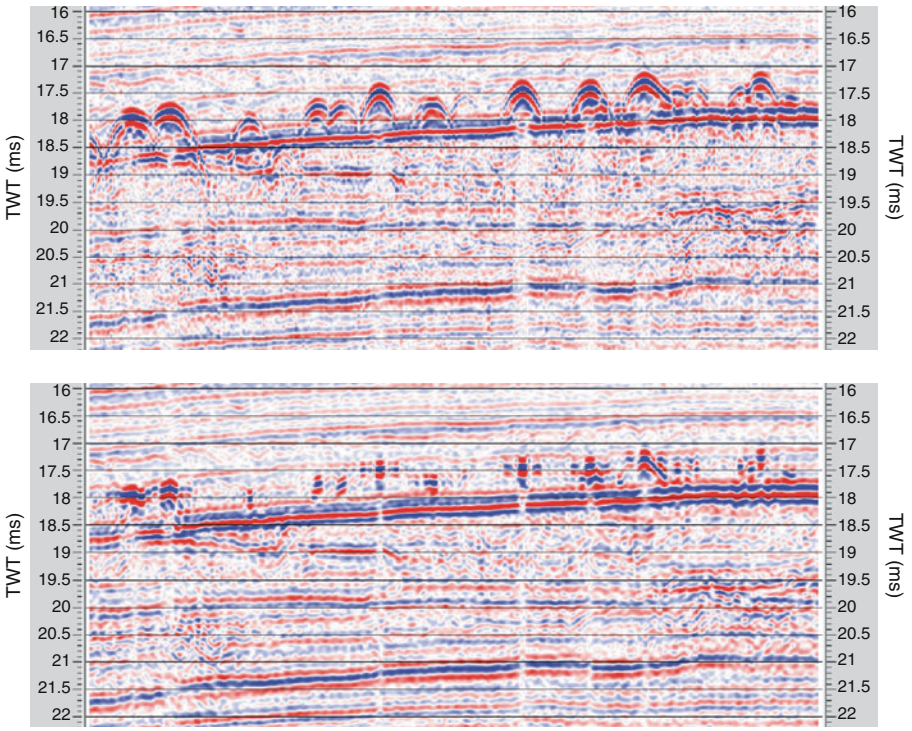
Single and Multichannel Seismics, Figure 6 Top image shows single-channel boomer data recorded on river Rhine with several decimeter amplitude waces. Wave action has been estimated from the cross-correlation of the seismic tarces and the necessary time shift is displayed in the middle graph. Lower image shows the seismic profile after correction for surface wave action. Note that the velocity pull-up and push-down effects due to the velocity difference between water and sediments are clearly visible along the reflector just below 15 ms in the lower image.



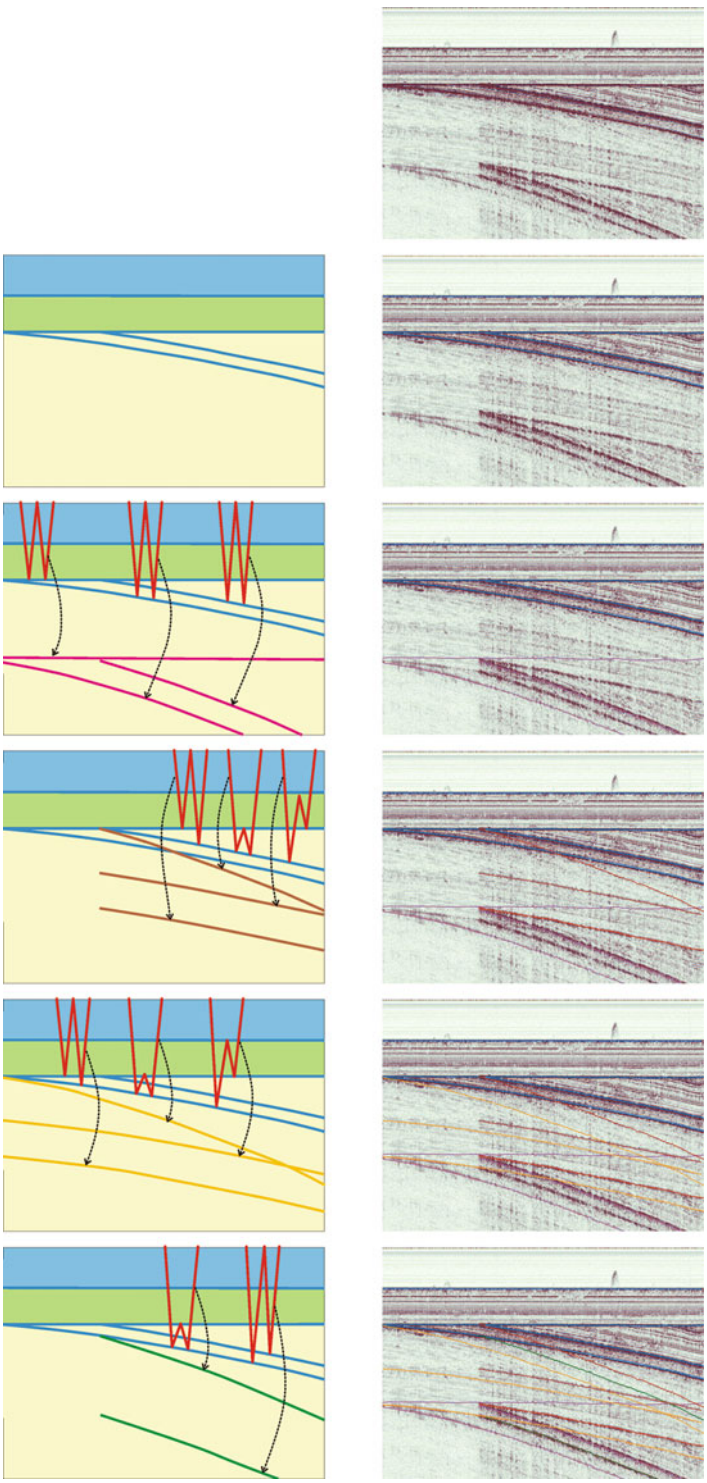
Single and Multichannel Seismics, Figure 7 Two images on the left show profiles without wavelet deconvolution applied. Note the "ringing" of the wavelet running parallel with the water bottom and intersecting genuine reflectors below the water bottom. Profiles on the right have been deconvolved using wavelet-based deterministic deconvolution. The ringing of the wavelet is completely eliminated.



Single and Multichannel Seismics, Figure 8 Wavelets before (**a**) and after (**b**) deterministic deconvolution applied and the deconvolution operator applied (**c**).



Single and Multichannel Seismics, Figure 9 Upper and lower images show the same detail of the seismic profile before and after migration, respectively. Note the collapse of the diffraction hyperbolae around 18 ms TWT. Lateral resolution of the profile is significantly improved by the migration process.



Single and Multichannel Seismics, Figure 10 Top right image shows the original seismic profile recorded in lake Balaton, Hungary. In the lake, approx. 4 m water and 4 m soft, unconsolidated mud can be found above the older Pliocene strata. This, together with the dipping strata in the Pliocene, presents an ideal condition for multiple generation. Top sketch marks with blue the interfaces considered for multiple generation. These are: water surface, water bottom, mud bottom, and two dipping intra-Pliocene strata. First, the free surface multiples (marked by *pink*) are calculated for each of these interfaces. This is followed by different peg-leg multiples calculated and marked by brown, yellow, and green lines on the profile. It is very interesting to observe that very many of the reflections recorded on the seismic profile are actually multiple energies generated by the above mentioned surfaces only.

provide enough energy for several kilometer deep penetration. Air guns are highly repeatable sources with considerable literature on measuring and designing their wavelet (Ziolkowski et al., 1982).

Bubble oscillation can be decreased or even completely diminished by the timely injection of a second volume of high-pressure air. Guns operating under this principle are called *Generated Injection Air Guns* or simply *GI guns*. GI guns have two chambers. The primary chamber (generator) produces the actual pressure impulse, while the secondary chamber (injector) is used to inject a second air volume near the maximum expansion of the first bubble in order to prevent its collapse.

Air gun and *GI* sources due to their higher energy and deeper penetration are mainly used for multichannel recording. A good comparison of single-channel boomer profile and a multichannel air gun profile is shown in Figure 5.

Processing considerations for high-resolution single- and multichannel seismic profiles

Static correction is always a key issue for proper imaging of the subsurface with seismic waves. This is especially true for high-resolution surveys. Offshore seismic surveys present a special static problem due to the wave motion. This equally effects single- and multichannel surveys, but the higher the resolution of the seismics, the more severe the problem can be. Coherency of the reflections in a decimeter resolution boomer survey can be severely affected by surface waves. Heave motion detectors offer a compensation via measuring the wave motions together with the seismic profile; however, if no heave motion is available for the survey, the necessary correction can be calculated from the data. Example of this is shown in Figure 6.

High repeatability of many of the sources (boomer, air guns) offers good opportunity for wavelet determination and application of deterministic *deconvolution* using the calculated wavelet. This is illustrated in Figures 7 and 8 showing two seismic profiles before and after deterministic wavelet deconvolution and the calculated wavelets and deconvolution operators.

Migration is also a key processing step not only for the multichannel, but also for the single-channel seismic profiles. In the later case, estimation of the migration velocity is more complicated, as there is no a priori information available from stacking velocity analysis. Velocity estimates from other measurements or migration velocity analysis can be used for migrating single-channel data. An example of this is shown in Figure 9.

Multiples generated by the free water surface present a significant problem for single-channel seismic profiles. Several processing algorithms exist for multichannel seismic data in order to suppress multiple energy, but almost all of them fail in case of single-channel seismic profiles. Figure 10 shows a single-channel profile recorded in lake Balaton, Hungary with significant multiple energy.

Summary

Seismic surveys provide one of the most detailed image of the subsurface from shallow (few tens of meters) to deep (several kilometers or even tens of kilometers) intervals. Although most of the recent seismic surveys utilize multichannel recording, high-resolution single-channel seismic surveys can also provide a cost-effective solution for offshore surveys. Main advantage of single-channel seismics is high-resolution imaging and relative simplicity of the survey. Multichannel surveys, on the other hand, can significantly improve the signal-to-noise ratio of the seismic profile by applying multifold imaging and at the same time also provide additional information, for example, estimates of the velocity field.

Bibliography

Allen, F. T., 1972. Some characteristics of marine sparker seismic data. *Geophysics*, **37**, 462–470.

Edgerton, H. F., and Hayward, E. C., 1964. The Boomer sonar source for seismic profiling. *Journal of Geophysical Research*, **69**, 3033–3042.

Giles, B. F., 1968. Pneumatic acoustic energy source. *Geophysical Prospecting*, **16**, 21–53.

Gutowski, M., Bull, J., Hestock, T., Dix, J., Hogarth, P., Leighton, T., and White, P., 2002. Chirp sub-bottom profiler source signature design and field testing. *Marine Geophysical Researches*, **23**, 481–492.

Karcher, J. C., 1987. The reflection seismograph: its invention and use in the discovery of oil and gas fields. *The Leading Edge*, **6**(11), 10–19.

Keppner, G., 1991. Ludger mintrop. *The Leading Edge*, **10**(9), 21–28.

Müller, C., Woelz, S., Ersoy, Y., Boyce, J., Jokisch, T., Wendt, G., and Rabbel, W., 2009. Ultra-high-resolution marine 2D-3D seismic investigation of the Liman Tepe/Karantina Island archaeological site (Urla/Turkey). *Journal of Applied Geophysics*, **68**, 124–134.

Plets, R. M. K., Dix, J. K., Adams, J. R., Bull, J. M., Henstock, T. J., Gutowski, M., and Best, A. I., 2009. The use of a high-resolution 3D Chirp sub-bottom profiler for the reconstruction of the shallow water archeological site of the Grace Dieu (1439), River Hamble, UK. *Journal of Archeological Science*, **36**, 408–418.

Sacchi, M., Molisso, F., Violante, C., Esposito, E., Insinga, D., Lubritto, C., Porfido, S., and Tóth, T., 2009. Insights into flood-dominated fan-deltas: very high-resolution seismic examples off the Amalfi cliffed coast, eastern Tyrrhenian Sea. *Geological Society London, Special Publications*, **322**, 33–71.

Schock, S. G., and LeBlanc, L. R., 1990. Chirp sonar: new technology for sub-bottom profiling. *Sea Technology*, **31**(9), 35–43.

Simpkin, P. G., and Davis, A., 1993. For seismic profiling in very shallow water, a novel receiver. *Sea Technology*, **34**(9), 21–28.

Vardy, M. E., Dix, J. K., Henstock, T. J., Bull, J. M., and Gutowski, M., 2008. Decimeter-resolution 3D seismic volume in shallow water: a case study in small-object detection. *Geophysics*, **73**, B33–B40.

Ziolkowski, A., Parkes, G., Hatton, L., and Haugland, T., 1982. The signature of an air gun array. *Geophysics*, **47**, 1413–1421.

Cross-references

Seismic Data Acquisition and Processing
Seismic Velocity-Density Relationships
Seismic Velocity-Temperature Relationships
Seismic, Migration

SLOW EARTHQUAKE

Teruyuki Kato
Earthquake Research Institute, The University of Tokyo, Bunkyo-ku, Tokyo, Japan

Definition

An earthquake is a sudden rupture of a fault embedded in the crust or upper mantle, and radiates elastic waves toward the earth's surface to vibrate the ground. Ground vibration includes wide spectrum of frequencies from tens of hertz to hundredth of seconds or lower. The ground vibration is recorded by a pendulum called a seismograph. Since the pendulum has its own frequency characteristics, components of longer periods becomes less visible and are not recognizable in the longest period.

Some types of fault ruptures have predominant component in long-period seismogram and such earthquakes are called slow earthquakes. The earthquake that does not radiate enough energy to be recorded on seismogram is called a silent earthquake or slow slip event (SSE). In this definition, observed pre-seismic or post-seismic slip, or creep events may all be categorized in the silent earthquake or SSE.

Early studies on slow and silent earthquake

Earlier studies of slow earthquake were made by examining difference in magnitude determined by different frequency ranges (e.g., Kanamori, 1972; Kanamori and Stewart, 1979). Seismic magnitude of large earthquakes estimated from shorter period wave such as body waves are significantly smaller than that determined from longer period seismic waves. Ultimately, magnitude determined by seismic wave analysis is often much smaller than the moment magnitude or that estimated by the geodetic data inversion analysis that includes longest period (or permanent displacement) data. These have been interpreted such that the significant amount of moment release was achieved by the component of fault rupture that is slow enough not radiating much seismic wave. Kanamori and Stewart (1979), for example, showed that a series of such slow events were triggered after the June 6, 1960, Chilean earthquake (Mw 9.5).

One of such difference could be manifested by a type of earthquake called "tsunami earthquake." The tsunami earthquake is characterized by anomalously high tsunami generation than that expected from body wave or surface wave magnitude. This can be interpreted by source characteristic such that the fault rupture was slow enough not radiating shorter seismic wave but rapid enough to generate tsunami. The 1896 Sanriku, Japan, earthquake (Ms 7.2, Mw 8.0), 1992 Nicaragua earthquake (Ms 7.0, Mw 7.6) are examples (e.g., Kanamori and Kikuchi, 1993).

If the rupture is much slower and does not radiate any seismic energy, then it is called a silent earthquake. Beroza and Jordan (1990) tested if such very slow or silent earthquake could be detected using spectra of Earth's free oscillations. Another trial is the comparison of long-term moment release rate at subduction zones. If all of strain accumulated along plate boundaries is released by seismic energy, then the released moment at the time of earthquake should be equal to the accumulated moment. Kawasaki et al. (2001) examined this hypothesis and found that the total moment released by earthquakes along the plate boundary is far smaller than the totally accumulated moment in an interseismic stage. This means that there are other unknown slow processes that release remaining moment at the plate boundaries.

One of possible mechanisms of slow strain release would be post-seismic transient displacement. Earlier discovery of post-seismic crustal deformation was found due to the 1946 Nankai, Japan, earthquake (Mw 8.1) (e.g., Okada and Nagata, 1953). The cause of such a post-seismic crustal deformation was interpreted as a retarded slip after the earthquake (Fitch and Scholz, 1971). Kawasaki et al. (1995) asserted that a post-seismic slow rupture was generated after the 1992 Sanriku, Japan, earthquake (Ms 6.9), by examining the strain records observed at the Esashi, Northern Japan, station. However, the question if such post-seismic transients on the records of strain sensors might stem from instrumental drifts after a strong shake was not well resolved.

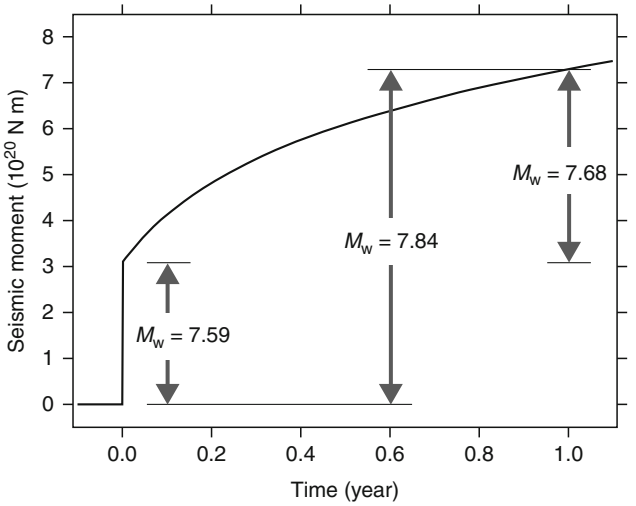
More direct and convincing finding of slow slip events was brought in 1990s by the advent of the Global Positioning System (GPS). Geographical Survey Institute of Japan (GSI) first deployed nationwide GPS network to monitor crustal deformation of the Japanese Islands starting in 1992. The network was much densified in 1996 with more than 600 permanent GPS sites together with near real-time monitoring facility. The network is called GPS Earth Observation Network System (GEONET). It has been further augmented to more than 1,200 sites all over the Japanese Islands. This nationwide array of continuous GPS observation provided such a powerful tool to discover slow slip events around the Japanese Islands, as described in the following section.

Also, the USA deployed a large array of GPS together with strain and tilt meters along the western coast of the North American continent and is called Plate Boundary Observatory (PBO), which also enabled discoveries of slow slip events.

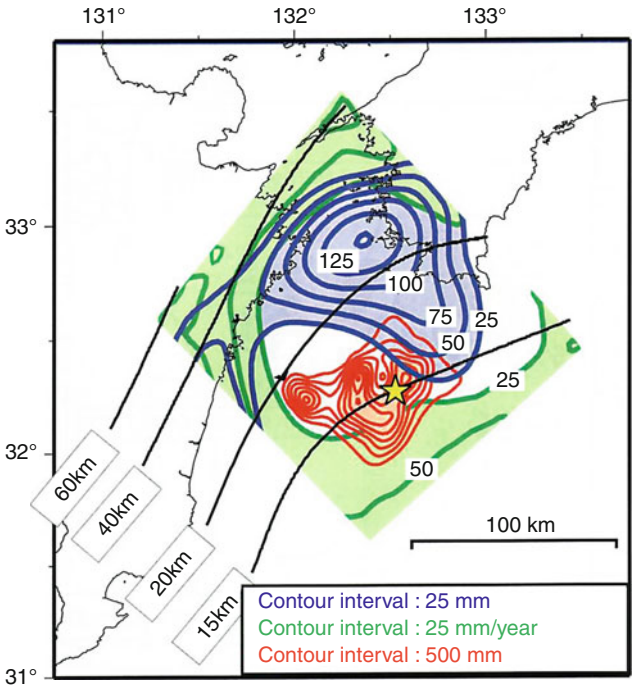
Rapid developments of GPS arrays all over the world, in particular along the Pacific Rim areas, enabled us to discover that silent earthquakes or slow slip events (SSE) are occurring at various areas around the glove, in particular, along the subducting plate interface. Comprehensive description on where do those slow events are taking place are found in Schwartz and Rokosky (2007) and Schwartz (2009).

Long-term slow slip events

A variety of slow slip events, including post-seismic slip, whose time durations are months to years have been discovered along the subducting plate boundaries.



Slow Earthquake, Figure 1 Moment release rate due to the 1994 Sanriku-Haruka-oki, Japan, earthquake. Figure suggests that the moment released in 1 year after the earthquake is bigger than the co-seismic moment release (Heki et al., 1997).



Slow Earthquake, Figure 2 Slip distributions on the subducting plate interface shown by equi-depth contours, projected on the earth's surface: (*Blue*) The slip distribution due to the 1997 slow event at the Bungo Channel, Japan (After Yagi and Kikuchi, 2003). (*Red*) the slip distribution due to the 1968 Hyuganada earthquake, and (*Green*) area of steady slip. (Courtesy by Y. Yagi.)

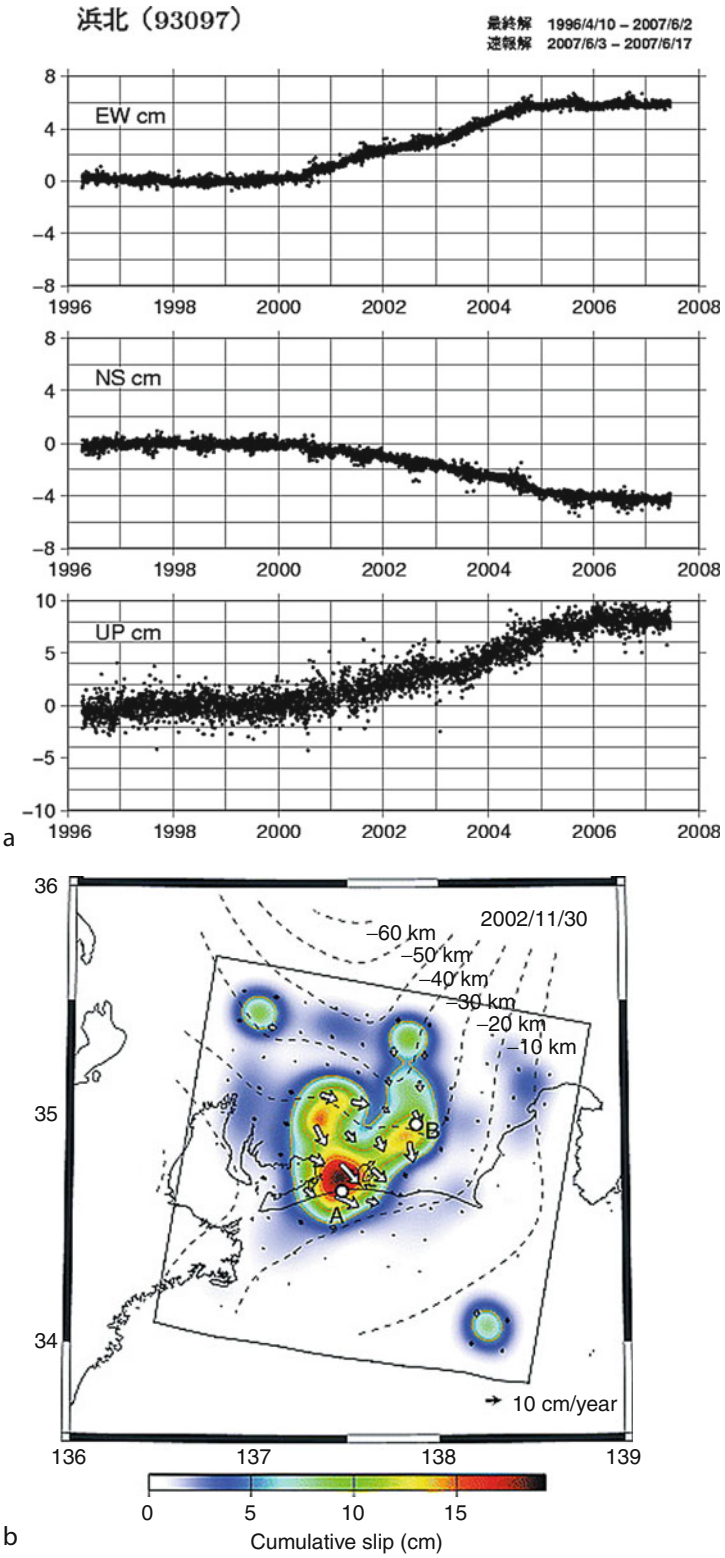
The first eminent case of large slow slip events found using GPS network data would be the post-seismic slip due to the 1994 Sanriku-haruka-oki earthquake (Heki et al., 1997). Heki et al. (1997) analyzed the GEONET data and found that the moment released by the post-seismic slow slips is bigger than that released at the time of earthquake (Figure 1). Then, Hirose et al. (1999) found a slow slip event in the Bungo Channel, western part of Japan, that occurred after the two 1996 Hyuganada earthquakes of Mw 6.6 on October 19 and Mw 6.7 on December 3 that occurred in the south of the region. This slip continued about 1 year in 1997 and was assumed to be due to a slip on the subducting Philippine Sea plate. Their geodetic inversion suggested that the slip occurred on about 60 km × 60 km surface with the maximum fault slip of about 18 cm, which is equivalent to Mw 6.6 earthquake (Figure 2).

Other examples of slow events found in the Japanese region include (1) east off the Boso Peninsula where slow slip events have repeated in 1996 and in 2002 and they were followed by earthquake swarms (Sagiya, 2004, Ozawa et al., 2003) and (2) a post-seismic transient slip after the 2003 Tokachi earthquake (Mw 8.0) (Miyazaki et al., 2004).

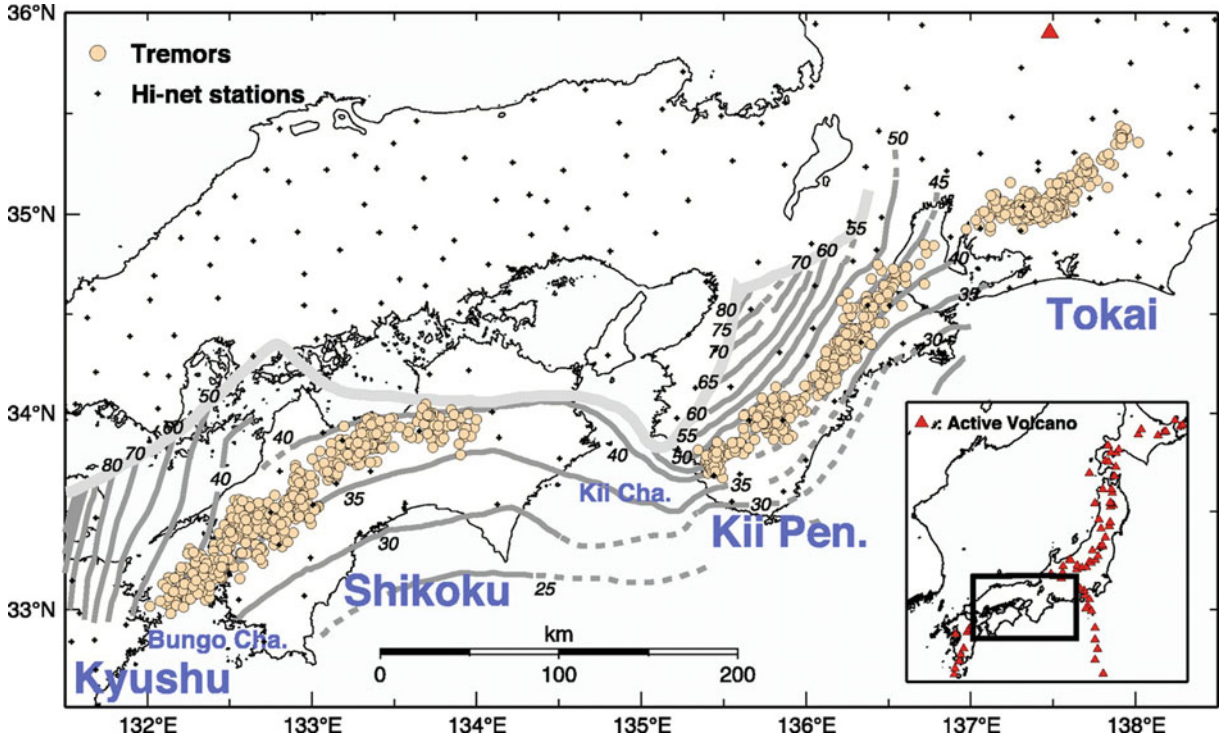
Lowry et al. (2001) found a transient movement at a continuous GPS site established at Cayaco, Guerero region, Mexico in early 1998. The area has been considered as a seismic gap. The event lasted about several months with 2 mm east, 26 mm south, and 16 mm up of displacements. They suggested that the displacement is consistent with a slip along the subduction interface that propagated from east to west. The transient displacement repeated in 2001–2002 for about 6 months. Kostoglodov et al. (2003) and Yoshioka et al. (2004) further investigated the mechanism of the event. Another repeated transient event in Mexico has been found in Oaxaca segment, hundreds of kilometers southeast of Guerrero segment, with more dense GPS array, and was studied by Brudzinski et al. (2007) and Correa-Mora et al. (2008).

Other long-term slow events have been found in Costa Rica (Protti et al., 2004), Alaska-Aleutian (Ohta et al., 2006), and New Zealand (Douglas et al., 2005). SSEs have also been found along the shallower part of the San Andreas fault (e.g., Wesson, 1987; Gladwin et al., 1994; Linde et al., 1996), and in Italy (Crescentini et al., 1999; Amoruso et al., 2002). Readers are asked to refer to Schwartz and Rokosky (2007) and Schwartz (2009) for a detailed review.

Yagi et al. (2001), and Yagi and Kikuchi (2003) used GPS data to examine areal distribution of co-seismic slip, post-seismic slip, and slow slip events for two Hyudanada earthquakes of 1996 (Mw 6.6 and Mw 6.7). They found that these areas complementarily share the areas and postulated that this complementarity may come from different constitutive parameters of slip at respective regions (Figure 2). A similar complementary region of seismic slip and aseismic slip is found also in the northeastern Japan (Yagi et al., 2003). Correa-Mora et al. (2008) indicated that the repeated transients at the Oaxaca segment occurred at the lower extension of the locked part of the



Slow Earthquake, Figure 3 (**a**) A sample GPS time series at the Hamakita site for the Tokai slow slip event. Permanent trend for the period from 1996 to 2000 was removed. Slow slip event was attributed in the period from late 2000 to early 2005. (From webpage of the Geographical Survey Institute) (**b**) Cumulative slip inferred from the extended Network Inversion Filter. *Arrows* show the direction and magnitude of slip of the upper plate (Miyazaki et al., 2006).



Slow Earthquake, Figure 4 Distribution of low-frequency tremor shown in dots (Obara, 2002).

seismogenic zone. These lines of evidence may suggest distinct physical rock properties along the different segments of the subducting plate interface.

The largest and the longest slow slip event, that has ever been found, lasted for about 5 years. It was recorded along the northern margin of the subducting Philippine Sea plate, namely, Tokai district (Figure 3). The event started in the middle of 2000 and continued until mid-2005. Both Ozawa et al. (2002) and Miyazaki et al. (2006) employed a state-space model to delineate temporal evolution of the slow slip event (Segall and Matthews, 1997). The slip amounted to be about several tens of centimeters, which is equivalent to moment magnitude of about 7.2–7.3 earthquake. Figure 3 shows a sample time series of coordinate change at a GEONET station that is located right above the slowly slipping region.

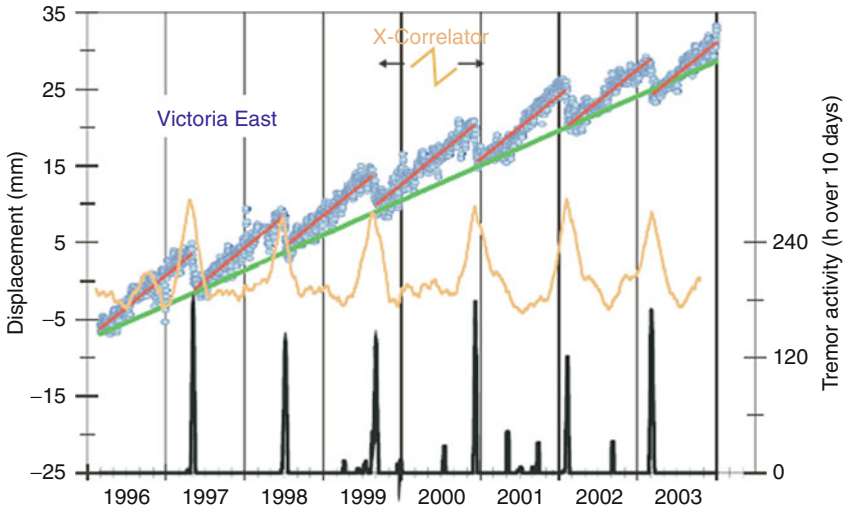
Low-frequency tremor and short-term slow slip

Obara (2002) discovered so-called deep low-frequency tremor using a nationwide seismic array in Japan, which is named as HiNet (High Sensitivity Seismograph Network) constituting of about 600 sites all over Japan. The deep low-frequency tremor is about 1 h of tremor activity, which is located nearly at the plate interface of 35–40 km in depth (Figure 4). The tremor has 0.2–2 Hz in predominant frequency. And then, the tremor was found to be accompanied by small silent slips of a few to several days (Obara and Hirose, 2006; Hirose and Obara, 2006).

Similar nonvolcanic tremor accompanied by short-term slow events were found also in the Northwestern Pacific (Dragart et al., 2001; Rogers and Dragart, 2003; Miller et al., 2002). Dragart et al. (2001) indicated that the SSE rather regularly repeats with interval of about 14 months and is associated with seismic swarm activities (Figure 5). The estimated slip is about 3 cm. Such slow events associated with swarm activities are called episodic tremor and slip (ETS) (Rogers and Dragart, 2003).

Hirose and Obara (2005) found that mid-term (about a month) and short-term (1 or 2 weeks) of slow slip events are accompanied by low-frequency tremors. The short-term silent slip was too small to be monitored by GPS but can be observed on highly sensitive tiltmeters, which are mostly embedded in HiNet boreholes. The estimated amount of slip was only 1–2 cm. They also found repeated similar short-term slow slip events associated with low-frequency tremors at the Tokai region (Hirose and Obara, 2006; Figure 6). Observed slow events were modeled by a slip at the depth of 25 km and the slip amounted only 1.2 cm.

Several lines of evidence from seismological studies suggest that the occurrence of low-frequency tremors and slow slip events may somehow be related to the existence of fluid flow. Shelly et al. (2006) and Obara and Hirose (2006) suggested that episodic tremors that include low-frequency earthquakes are generated by a slow shear slip on the plate interface. Fluids might play a key role for the generation of such low-frequency earthquakes.



Slow Earthquake, Figure 5 Frequently and regularly repeating slow slip events and low-frequency tremors along the Cascadia subduction zones (Dragart et al., 2004).

Matsubara et al. (2009) and Kato et al. (2010) also indicated that the high-pressure fluids released by dehydration from the subducting oceanic crust generates those tremors and slow slip events.

In addition to the low-frequency tremor that occurs at a deeper extension of the seismogenic zone, low-frequency earthquakes (LFE) of a characteristic period of 10–20 s have also been observed at very shallow areas at the updip extension of the seismogenic zones along the subducting plate. They are found along the Nankai trough (e.g., Ishihara, 2003; Obara and Ito, 2005) and at the junction between the Japan trench and the Kuril trench (Asano et al., 2008). Considering that the LFEs along the Nankai trough are occurring within the accretionary prism and have higher dip angle compared with the subucting plate interface, Ito and Obara (2006a) hypothesized that they may occur in the spray faults (or out-of-sequence thrusts) in the wedge.

Existence of fluid would have a key role for generating such low-frequency events, for both deep and shallow low-frequency earthquakes (e.g., Davis et al., 2006; Ito and Obara, 2006b).

Pre-seismic slip

A number of possible pre-seismic slips have been reported. Roeloffs (2006), among others, compared ten distinct examples of reported pre-seismic deformation rate changes before large earthquakes. For example, anomalous crustal deformations were observed before the 1944 Tonankai and 1946 Nankai earthquakes in Japan. Linde and Sacks (2002) showed that these deformations are consistent with the assumed aseismic slip of about 2 m in the down dip extension of the seismically slipped interface. Kanamori and Cipar (1974) and Linde and Silver (1989) postulated that a slow event had preceded the 1960 Chile earthquake.

It might be pointed out that the mechanism of precursory slow slip would be physically the same as slow slip event. Only the difference would be that while the precursory slip is an accelerating phase toward the rapid rupture, slow slip event is the slip that is not grown to a rapid rupture. Studies to clarify the mechanism that creates such difference are yet to be conducted in the future.

Scaling relation

An important question raised is if the discovered slow earthquakes or slow events are just the slow version of "regular" or "normal" earthquakes. If this is the case, those slow events may have the similar scaling relations as the other regular earthquakes. Ide et al. (2007), however, suggested that the slow earthquakes including silent earthquake, slow slip, etc., had different scaling relation and spectral behavior compared with normal earthquakes (Figure 7). They suggested that the characteristic duration (T sec) and the released moment (M_0 Nm) has linear relationship:

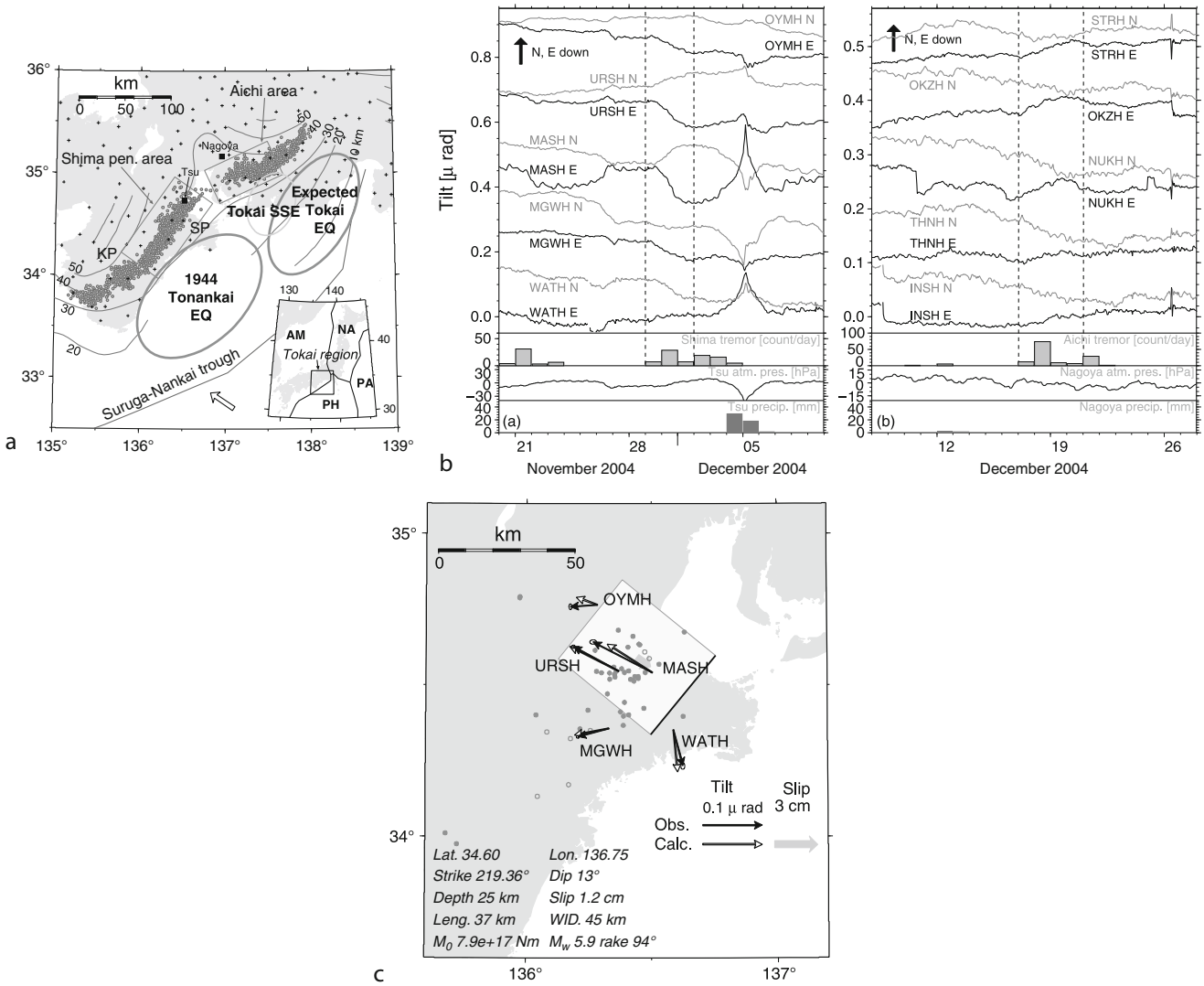
$$M_0 \approx T \times 10^{12-13} \quad (1)$$

While this relation for the regular earthquake is

$$M_0 \approx T^3 \times 10^{15-16} \quad (2)$$

Thus, the characteristics of moment rate spectrum are also different between these events.

This finding may tell us that the slow events are not just a slower extension of seismic slip but rather has a physically different nature. They suggest that such difference would be interpreted by "a constant low-stress drop model or a diffusional constant-slip model" (Ide



Slow Earthquake, Figure 6 (**a**) Map of studied area. (**b**) Observed tilt changes in November and December 2004, due to short-term slow slip events in the Tokai region. Periods of slow events are bounded by *dashed lines*. Recorded low-frequency tremors are also shown. (**c**) Modeled source region and estimated parameters. Observed and predicted tilt changes are shown in vectors (Hirose and Obara, 2006).

et al., 2007). Though the idea is yet to be tested, this might bring us a new insight for understanding the mechanism of fault slip or "earthquake" itself.

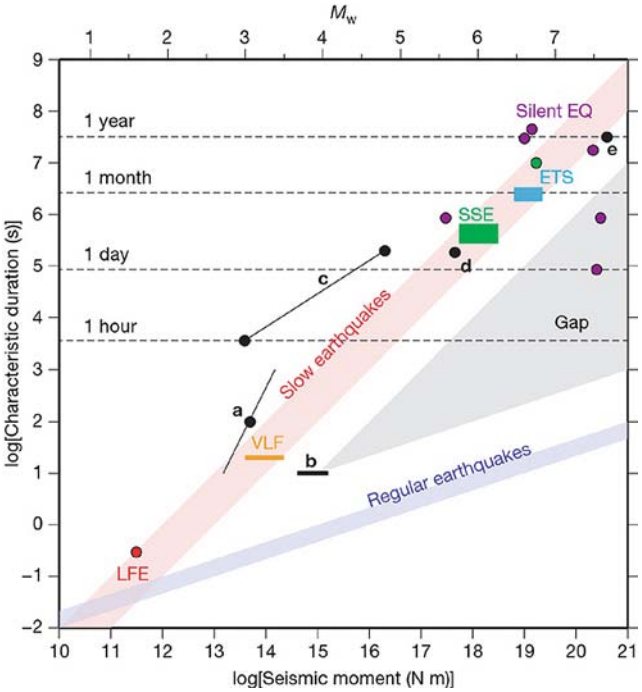
Mechanism of slow earthquake

The mechanism of fault slip has been extensively studied in the field of rock mechanics including theoretical treatments. It is inevitable to clarify the constitutive relations of fault ruptures to understand what mechanism controls the occurrence of slow earthquake (or earthquake itself). Numerous experimental and theoretical studies have been conducted for this purpose in decades. Among various proposed constitutive laws, the rate- and state-dependent friction law has been widely accepted (Dieterich, 1979; Ruina, 1983). The law is expressed by the following equations:

$$\mu\left(=\frac{\tau}{\sigma}\right)=\mu^{-}+a\ln\left(\frac{V}{V^{*}}+\theta\right) \tag{3}$$

$$\frac{\mathrm{d}\theta}{\mathrm{d}t}=-\left(\frac{V}{L}\right)\left[\theta+b\ln\left(\frac{V}{V^{*}}\right)\right] \tag{4}$$

Where, μ is friction coefficient, τ is shear stress, σ is normal stress, μ^{-} is friction coefficient at V^{*} where V^{*} is constant, L is characteristic slip distance, and θ is so-called state variable that represents the state along the fault



Slow Earthquake, Figure 7 Comparison between seismic moment and the characteristic duration of various slow earthquakes. LFE (*red*), VLF (*orange*), SSE (*green*), and ETS (*light blue*) are low-frequency earthquake, very low frequency earthquake, slow slip event, and episodic tremor and slip, respectively. See Ide et al. (2007) for more details.

surface. Coefficients a and b are constant. When the state θ does not change with time, it is called steady state and the friction coefficient in such steady state μ^{ss} is shown as

$$\mu^{ss} = \mu_0 + (a - b) \ln\left(\frac{V}{V^-}\right). \quad (5)$$

A simple spring-slider model is often used to model the fault slips (Ruina, 1983; Rice and Ruina, 1983). The system is characterized by the normal stress, shear stress, and the stiffness (or spring constant) k. Studies suggest that the type of slip, either rapid or slow slip, is controlled by several parameters. From Equation 5, it is suggested that, if $(a - b)$ is negative, the friction coefficient μ_{ss} becomes smaller when slip velocity V increases, which is called velocity weakening. In case of velocity weakening, system stiffness k controls the slip mode; if k is smaller than a critical value k_c which is given by $(b - a)\ \sigma/L$ (e.g., Dieterich, 1979), then the slip is unstable. Even in case of velocity weakening, the slip becomes slow if $(a - b)$ is close to zero. When $(a - b) > 0$, it is called velocity strengthening and the slip is decelerated.

Marone et al. (1991) used the case of velocity strengthening to interpret the post-seismic transient slow slip along the San Andreas fault and estimated the thickness of surface velocity strengthening layer to be 2–5 km. Yoshida and Kato (2003) used two tethered spring-sliders model to investigate fault interactions and examined what kind of conditions controls the occurrence of slow slip events. They showed that the repeated slow slip events may occur if the stiffness of the first block k is smaller than k_c and the second block is such that k is slightly bigger than k_c (which means that the condition is close to the boundary between stable and unstable slip). They also examined the case that an earthquake triggers the slow slip events in the adjacent region as was observed in the case of 2003 Tokachi-oki earthquake.

Further application of the rate- and state-dependent friction law to the numerical simulation of slip along the fault has been developed by a number of researchers. Tse and Rice (1986), for example, applied the relation to the two-dimensional San Andreas fault and suggested that unstable sliding may be limited to the depth shallower than 5–7 km and the steady slow slips prevails below 13–15 km of depth. Three-dimensional application was, for example, done by Rice (1993) and Stuart and Tullis (1995).

Concluding remarks

Although the slow earthquake has been studied for decades, its rapid progress was realized since the advent of the Global Positioning System observations in the middle of 1990s. Slow slip events would be one of the most important discoveries in seismology in the recent decades, thanks to GPS. In addition, the discoveries of other low-frequency events by high-sensitivity seismic arrays provided another important progress in understanding the mechanism of earthquake generation.

It may be reminded that the whole earthquake cycle was once categorized into four stages, namely, inter-seismic, pre-seismic, co-seismic, and post-seismic stages, and they had been studied rather independently. However, these stages are now recognized as cross-sectional views of single continuous series of progress of slips along the fault under the unified governing fault constitutive law.

Synthetic image of earthquake from high frequency that has been recorded by seismometers to very low frequency including transient slow events provided seismologists with opportunities of ultimate understanding of what "earthquake" is. In order, however, to realize this, a lot of more advanced researches are to be conducted and the results should be synthesized; for example, detailed deep sounding of structures of the lower crust and upper mantle where brittle and ductile ruptures are taking place should be conducted to clarify the field of earthquakes. Also, laboratory experiments for developing more conclusive governing laws of rapid and slow ruptures are indispensable. In addition, further findings of slow slip events and slow earthquakes at various fields in the world will provide deeper insights on the mechanism of earthquake generation. Such progresses might lead to a further development toward the most difficult seismological problem of earthquake prediction.

Acknowledgments

The author is indebted to Professor Naoyuki Kato and Professor Kazushige Obara of Earthquake Research Institute, the University of Tokyo, for their helpful discussions and providing information to the author. This study was supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan, under its Observation and Research Program for Prediction of Earthquakes and Volcanic Eruptions.

Bibliography

Amoruso, A., Crescentini, L., Morelli, A., and Scarpa, R., 2002. Slow rupture of an aseismic fault in a seismogenic region of Central Italy. *Geophys Res Lett*, **29**(24), 2219, doi:10.1029/2002GL016027.

Asano, Y., Obara, K., and Ito, Y., 2008. Spatiotemporal distribution of very-low frequency earthquakes in Tokachi-oki near the junction of the Kuril and Japan trenches revealed by using array signal processing. *Earth, Planets and Space*, **60**, 871–875.

Beroza, G. C., and Jordan, T., 1990. Searching for slow and silent earthquakes. *Journal of Geophysical Research*, **95**, 2485–2510.

Brudzinski, M., Cabral-Cano, E., Correa-Mora, F., DeMets, C., and Márquez-Azúa, B., 2007. Slow slip transients along the Oaxaca subduction segment from 1993 to 2007. *Geophysical Journal International*, **171**, 523–538.

Correa-Mora, F., DeMets, C., Cabral-Cano, E., and Márquez-Azúa, B., 2008. Interplate coupling and transient slip along the subduction interface beneath Oaxaca, Mexico. *Geophysical Journal International*, **175**, 269–290.

Crescentini, L., Amoruso, A., and Scarpa, R., 1999. Constraints on slow earthquake dynamics from a swarm in central Italy. *Science*, **286**, 2132–2134.

Davis, E. E., Becker, K., Wang, K., Obara, K., Ito, Y., and Kinoshita, M., 2006. A discrete episode of seismic and aseismic deformation of the Nankai trough subduction zone accretionary prism and incoming Philippine Sea plate. *Earth and Planetary Science Letters*, **242**, 73–84.

Dieterich, J. H., 1979. Modeling of rock friction, 1. Experimental results and constitutive equations. *Journal of Geophysical Research*, **84**, 2161–2168.

Douglas, A., Beavan, J., Wallace, L., and Townend, J., 2005. Slow slip on the northern Hikurangi subduction interface, New Zealand. *Geophysical Research Letters*, **32**, L16305, doi:10.1029/2005GL023607.

Dragart, H., Wang, K., and James, T. S., 2001. A silent slip event on the deeper Cascadia subduction interface. *Science*, **292**, 1525–1528.

Dragart, H., Wang, K., and Rogers, G., 2004. Geodetic and seismic signatures of episodic tremor and slip in the northern Cascadia subduction zone. *Earth Planets Space*, **56**, 1143–1150.

Fitch, T. J., and Scholz, C. H., 1971. Mechanism of underthrusting in southwest Japan: a model of convergent plate interactions. *Journal of Geophysical Research*, **76**, 7260–7292.

Gladwin, M. T., Gwyther, R. L., Hart, R. H. G., and Brechenridge, K. S., 1994. Measurements of the strain field associated with episodic creep events on the San Andreas fault at San Juan Bautista, California. *Journal of Geophysical Research*, **99**, 4559–4564.

Heki, K., Miyazaki, S., and Tsuji, H., 1997. Silent fault slip following an interpolate thrust earthquake at the Japan Trench. *Nature*, **386**, 595–598.

Hirose, H., and Obara, K., 2005. Repeating short- and long-term slow slip events with deep tremor activity around the Bungo Channel region, southwest Japan. *Earth, Planets and Space*, **57**, 961–972.

Hirose, H., and Obara, K., 2006. Short-term slow slip and correlated tremor episodes in the Tokai region, central Japan. *Geophysical Research Letters*, **33**, L17311, doi:10.1029/2006GL026579.

Hirose, H., Hirahara, K., Kimata, F., Fujii, N., and Miyazaki, S., 1999. A slow thrust slip event following the two 1996 Hyuganada earthquakes beneath the Bungo Channel, southwest Japan. *Geophysical Research Letters*, **26**(21), 3237–3240.

Ide, S., Beroza, G. C., Shelly, D. R., and Uchide, T., 2007. A scaling law for slow earthquakes. *Nature*, **447**, 76–79, doi:10.1038/nature05780.

Ishihara, Y., 2003. Major existence of very low frequency earthquakes in background seismicity along subduction zone of Southwestern Japan, *Eos Transactions of AGU*, **84** (46), Fall Meet. Suppl., Abstract S41C-0107.

Ito, Y., and Obara, K., 2006a. Dynamic deformation of the accretionary prism excites very low frequency earthquakes. *Geophysical Research Letters*, **33**, L02311, doi:10.1029/2005GL025270.

Ito, Y., and Obara, K., 2006b. Very low frequency earthquakes within accretionary prisms are very low stress-drop earthquakes. *Geophysical Research Letters*, **33**, L09302, doi:10.1029/2006GL025883.

Kanamori, H., 1972. Mechanism of tsunami earthquakes. *Physics of the Earth and Planetry Interiors*, **6**, 346–359.

Kanamori, H., and Cipar, J., 1974. Focal process of the great Chilean earthquake May 22, 1960. *Physics of the Earth and Planetary Interiors*, **9**, 127–136.

Kanamori, H., and Kikuchi, M., 1993. The 1992 Nicaragua earthquake: a slow tsunami earthquake associated with subducted sediments. *Nature*, **361**, 714–716.

Kanamori, H., and Stewart, G. S., 1979. A slow earthquake. *Physics of the Earth and Planetary Interiors*, **18**, 167–175.

Kawasaki, I., Asai, Y., Tamura, Y., Sagiya, T., Mikami, N., Okada, Y., Sakata, M., and Kasahara, M., 1995. The 1992 Sanriku-Oki, Japan, ultra-slow earthquake. *Journal of Physics of Earth*, **43**, 105–116.

Kato, A., Iidaka, T., Ikuta, R., Yoshida, Y., Katsumata, K., Iwasaki, T., Sakai, S., Thurber, C., Tsumura, N., Yamaoka, K., Watanabe, T., Kunitomo, T., Yamazaki, F., Okubo, M., Suzuki, S., and Hirata, N., 2010. Variations of fluid pressure within the subducting oceanic crust and slow earthquakes. *Geophys Res Lett*, **37**, L14310, doi:10.1029/2010GL043723.

Kawasaki, I., Asai, Y., and Tamura, Y., 2001. Space-time distribution of interplate moment release including slow earthquakes and the seismo-geodetic coupling in the Sanriku-oki region along the Japan trench. *Tectonophysics*, **330**, 267–283.

Kostoglodov, V., Singh, S. K., Santiago, J. A., Franco, S. I., Larson, K. M., Lowry, A. R., and Bilham, R., 2003. A large silent earthquake in the Guerrero seismic gap, Mexico. Geophysical Research Letters, **30**(15), 1807, doi: 10.1029/2003GL017219

Linde, A. T., and Silver, P. G., 1989. Elevation changes and the great 1960 Chilean earthquake: support for aseismic slip. *Geophysical Research Letters*, **16**, 1305–1308.

Linde, A. T., Gladwin, M. T., Johnston, M. J. S., Gwyther, R. L., and Bilham, R. G., 1996. A slow earthquake sequence on the San Andreas Fault. *Nature*, **383**, 65–68.

Linde, A. T., and Sacks, I. S., 2002. Slow earthquakes and great earthquakes along the Nankai trough. *Earth Planet Sci Lett*, **203**, 265–275.

Lowry, A. R., Larson, K. M., Kostoglodov, V., and Bilham, R., 2001. Transient fault slip in Guerrero, southern Mexico. *Geophysical Research Letters*, **28**, 3753–3756.

Marone, C. J., Scholz, C. H., and Bilham, R., 1991. On the mechanism of earthquake afterslip. *Journal of Geophysical Research*, **96**, 8441–8452.

Matsubara, M., Obara, K., and Kasahara, K., 2009. High-Vp/Vs zone accompanying non-volcanic tremors and slow-slip events beneath southwestern Japan. *Tectonophysics*, **472**, 6–17.

Melbourne, T. I., Webb, F. H., Stock, J. M., and Reigber, C., 2002. Rapid postseismic transients in subduction zones from continuous GPS. *Journal of Geophysical Research*, **107**(B10), 2241, doi:10.1029/2001JB000555.
Miller, M. M., Melbourne, T., Johnson, D. J., and Sumner, W. Q., 2002. Periodic slow earthquakes from the Cascadia subduction zone. *Science*, **295**, 2423.
Miyazaki, S., Segall, P., Fukuda, J., and Kato, T., 2004. Space time distribution of afterslip following the 2003 Tokachi-oki earthquake: implications for variations in fault zone frictional properties. *Geophysical Research Letters*, **31**, L06623, doi:10.1029/2003GL019410.
Miyazaki, S., Segall, P., McGuire, J. J., Kato, T., and Hatanaka, Y., 2006. Spatial and temporal evolution of stress and slip rate during the 2000 Tokai slow earthquake. *Journal of Geophysical Research*, **111**, B03409, doi:10.1029/2004JB003426.
Obara, K., 2002. Nonvolcanic deep tremor associated with subduction in southwest Japan. *Science*, **296**, 1679–1681.
Obara, K., and Hirose, H., 2006. Non-volcanic deep low-frequency tremors accompanying slow slips in the southwest Japan subduction zone. *Tectonophysics*, **417**, 33–51.
Obara, K., and Ito, Y., 2005. Very low frequency earthquakes excited by the 2004 off the Kii peninsula earthquakes: a dynamic deformation process in the large accretionary prism. *Earth, Planets and Space*, **57**, 321–326.
Ohta, Y., Freymueller, J. T., Hreinsdottir, Sigrun, and Suito, H., 2006. A large slow slip event and the depth of the seismogenic zone in the south central Alaska subduction zone. *Earth and Planetary Science Letters*, **247**, 108–116.
Okada, A., and Nagata, T., 1953. Land deformation of the neighborhood of Muroto point after the Nankaido great earthquake in 1946. *Bulletin of Earthquake Research Institute*, **31**, 169–177.
Ozawa, S., Murakami, M., Kaidzu, M., Tada, T., Hatanaka, Y., Yarai, H., and Nishimura, T., 2002. Detection and monitoring of ongoing aseismic slip in the Tokai region, central Japan. *Science*, **298**, 1009–1012.
Ozawa, S., Miyazaki, S., Hatanaka, Y., Imakiire, T., Kaidzu, M., and Murakami, M., 2003. Characteristic silent earthquakes in the eastern part of the Boso peninsula, Central Japan. *Geophys Res Lett*, **30**(6), 1283, doi:10.1029/2002GL016665.
Protti, M., Gonzales, V., Kato, T., Iinuma, T., Miyazaki, S., Obana, K., Kaneda, Y., LaFemina, P., Dixon, T., and Schwartz, S., 2004. A creep event on the shallow interface of the Nicoya Peninsula, Costa Ricaseismogenic zone. Eos Transactions of AGU, **85**(47), Fall Meet. Suppl., Abstract S441D-07.
Rice, J. R., 1993. Spatio-temporal complexity of slip on a fault. *Journal of Geophysical Research*, **98**, 9885–9907.
Rice, J. R., and Ruina, A. L., 1983. Stability of steady frictional slipping. *Journal of Applied Mechanics*, **50**, 343–349.
Roeloffs, E. A., 2006. Evidence for aseismic deformation rate changes prior to earthquakes. *Annual Review of Earth and Planetary Sciences*, **34**, 591–627.
Rogers, G., and Dragart, H., 2003. Episodic tremor and slip on the Cascadia subduction zone: the chatter of silent slip. *Science*, **300**, 1942–1943.
Ruina, A. L., 1983. Slip instability and state variable friction laws. *Journal of Geophysical Research*, **88**, 10359–10370.
Sagiya, T., 2004. Interplate coupling in the Kanto district, Central Japan, and the Boso Peninsula silent earthquake in May 1996. *Pure and Applied Geophysics*, **161**, 2327–2342.
Schwartz, S. Y., 2009. Episodic aseismic slip at plate boundaries. In Kanamori, H. (ed.), Schubert G. (ed.-in-chief), *Earthquake Seismology*. Oxford: Elsevier, Vol. 4, pp. 445–472.
Schwartz, S. Y., and Rokosky, J. M., 2007. Slow slip events and seismic tremor at circum-Pacific subduction zones. *Review of Geophysics*, **45**, RG3004.
Segall, P., and Matthews, M., 1997. Time dependent inversion of geodetic data. *Journal of Geophysical Research*, **102**(B10), 22391–22409.
Shelly, D. R., Beroza, G. C., Ide, S., and Nakamura, S., 2006. Low-frequency earthquakes in Shikoku, Japan, and their relationship to episodic tremor and slip. *Nature*, **442**, 188–191.
Stuart, W. D., and Tullis, T. E., 1995. Fault model for preseismic deformation at Parkfield, California. *Journal of Geophysical Research*, **100**, 24079–24099.
Tse, S. T., and Rice, J. R., 1986. Crustal earthquake instability in relation to the depth variation of frictional slip properties. *Journal of Geophysical Research*, **91**, 9452–9472.
Wesson, R. L., 1987. Modeling aftershock migration and afterslip of the San Juan Bautista, California, earthquake of October 3, 1972. *Tectonophysics*, **144**, 215–229.
Yagi, Y., and Kikuchi, M., 2003. Partitioning between co-seismic slip and asesmic slip. *Journal of Geography*, **112**(6), 828–836 (in Japanese with English abstract).
Yagi, Y., Kikuchi, M., and Sagiya, T., 2001. Co-seismic slip, post-seismic slip, and aftershocks associated with two large earthquakes in 1996 in Hyuga-nada, Japan. *Earth, Planets and Space*, **53**, 793–803.
Yagi, Y., Kikuchi, M., and Nishimura, T., 2003. Co-sesmic slip, post-seismic slip, and largest aftershock associated with the 1994 Sanriku-haruka-oki, Japan, earthquake. *Geophysical Research Letters*, **30**(22), 2177, doi:10.1029/2003GL018189.
Yoshida, S., and Kato, N., 2003. Episodic aseismic slip in a two-degree-of-freedom block-spring model. *Geophysical Research Letters*, **30**(13), 1681, doi:10.1029GL017439.
Yoshioka, S., Mikumo, T., Kostoglodov, V., Larson, K. M., Lowry, A. R., and Singh, S. K., 2004. Interplate coupling and a recent aseismic slow slip event in the Guerrero seismic gap of the Mexican subduction zone, as deduced from GPS data inversion using a Baysian information criterion. *Physics of the Earth and Planetary Interiors*, **146**, 513–530.

Cross-references

Earthquakes, Source Theory
GPS, Data Acquisition and Analysis
GPS, Tectonic Geodesy
Seismic Quiescence and Activation

SPHERICAL HARMONIC ANALYSIS APPLIED TO POTENTIAL FIELDS*

Nikolaos K. Pavlis
Geodesy and Geophysics Basic and Applied Research, National Geospatial-Intelligence Agency (NGA), Reston, VA, USA

Synonyms

Fourier expansion on the sphere; Spherical harmonic decomposition; Spherical harmonic expansion; Spherical harmonic transform

Definition

A square integrable function, defined on a surface that has a one-to-one correspondence with the unit sphere, may be

represented as a linear combination of *Surface Spherical Harmonic functions*. *Spherical Harmonic Analysis* (SHA) is the process by which the coefficients defining this linear combination are determined. These coefficients constitute the *Surface Spherical Harmonic spectrum* of the function. Functions that satisfy Laplace's partial differential equation are called *harmonic*. These can be decomposed into series of *Solid Spherical Harmonic functions*. Although SHA is *not* restricted to functions that represent potential fields, it plays a special role in the determination of such fields through the solution of *Boundary Value Problems* (BVP).

Introduction: basic formulas

A scalar function V, representing a potential field such as the Earth's gravitational or magnetic field, satisfies Laplace's partial differential equation (PDE), in the space outside of the sources generating the field. In geocentric spherical polar coordinates (r, θ, λ), where r is the radial distance, θ is the geocentric colatitude (defined as 90° minus the geocentric latitude), and λ is the longitude, Laplace's PDE takes the form (Heiskanen and Moritz, 1967, Equation 1–41):

$$\Delta V \equiv \nabla^2 V = \frac{\partial^2 V}{\partial r^2} + \frac{2}{r}\frac{\partial V}{\partial r} + \frac{1}{r^2}\frac{\partial^2 V}{\partial \theta^2} + \frac{\cot\theta}{r^2}\frac{\partial V}{\partial \theta} + \frac{1}{r^2 \sin^2\theta}\frac{\partial^2 V}{\partial \lambda^2} = 0. \tag{1}$$

Equation 1 may be solved using the method of separation of variables. For the space outside a sphere of radius $r = a$, its solution may be written in the form (see Heiskanen and Moritz, 1967 for a complete derivation):

$$V(r,\theta,\lambda) = \sum_{n=0}^{\infty}\left(\frac{a}{r}\right)^{n+1}\sum_{m=-n}^{n} v_{nm} Y_{nm}(\theta,\lambda). \tag{2}$$

The functions $\left[(a/r)^{n+1} Y_{nm}(\theta,\lambda)\right]$ are called *Solid Spherical Harmonic functions*. The *Surface Spherical Harmonic functions* $Y_{nm}(\theta,\lambda)$ are defined as:

$$Y_{nm}(\theta,\lambda) = \overline{P}_{n|m|}(\cos\theta)\cdot\left\{\begin{array}{l}\cos m\lambda\\ \sin|m|\lambda\end{array}\right\}\begin{array}{l}\text{if } m \geq 0\\ \text{if } m < 0\end{array}. \tag{3}$$

$\overline{P}_{n|m|}(\cos\theta)$ is the fully normalized *Associated Legendre function* of the first kind, of degree n and order $|m|$ (Heiskanen and Moritz, 1967, Sects. 1–11 and 1–14). *Fully normalized* Associated Legendre functions are commonly used in geodesy. These are related to their unnormalized counterparts $P_{n|m|}(\cos\theta)$ by:

$$\overline{P}_{n|m|}(\cos\theta) = \sqrt{(2-\delta_{|m|0})(2n+1)\frac{(n-|m|)!}{(n+|m|)!}}\cdot P_{n|m|}(\cos\theta), \tag{4}$$

where:

$$\delta_{|m|0} = \left\{\begin{array}{ll}1 & \text{if } m = 0\\ 0 & \text{if } m \neq 0\end{array}\right. . \tag{5}$$

For the surface spherical harmonic functions Y_{nm}, this normalization implies that:

$$\frac{1}{4\pi}\iint_{\sigma} Y_{nm}^2(\theta,\lambda)\,d\sigma = 1, \tag{6}$$

with the integration being performed over the unit sphere σ, whose area element is $d\sigma$ ($d\sigma = \sin\theta d\theta d\lambda$). Note that in geomagnetism, the *Schmidt seminormalized* Associated Legendre functions (Blakely, 1995, p. 113) $\breve{P}_{n|m|}(\cos\theta)$ are used instead. These are defined by:

$$\breve{P}_{n|m|}(\cos\theta) = \sqrt{(2-\delta_{|m|0})\frac{(n-|m|)!}{(n+|m|)!}}\cdot P_{n|m|}(\cos\theta). \tag{7}$$

Surface spherical harmonics constitute a set of *orthogonal* basis functions on the unit sphere, i.e.:

$$\iint_{\sigma} Y_{nm}(\theta,\lambda) Y_{sr}(\theta,\lambda)\,d\sigma = 0 \quad \text{if } n \neq s \text{ or } m \neq r \text{ or both.} \tag{8}$$

This property of *orthogonality* permits the determination of the spherical harmonic coefficients v_{nm}, which appear in Equation 2, as follows. Assume for the sake of this discussion that the function $V(r,\theta,\lambda)$ is observable everywhere on the surface of the sphere $r = a$. Equation 2 then takes the form:

$$V(a,\theta,\lambda) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} v_{nm} Y_{nm}(\theta,\lambda). \tag{9}$$

In this case, the orthogonality of surface spherical harmonics implies that:

$$v_{nm} = \frac{1}{4\pi}\iint_{\sigma} V(a,\theta,\lambda) Y_{nm}(\theta,\lambda)\,d\sigma. \tag{10}$$

The coefficients v_{nm} constitute the *surface spherical harmonic spectrum* of the function $V(r,\theta,\lambda)$, pertinent to the surface of the sphere of radius $r = a$. It should be emphasized that *spherical harmonic expansions*, as the one given in Equation 9, are *not* restricted to functions satisfying Laplace's equation (*harmonic functions*). Any square integrable function $f(\theta,\lambda)$, defined over a surface S that has a one-to-one correspondence with the unit sphere, may be expanded in surface spherical harmonics as:

$$f^S(\theta,\lambda) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} f_{nm}^S Y_{nm}(\theta,\lambda), \tag{11}$$

with the coefficients f_{nm}^S given by:

$$f_{nm}^S = \frac{1}{4\pi} \iint_\sigma f^S(\theta,\lambda) Y_{nm}(\theta,\lambda) d\sigma. \tag{12}$$

The superscript "S" emphasizes the dependence of the spectrum f_{nm}^S of the function $f(\theta,\lambda)$, on the particular surface S over which its values $f^S(\theta,\lambda)$ are given. Jekeli (1988) carefully distinguishes between the direct Legendre transform (i.e., the spectrum) of an arbitrary square integrable function $f(\theta,\lambda)$, defined on a surface S that has a one-to-one correspondence with the unit sphere, and the special case of the solution of Laplace's equation when the boundary data reside on the surface of the sphere $r = a$. The former is expressed in Equations 11 and 12 above; the latter in Equations 9 and 10. Obviously the two quantities take the same form with an appropriate choice of coordinates. Furthermore, note that while rather arbitrary functions can be expanded in surface spherical harmonics, *only* harmonic functions (i.e., functions satisfying Laplace's equation) can be expanded into solid spherical harmonics, within their region of harmonicity.

The total power of the function $f(\theta,\lambda)$, defined over the surface S, is defined to be:

$$M\left\{\left[f^S\right]^2\right\} = \frac{1}{4\pi} \iint_\sigma \left[f^S(\theta,\lambda)\right]^2 d\sigma. \tag{13}$$

A generalization of Parseval's theorem permits the computation of $M\left\{\left[f^S\right]^2\right\}$ from the spectrum f_{nm}^S, by:

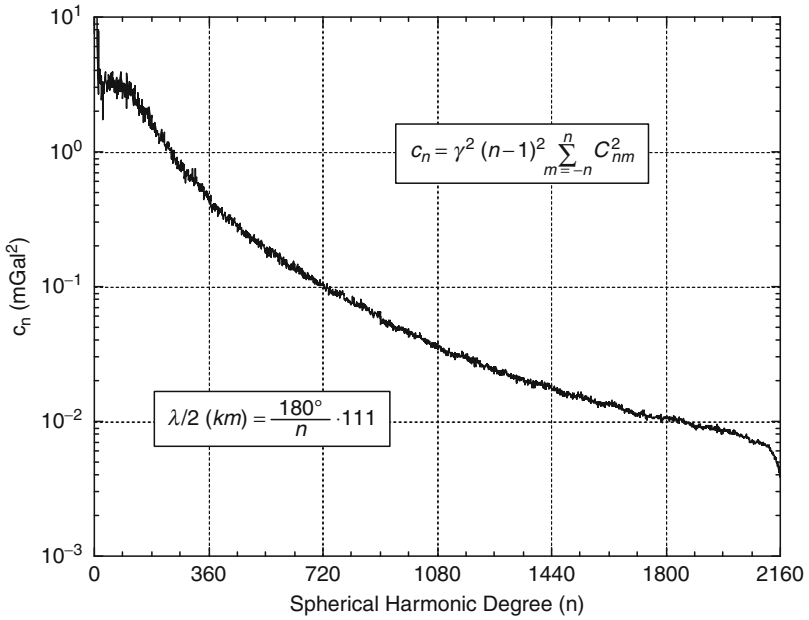
$$M\left\{\left[f^S\right]^2\right\} = \sum_{n=0}^{\infty} f_n^S = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \left[f_{nm}^S\right]^2. \tag{14}$$

The quantities f_n^S, defined as:

$$f_n^S = \sum_{m=-n}^{n} \left[f_{nm}^S\right]^2, \tag{15}$$

represent the total power (or variance) per spherical harmonic degree. In geodetic literature, they are known as *degree variances*. A direct correspondence exists between spherical harmonic degree and spatial wavelength. Therefore, the degree variances enable the study of the distribution of power within the function $f^S(\theta,\lambda)$, in terms of spatial wavelength. For example, Figure 1 (Pavlis et al., 2008) shows the degree variances implied by the Earth Gravitational Model 2008 (EGM2008) for the *free-air gravity anomaly* signal.

Spherical harmonic expansions of functions, such as the Earth's topography (heights above and depths below the Mean Sea Level), surface temperature, atmospheric pressure, etc., are very useful since they permit the study of the spectral properties of these functions. Furthermore, the decomposition of these functions into their spectral components permits the efficient application of *convolutions* on the sphere (Driscoll and Healy, 1994), and of



Spherical Harmonic Analysis Applied to Potential Fields, Figure 1 Degree variances (c_n) of the free-air gravity anomaly signal (in mGal^2) implied by the EGM2008 global gravitational model, as a function of spherical harmonic degree n. The degree variances refer to the surface of a sphere of radius $a = 6,378,136.3$ m. The relationship between half-wavelength $(\lambda/2)$ resolution and spherical harmonic degree is provided in the *insert*. In SI units, $1\ \mathrm{mGal} = 10^{-5}\ \mathrm{m\,s^{-2}}$.

band-pass filtering of data, with strict control on the filter's spherical harmonic-degree bandwidth. For functions representing potential fields, in addition to these spectral analysis applications, spherical harmonic analysis offers one approach for the solution of the underlying *Boundary Value Problem* that is associated with the specific *functional* of the field that has been observed. We discuss these aspects next.

Boundary Value Problems (BVP) of potential theory

In general, a BVP consists of a differential equation that is subject to a set of boundary conditions. In potential theory, the unknown (scalar) function to be determined is the potential, V, which satisfies Laplace's PDE, in the space outside of its generating sources. The function V should vanish at infinity as the reciprocal of the distance between the point in question and the generating source element. The boundary conditions, in order to be of practical use, should represent quantities related to V (*field functionals*) that are observable over some surface S. Two general approaches exist for the solution of a BVP:

(a) Solution using spectral analysis, i.e., spherical harmonics
(b) Solution using *Greens functions*, i.e., integral formulas

Traditionally, in gravimetric geodesy, the following three BVP of potential theory have been considered:

1. First BVP of potential theory, or *Dirichlet*'s problem (Heiskanen and Moritz, 1967, Sects. 1–16): given an arbitrary function f on a surface S, determine a function V that is harmonic either inside or outside S and which assumes on S the values of the prescribed function f. If the surface S is a sphere of radius $r = a$, the solution in terms of spherical harmonics, for the region outside S, is given by Equation 2. The corresponding Greens function solution is given by *Poisson*'s integral formula (Heiskanen and Moritz, 1967, Equation 1–89).
2. Second BVP of potential theory, or *Neumann*'s problem (Heiskanen and Moritz, 1967, Sects. 1–17): here, instead of the values of the potential V itself, one is given on a surface S the values of its normal derivative $\partial V/\partial n$. The normal derivative is the derivative along the outward-directed surface normal n to S. If the surface S is a sphere of radius $r = a$, the solution in terms of spherical harmonics, for the region outside S, is given by:

$$V(r,\theta,\lambda) = -a\sum_{n=0}^{\infty}\frac{1}{n+1}\left(\frac{a}{r}\right)^{n+1}\sum_{m=-n}^{n}u_{nm}Y_{nm}(\theta,\lambda), \tag{16}$$

where u_{nm} is the spectrum of the $\partial V/\partial n$ boundary values, given on the sphere $r = a$, i.e.:

$$\left(\frac{\partial V}{\partial n}\right)_{r=a} = \left(\frac{\partial V}{\partial r}\right)_{r=a} = \sum_{n=0}^{\infty}\sum_{m=-n}^{n}u_{nm}Y_{nm}(\theta,\lambda). \tag{17}$$

The values of u_{nm} can be obtained from:

$$u_{nm} = \frac{1}{4\pi}\iint_{\sigma}\left(\frac{\partial V}{\partial r}\right)_{r=a}Y_{nm}(\theta,\lambda)d\sigma. \tag{18}$$

The corresponding Greens function solution to the second BVP is given by *Hotine*'s integral formula (Hotine, 1969).

3. Third BVP of potential theory (Heiskanen and Moritz, 1967, Sects. 1–17): here, a linear combination of V and of its normal derivative $\partial V/\partial n$ is given on the surface S, i.e., $hV + k(\partial V/\partial n)$ is given on S. Again, if the surface S is a sphere of radius $r = a$, the solution in terms of spherical harmonics, for the region outside S, is given by:

$$V(r,\theta,\lambda) = \sum_{n=0}^{\infty}\frac{1}{h-(k/a)(n+1)}\left(\frac{a}{r}\right)^{n+1}\sum_{m=-n}^{n}w_{nm}Y_{nm}(\theta,\lambda), \tag{19}$$

where w_{nm} is the spectrum of the $hV + k(\partial V/\partial n)$ boundary values, given on the sphere $r = a$, i.e.:

$$\left(hV + k\frac{\partial V}{\partial n}\right)_{r=a} = \left(hV + k\frac{\partial V}{\partial r}\right)_{r=a} = \sum_{n=0}^{\infty}\sum_{m=-n}^{n}w_{nm}Y_{nm}(\theta,\lambda). \tag{20}$$

The values of w_{nm} can be obtained from:

$$w_{nm} = \frac{1}{4\pi}\iint_{\sigma}\left(hV + k\frac{\partial V}{\partial r}\right)_{r=a}Y_{nm}(\theta,\lambda)d\sigma. \tag{21}$$

This particular BVP is of great importance to physical geodesy. The determination of the *geoid* given *gravity anomalies* as boundary data – called the BVP of physical geodesy – represents a specific case of this third BVP, where $h = -2/a$ and $k = -1$. With these values for h and k, if the surface S is a sphere of radius $r = a$, the solution in terms of spherical harmonics, for the region outside S, is given by:

$$V(r,\theta,\lambda) = a\sum_{n=0}^{\infty}\frac{1}{n-1}\left(\frac{a}{r}\right)^{n+1}\sum_{m=-n}^{n}g_{nm}Y_{nm}(\theta,\lambda), \tag{22}$$

where g_{nm} is the spectrum generated from the SHA of the gravity anomaly boundary values that are given on the sphere $r = a$. The corresponding Greens

function solution to the BVP of physical geodesy is given by *Stokes'* integral formula (Heiskanen and Moritz, 1967, Sects. 2–16).

Other BVP may also be formulated and solved, which correspond to other observable boundary data. For example, one may consider the *tensor* of second-order gradients of the gravitational potential (Moritz, 1980, p. 4) (or some subset of the elements of this tensor) as boundary data, which give rise to BVP corresponding to *gravity gradiometry*.

Equations 9, 17, and 20 are all of the same form as Equation 11. Therefore, provided that the boundary data are given on, or can be reduced to, the surface of a sphere of radius $r = a$, surface SHA may be used to determine the surface spherical harmonic spectrum of the field. This surface spectrum, multiplied by the appropriate radial terms that appear in Equations 2, 16, and 19, provides then a solution in terms of solid spherical harmonics, to the corresponding BVP. The convergence of the series 2, 16, and 19 is guaranteed in the space outside of the sphere $r = a$.

Laplace's PDE can also be expressed in terms of the *ellipsoidal coordinates* (u, δ, λ), where u is the semi-minor axis of the confocal ellipsoid, δ is the *reduced* colatitude (defined as 90° minus the reduced latitude), and λ is the longitude (Heiskanen and Moritz, 1967, Sects. 1–19). The Cartesian coordinates (x, y, z) of a point are related to its spherical (r, θ, λ) and ellipsoidal (u, δ, λ) coordinates as follows:

$$\begin{aligned} x &= r\sin\theta\cos\lambda = \left(u^2 + E^2\right)^{1/2}\sin\delta\cos\lambda \\ y &= r\sin\theta\sin\lambda = \left(u^2 + E^2\right)^{1/2}\sin\delta\sin\lambda, \\ z &= r\cos\theta = u\cos\delta \end{aligned} \tag{23}$$

where E is the constant linear eccentricity of the family of confocal ellipsoids defining the coordinate system. Expressed in the (u, δ, λ) coordinate system, Laplace's PDE can also be solved using the method of separation of variables. The solution, for the space exterior to a reference ellipsoid with semi-minor axis $u = b$, is given by (Heiskanen and Moritz, 1967, Sects. 1–20):

$$V(u,\delta,\lambda) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n}\frac{Q_{n|m|}(i(u/E))}{Q_{n|m|}(i(b/E))}z_{nm}Y_{nm}(\delta,\lambda). \tag{24}$$

$Q_{n|m|}(i(u/E))$ is the *Associated Legendre function* of the second kind, of degree n and order $|m|$ (Heiskanen and Moritz, 1967, Sects. 1–12). i is the imaginary unit $(i = \sqrt{-1})$. On the surface of the reference ellipsoid $u = b$, Equation 24 becomes:

$$V(b,\delta,\lambda) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} z_{nm}Y_{nm}(\delta,\lambda). \tag{25}$$

Equation 25 is of the same general form as Equations 9 and 11, with the important distinction that δ now represents the *reduced* colatitude. The surface *ellipsoidal* harmonic spectrum z_{nm} can be determined from:

$$z_{nm} = \frac{1}{4\pi}\iint_{\sigma} V(b,\delta,\lambda)Y_{nm}(\delta,\lambda)d\sigma, \tag{26}$$

in exactly the same fashion as in the case of the surface *spherical* harmonic spectrum in 10 and 12. Note that the integration is again over the unit sphere, whose area element is now defined as $d\sigma = \sin\delta d\delta d\lambda$.

Spherical harmonics are used extensively in geodesy because of their simplicity, and because the Earth, to a first-order approximation, is a sphere. Since, to a second-order approximation, the Earth resembles more an ellipsoid of revolution, ellipsoidal harmonics are more suitable for the solution of BVP, when the boundary data reside on or near the surface of the Earth. The transformation between ellipsoidal and spherical harmonics can be performed using the exact analytical relationships formulated by Jekeli (1988) and implemented by Gleason (1988). In this fashion, the use of the rather complicated solid ellipsoidal harmonics of Equation 24 may be avoided altogether.

Spherical Harmonic Analysis (SHA): numerical techniques

Let us consider Equations 11 and 12. Equation 11 is representative of Equations 9, 17, 20, and 25, while Equation 12 is representative of Equations 10, 18, 21, and 26. Each one of Equations 11 and 12 suggests an alternative approach for the estimation of the coefficients f^S_{nm}, as we discuss next.

Least Squares Adjustment. The first approach uses the linear mathematical model of Equation 11 to set up a system of *observation equations*. With f^S_{obs} representing the observable quantity and $\hat{f}^S_{nm}$ the estimates of the unknown parameters, the observation equations can be written as:

$$\upsilon(\theta,\lambda) = \sum_{n=0}^{N}\sum_{m=-n}^{n}\hat{f}^S_{nm}Y_{nm}(\theta,\lambda) - f^S_{obs}(\theta,\lambda), \tag{27}$$

where $\upsilon(\theta,\lambda)$ is the residual associated with the observation $f^S_{obs}(\theta,\lambda)$. Notice that in Equation 27 the summation over spherical harmonic degree was truncated to a finite maximum degree N. In vector-matrix notation, Equation 27 takes the form:

$$\mathbf{v} = \mathbf{A}\hat{\mathbf{x}} - \mathbf{l}. \tag{28}$$

$\mathbf{v}$ is the vector representing the residuals $\upsilon(\theta,\lambda)$, $\mathbf{A}$ is the *design matrix* containing the terms that multiply the unknown parameters $\hat{f}^S_{nm}$, $\hat{\mathbf{x}}$ is the vector representing the unknown parameters $\hat{f}^S_{nm}$, and $\mathbf{l}$ is the vector containing the observations $f^S_{obs}(\theta,\lambda)$. In Equation 28, only the matrix

$\mathbf{A}$ and the vector $\mathbf{l}$ are known quantities. The linear system (28) can be solved using the method of *Least Squares Adjustment*. If $\boldsymbol{\Sigma}$ denotes the variance–covariance matrix of the errors associated with the observations $f^S_{obs}(\theta,\lambda)$, then, minimization of the quadratic form:

$$\phi = \mathbf{v}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{v}, \tag{29}$$

subject to the condition (28), leads to the *normal equations*:

$$\mathbf{N}\hat{\mathbf{x}} = \mathbf{U}, \text{where} \quad \mathbf{N} = \mathbf{A}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{A} \quad \text{and} \quad \mathbf{U} = \mathbf{A}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{l}. \tag{30}$$

A scaled version of the inverse of the variance–covariance matrix is the weight matrix $\mathbf{P}$. $\mathbf{P} = \sigma_0^2\boldsymbol{\Sigma}^{-1}$ and σ_0^2 is called the a priori variance of unit weight. σ_0^2 represents the error variance of an observation whose weight is equal to 1. The solution of the system of normal equations is:

$$\hat{\mathbf{x}} = \mathbf{N}^{-1}\mathbf{U}. \tag{31}$$

The variance-covariance matrix of the errors associated with the estimated values $\hat{\mathbf{x}}$ is given by:

$$\boldsymbol{\Sigma}_{\hat{\mathbf{x}}} = \mathbf{N}^{-1} = \left(\mathbf{A}^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{A}\right)^{-1}. \tag{32}$$

Carl Friedrich Gauss (1777–1855), who is credited with the development of the Least Squares Adjustment technique, used this approach in the analysis of geomagnetic data. Herein, we abbreviate the Least Squares Adjustment technique by LS.

Although in the previous discussion we presented the application of the LS method to a mathematical model that describes a *surface* spherical harmonic expansion, the LS method could be applied equally well to mathematical models describing *solid* spherical harmonic expansions, like those of Equations 2, 16, and 19. This means that one may use the LS method to estimate the spectra of potential fields, even when the boundary data are arbitrarily located in the three-dimensional space. In contrast, the *Numerical Quadrature* technique that we discuss next, when used for the determination of potential fields, requires that the boundary data are located over surfaces that are coordinate surfaces in the coordinate system in which Laplace's equation is formulated and solved ($r = const.$ for spherical harmonics, $u = const.$ for ellipsoidal harmonics). In theory, these coordinate surfaces must also encompass all the sources generating the field.

Numerical Quadrature. In the above discussion of the LS technique, the distribution, both in terms of location and in terms of number, of the measured values $f^S_{obs}(\theta,\lambda)$ remained unspecified. Consider now an equiangular ($\Delta\theta = \Delta\lambda$) partition of the unit sphere, along meridians ($\lambda = const.$) and parallels ($\theta = const.$). Such a partition creates a grid of $L \times 2L$ cells on the unit sphere, where $L = \pi/\Delta\theta$. Let us assume that the function $f(\theta,\lambda)$ was sampled in such a way that, for each one of the $L \times 2L$ cells, a value $f^S_{obs}(\theta,\lambda)$ exists that corresponds, for example, to the center of that cell. Such a sampling suggests a possible *discretization* of the surface integral of Equation 12, leading to the following *Numerical Quadrature* (NQ) formula (Colombo, 1981, Equation 1.5):

$$\tilde{f}^S_{nm} = \frac{1}{4\pi}\sum_{i=0}^{L-1}\sum_{j=0}^{2L-1} f^S_{obs}(\theta_i,\lambda_j)\,Y_{nm}(\theta_i,\lambda_j)\,\Delta\sigma_i, \tag{33}$$

where the area element $\Delta\sigma_i$ of the cells residing on the ith "row" is given by:

$$\Delta\sigma_i = \Delta\lambda\int_{\theta_i}^{\theta_{i+1}} \sin\theta d\theta = \Delta\lambda\cdot(\cos\theta_i - \cos\theta_{i+1}). \tag{34}$$

Equation 33 is a simple NQ formula, applicable to data $f^S_{obs}(\theta,\lambda)$ that represent point values.

In general, the estimate of the spectrum obtained either by using LS or by using NQ, will be different from the "true" spectrum f_{nm} of the function $f(\theta,\lambda)$. Furthermore, even when the two techniques are applied to the same data $f^S_{obs}(\theta,\lambda)$, their estimates will be, in general, different, i.e.:

$$\hat{f}^S_{nm} \neq \tilde{f}^S_{nm}. \tag{35}$$

It is therefore appropriate to consider the specific errors that affect $\hat{f}^S_{nm}$ and $\tilde{f}^S_{nm}$, compare the LS and NQ techniques, and determine if it is possible to design and formulate any other SHA technique(s) that would satisfy some desirable optimality criterion. This study was essentially carried out by Colombo (1981), in his treatise of the subject, entitled *Numerical Methods for Harmonic Analysis on the Sphere*. We summarize in the following Colombo's developments and provide some examples of practical application of his formulations. Before this, however, we need to define and discuss briefly the error sources that affect any estimate of the spectrum of a function that is obtained on the basis of observed data, regardless of the technique that is used to obtain this spectrum.

- *Error Sources.* Two types of errors affect any estimate of the spectrum that is obtained from observations of a function: (a) Sampling error, and, (b) Propagated noise. The sampling error arises due to the fact that in reality, observations can only be made at a finite number of discrete points, while Equations 11 and 12 require the function $f(\theta,\lambda)$ to be known at *every* point on the surface S. Propagated noise, on the other hand, arises due to the fact that observations can never be error-free. Therefore, errors in the observations (both random and systematic) will generally propagate into the spectrum that is estimated from them.

Optimal Estimation. The particular way that sampling errors and propagated noise affect the estimated spectrum depends on the particular technique used to estimate the

spectrum. Colombo (1981) designed and formulated a linear estimator of the spectrum, which is optimal in the sense that it minimizes the sum of the squares of the sampling error and the propagated noise. His *Optimal Estimation* (OE) technique is a form of Numerical Quadrature, where the *optimal* quadrature weights are defined using the formalism of *Least Squares Collocation* (LSC) (Moritz, 1980). By exploiting the geometry of regular grids (i.e., grids where at least the longitude increment is constant) of data located over surfaces that have rotational symmetry (such as a sphere or a rotational ellipsoid), and by imposing certain conditions on the *covariance functions* of the signal and the data noise, Colombo (1981) demonstrated that the variance–covariance matrix of such data sets consists of *Toeplitz-circulant* blocks. These matrix blocks can be formed and inverted very efficiently, even when the size of the data set is very large. Colombo's OE technique was used in the gravitational field expansion developed by Hajela (1984), and in the OSU86C/D global gravitational models developed by Rapp and Cruz (1986a). Additional discussion about the estimation of gravitational spectra using LSC can be found in Tscherning (2001).

Colombo (1981, p. 76) also investigated the use of a semiempirical set of numerical quadrature weights, which depend only on the spherical harmonic degree, and can be evaluated very easily as a function of Pellinen's smoothing factors (Colombo, 1981, p. 85). Although suboptimal, these numerical quadrature weights proved to be quite adequate and were used in several gravitational modeling studies. The OSU86E/F (Rapp and Cruz, 1986b) and the OSU89A/B (Rapp and Pavlis, 1990) global gravitational models were developed using these semiempirical numerical quadrature weights.

Practical aspects

Several factors complicate the application of SHA techniques when analyzing real data. These are related on the one hand to the inherent properties of the alternative SHA techniques, and on the other to the properties of the data to be analyzed. The former have to do with the computational capabilities and requirements (i.e., the flexibility and efficiency) of the different SHA techniques. The latter include issues related to the distribution and type (point or area-mean values) of the available data, the geometry of the grid to which the original data are reduced, the spectral properties of the function $f(\theta, \lambda)$ that is being approximated, and the spectral properties of the errors that may affect the observed values $f^S_{obs}(\theta, \lambda)$. We discuss briefly these factors in the following paragraphs.

- *Computational Issues.* An inspection of Equations such as (11) indicates that a spherical harmonic expansion complete to degree and order N contains $(N+1)^2$ harmonic coefficients f^S_{nm} in total. Therefore, the normal matrix $\mathbf{N}$ of Equation 30 has dimensions $(N+1)^2 \times (N+1)^2$. The structure and characteristics of the matrix $\mathbf{N}$ depends on the geometry of the available data distribution and on the characteristics of the variance–covariance matrix $\mathbf{\Sigma}$ of the data errors. If the data to be analyzed are arbitrarily located in latitude and longitude, as is the case with most raw geophysical data collected in the field (e.g., gravimetric or magnetometric data), then, apart from symmetry, the matrix $\mathbf{N}$ would be fully occupied. The same is true (in general) if the error variance–covariance matrix $\mathbf{\Sigma}$ is arbitrary. The size of the normal matrix $\mathbf{N}$ is already formidable for expansions complete to degree and order 200 or so. Normal matrices for expansions to degree and order 2160, as the one corresponding to the EGM2008 model (Pavlis et al., 2008), are well beyond current computational capabilities. Nevertheless, provided that the size of the normal matrix can be handled computationally, LS adjustment and LSC are the *only* techniques that can handle arbitrarily located data, with arbitrary error variance–covariance matrices. NQ and OE require the data to be input in the form of a grid, and impose restrictions on the stochastic properties of the noise, represented within the matrix $\mathbf{\Sigma}$. In the following, we assume that the raw data have been preprocessed in some fashion, so that the input data to the various estimators are available in gridded form.

As we mentioned previously, unlike estimators based on orthogonality relations, LS permits also the use of *solid* spherical harmonics in the formation of observation equations. Assume for the sake of this example, that a grid of potential values V is available over the physical surface of the Earth (the topography). One may then form observation equations on the basis of the mathematical model given in Equation Equation 2. In such a case, even if the V values are given over a regular latitude-longitude grid, these values would still be located arbitrarily in the radial direction. This would then require the formation and inversion of a fully occupied normal matrix, as it is described by Pavlis (1988), who performed such estimations to degree 36, 50, and 70, using equiangular grids of free-air gravity anomalies defined on the surface of the Earth's topography.

- *Data Organization, Types, and Distribution.* In most geodetic and geophysical SHA applications, the data are organized in geographic grids. Equiangular grids are most often used, although grids with cells of equal area are also possible, and have been used in some cases. There is therefore a need to transition from the arbitrarily located point measurements, to "artificial" measurements referring to the nodes of some specified grid. In geodetic practice, this transition is accomplished by predicting what the "artificial" grid measurements would have been, given the available arbitrarily scattered data. The prediction is done using the LSC formalism, which requires some knowledge of the signal and noise covariance functions of the data (Moritz, 1980). One usually applies LSC to predict area-mean values for the grid cells, rather than point values. SHA can be applied equally well using either point or

area-mean gridded data. In the case of area-mean values, the mathematical models have to be modified appropriately, so that they accurately represent the area-mean character of the data. This is done by integrating the surface spherical harmonic functions over the specific latitudinal and longitudinal limits of each grid cell. From Equation 3 one has:

$$\int_{\theta_i}^{\theta_{i+1}} \int_{\lambda_j}^{\lambda_{j+1}} Y_{nm}(\theta,\lambda)\sin\theta d\theta d\lambda = \int_{\theta_i}^{\theta_{i+1}} \overline{P}_{n|m|}(\cos\theta)\sin\theta d\theta \int_{\lambda_j}^{\lambda_{j+1}} \begin{Bmatrix} \cos m\lambda \\ \sin|m|\lambda \end{Bmatrix} d\lambda \quad \begin{matrix} \text{if } m \geq 0 \\ \text{if } m < 0 \end{matrix}. \tag{36}$$

Efficient recurrence relationships for evaluating the integrals of the Associated Legendre functions have been derived by Paul (1978). The numerical treatment of the evaluation of Associated Legendre functions of very high degree (e.g., $n > 1,400$ or so) requires special care since some of the values of these functions become exceedingly small, which could cause numerical problems.

Even after the gridding of geophysical data collected on or near the Earth's surface, the radial location of these data, corresponding to the Earth's topography, remains arbitrary. Solutions of BVP that are based on orthogonality relations, formulated in either spherical or ellipsoidal harmonics, require these data to be located over a sphere or an ellipsoid, respectively. This artificial (approximately radial) "movement" of the data can be accomplished using *analytical continuation*. Wang (1987, 1988) discusses some techniques that can be applied to analytically continue gravity anomaly data.

In many geophysical problems where SHA techniques are applied, the available data do not suffice to cover completely the Earth. Data gaps can produce undesirable *leakage* effects in the spectral estimates, which by definition are of global character, and require global coverage. These undesirable effects are different depending on the particular technique used to estimate the spectrum. In LS, the spectral estimator "interprets" a data gap as an undefined data value with infinite error, i.e., zero weight (see Pavlis, 1988 for details). In NQ techniques, the estimator "interprets" a data gap as a data value that is equal to zero. In either case, the estimated spectrum would produce undesirable results when evaluated over the region of the data gap. It is therefore preferable to "fill in" data gaps with some reasonable values of the quantity under consideration. These "fill-in" values may be obtained on the basis of other data and/or models. For example, in gravity modeling, topographic elevations and models of the *isostatic compensation* may be used to "fill in" areas void of actual gravity measurements (cf. Pavlis and Rapp, 1990).

- *LS versus NQ techniques*. It is of interest to compare the LS and the NQ techniques from an analytical as well as a numerical perspective. Such studies have been reported by Rapp (1969, 1986), Colombo (1981), and Pavlis (1988). Sneeuw (1994) discussed the two techniques from a historical perspective. A brief review of the conclusions reached by these investigations follows. Detailed derivations and analyses supporting these conclusions can be found in the cited references. NQ is used here to identify the simple numerical quadratures formula with the semiempirical set of suboptimal quadrature weights proposed by Colombo (1981, p. 76).
  1. NQ determines each harmonic coefficient independently of all others. In contrast, LS estimates a correlated set of coefficients; thus, solutions to different maximum degrees will yield different values for the common harmonic coefficients. This will occur even if the input data are uncorrelated, and arises from the loss of orthogonality between the discrete (point or integrated) samples of the Associated Legendre functions (Pavlis, 1988, Sect. 4.2.1). Unlike the Legendre functions, discrete (point, as well as integrated) samples of $\cos m\lambda$ and $\sin m\lambda$ preserve their orthogonality in the interval $[0, 2\pi)$, as long as the sampling interval $\Delta\lambda$ is constant. It is this orthogonality property of sines and cosines, along with the equatorial symmetry and the parity properties of Legendre functions, which produce the sparsity patterns of the *Block-Diagonal* (*BD*) *Least Squares* technique that we discuss next.
  2. NQ cannot account for varying accuracies of the input data, while the LS estimator is capable of accounting for any (positive-definite) error variance–covariance matrix $\boldsymbol{\Sigma}$ associated with the input data.
  3. If $L(=\pi/\Delta\theta)$ denotes the *Nyquist degree* implied by the data sampling interval, then the normal equations formed based on observation equations like (27) become singular if $N \geq L$ (Colombo, 1981), where N is the maximum solved-for degree.
  4. LS estimation (and LSC) can recover *exactly* a set of coefficients from "synthetic" noiseless data, provided that the data are band-limited and their frequency content does not exceed the Nyquist degree. This property offers also one verification test that any software developed to perform SHA using the LS technique should pass (Pavlis, 1988, Sect. 4.2.2). The simple NQ technique is incapable of recovering the original "true" coefficients from which the "synthetic" data were computed, as Rapp's (1986) numerical experiments have also demonstrated.
  5. Unlike the LS technique, the simple NQ does not involve any matrix formation or inversion. In this

regard, the simple NQ is considerably less demanding computationally than LS. In addition, if the sampling interval $\Delta\lambda$ in the longitude direction is constant, then the NQ SHA algorithm can benefit enormously from the application of *Fast Fourier Transform* (FFT) techniques, as it was pioneered by Colombo (1981).

6. In contrast to the simple NQ, the LS technique produces not only the estimates of the coefficients, but also an estimate of their error variance–covariance matrix (see Equation 32). This matrix can then be used to propagate the errors of the estimated coefficients onto other quantities that may be computed from them.

Block-Diagonal (BD) Least Squares. From the previous discussion, it becomes clear that a SHA technique that would combine the computational efficiency of the simple NQ, with the rigor and flexibility of LS, is highly desirable. Such a technique would attempt to retain the advantages of both techniques, while avoiding their respective limitations. The Block-Diagonal (BD) Least Squares may be viewed as such a "best of both worlds" type of technique. Colombo (1981) has shown that if:

(a) The data reside on a surface of revolution (e.g., a rotational ellipsoid or a sphere)
(b) The grid is complete and the longitude increment is constant
(c) The data weights are longitude-independent
(d) The data weights are symmetric with respect to the equator

then, the elements of the normal matrix **N**, corresponding to the coefficients f^S_{nm} and f^S_{rq}, would be zero as prescribed by (see also Pavlis, 1988 for details):

$$[\mathbf{N}]_{f^S_{nm} f^S_{rq}} = 0 \quad \text{if } m \neq q, \quad \text{or if } m = q \text{ and } n - r = 2k + 1. \tag{37}$$

Note that in this notation the order subscript is a signed integer, whose sign identifies the type of coefficient (positive: cosine, negative: sine). If condition (d) does not hold true, then:

$$[\mathbf{N}]_{f^S_{nm} f^S_{rq}} = 0 \quad \text{if } m \neq q. \tag{38}$$

The sparsity patterns implied by Equations 37 and 38 will be referred to as BD1 and BD2 respectively. In addition, a BD3 pattern may be considered, defined by:

$$[\mathbf{N}]_{f^S_{nm} f^S_{rq}} = 0 \quad \text{if } |m| \neq |q|, \tag{39}$$

which admits nonzero off-diagonal elements across coefficients of different type within a given order. It is instructive to consider the computational efficiency implied by these patterns. Table 1, which is taken from Pavlis, pp. 8–5, in Lemoine et al. (1998), provides relevant statistics for an expansion complete from degree and order 0 to degree and order 360, excluding degree $n = 1$ terms. Such an expansion involves 130,318 unknown coefficients and the upper (or lower) triangular part of the symmetric but fully occupied normal matrix **N** has 8,491,455,721 elements in total.

The enormous computational savings that can be inferred from Table 1 make the BD approximations very attractive estimation techniques. These savings, however, come at the expense of the rigor exercised in the implementation of the SHA. Any approximation of a normal matrix by a BD structure should be simple enough to permit efficient numerical implementation, and, on the same time, rigorous enough to maintain the most important characteristics of the full matrix. Therefore, one has to consider carefully whether the approximations leading to the various BD sparsity patterns can be tolerated, given the characteristics of the particular data set that is analyzed. For example, in global gravitational modeling applications, the real-world gravity anomaly data to be analyzed comply *only* with the conditions (a) and (b) above. In fact, to comply even with the (a) and (b) conditions, "filling-in" techniques and analytical continuation have to be employed, since the original data set is neither complete, nor residing on any surface of revolution, since it refers to the Earth's topography. These aspects were carefully considered in the development of the OSU91A (Rapp et al., 1991) and the EGM96 (Lemoine et al., 1998) gravitational models. In both these models, different SHA techniques were used toward the estimation of different spectral portions of the models. This was due to the fact that different spectral portions of these models were determined on the basis of different gravimetric data sets.

In the case of *band-limited* data, one can show that the LS adjustment approach using a priori information is

Spherical Harmonic Analysis Applied to Potential Fields, Table 1 Statistics of normal matrices related to an expansion complete to degree $N = 360$ (excluding degree $n = 1$ coefficients) using different sparsity patterns

	Sparsity pattern		
Statistic	BD1	BD2	BD3
Total number of nonzero elements	7,905,721	15,746,100	31,362,241
Percentage of full matrix elements	0.09	0.19	0.37
Number of blocks	1,440	721	361
Number of unknowns in largest block	181	360	718
Number of elements in largest block	16,471	64,980	258,121

formally equivalent to LSC (Moritz, 1980, p. 166; Colombo, 1981, Sect. 2.13). In such a case, the BD approaches discussed previously are the Least Squares counterpart of Colombo's (1981) Optimal Estimation technique.

BD techniques of varying sophistication have been used to develop GPM2 (Wenzel, 1985), DGFI92A (Gruber and Bosch, 1992), GFZ95A (Gruber et al., 1996), the GPM98A, B, and C models (Wenzel, 1998, 1999), and the EGM2008 model (Pavlis et al., 2008).

Some of the BD structures of the normal matrix that were discussed above arise also in the analysis of data from certain configurations of *Satellite-to-Satellite Tracking* (SST) and of *Satellite Gravity Gradiometry* (SGG). This was also recognized and studied by Colombo (1984, 1989). Schuh (1996) provides a detailed study of the numerical solution strategies applicable to the analysis of such observables. SST data (in the "low-low" configuration) are currently available from the GRACE mission (GRACE, 1998), and SGG data are currently being collected from the GOCE mission (ESA, 1999). The availability of BD normal matrices resulting from the analysis of SST and SGG data is very important when considering the combination of these matrices with corresponding matrices derived from the analysis of global grids of gravimetric data. Such a combination of BD normal matrices permits the efficient development of global gravitational models to a very high degree (2160), as it was done in the case of EGM2008 (Pavlis et al., 2008).

Summary

Spherical Harmonic Analysis (SHA) is a critical element of the solution of Boundary Value Problems associated with potential fields. SHA is applicable not only to potential fields but also to a rather wide class of functions that are sampled over surfaces which have a one-to-one correspondence with the unit sphere. Least Squares (using either fully occupied or block-diagonal normal matrices), Numerical Quadrature, and Optimal Estimation techniques have been used to perform SHA within gravitational and magnetic field modeling efforts. The organization of the data to be analyzed in the form of geographic grids defined by meridians and parallels enables the use of highly efficient numerical algorithms for SHA. In particular, grids with constant longitudinal spacing of data permit the application of Fast Fourier Transform techniques, thereby increasing tremendously the efficiency of the SHA algorithms.

Bibliography

Blakely, R. J., 1995. *Potential Theory in Gravity and Magnetic Applications*. Cambridge: Cambridge University Press.

Colombo, O. L., 1981. *Numerical methods for harmonic analysis on the sphere*. Report 310, Department of Geodetic Science and Surveying, Ohio State University, Columbus.

Colombo, O. L., 1984. *The Global Mapping of Gravity with Two Satellites*. Netherlands Geodetic Commission, Publications on Geodesy, New Series, Vol. 7, No. 3. Delft.

Colombo, O. L., 1989. High resolution analysis of satellite gradiometry. In Rapp, R. H. (ed.), *Progress in the Determination of the Earth's Gravity Field*, Extended Abstracts from the meeting held in Ft. Lauderdale, Florida, September 13–16, 1988. Report 397, Department of Geodetic Science and Surveying, Ohio State University, Columbus.

Driscoll, J. R., and Healy, D. M., Jr., 1994. Computing Fourier transforms and convolutions on the 2-sphere. *Advances in Applied Mathematics*, **15**, 202–250.

ESA SP-1233 (1), 1999. *The Four Candidate Earth Explorer Core Missions – Gravity Field and Steady-State Ocean Circulation Mission*. Noordwijk: ESA Publications Division, ESTEC.

Gleason, D. M., 1988. Comparing ellipsoidal corrections to the transformation between the geopotential's spherical and ellipsoidal spectrums. *Manuscripta Geodaetica*, **13**, 114–129.

GRACE, 1998. – Gravity Recovery and Climate Experiment: Science and Mission Requirements Document, revision A, *JPLD-15928*, NASA's Earth System Science Pathfinder Program.

Gruber, T., and Bosch, W., 1992. *A new 360 gravity field model, presented at the XVII General*. Assembly of the European Geophysical Society, Edinburgh, Scotland.

Gruber, T., Anzenhofer, M., and Rentsch, M., 1996. The 1995 GFZ high resolution gravity model. In Rapp, R. H., Cazenave, A. A., and Nerem, R. S. (eds.), *Global Gravity Field and Its Temporal Variations*. Berlin/Heidelberg: Springer. IAG Symposia, Vol. 116.

Hajela, D. P., 1984. *Optimal estimation of high degree gravity field from a global set of 1° × 1° anomalies to degree and order 250*. Report 358, Department of Geodetic Science and Surveying, Ohio State University, Columbus.

Heiskanen, W. A., and Moritz, H., 1967. *Physical Geodesy*. San Francisco: W.H. Freeman.

Hotine, M., 1969. *Mathematical Geodesy*. Washington, DC: U.S. Department of Commerce. ESSA Monograph, Vol. 2.

Jekeli, C., 1988. The exact transformation between ellipsoidal and spherical harmonic expansions. *Manuscripta Geodaetica*, **13**, 106–113.

Lemoine, F. G., Kenyon, S. C., Factor, J. K., Trimmer, R. G., Pavlis, N. K., Chinn, D. S., Cox, C. M., Klosko, S. M., Luthcke, S. B., Torrence, M. H., Wang, Y. M., Williamson, R. G., Pavlis, E. C., Rapp, R. H., and Olson, T. R., 1998. The development of the joint NASA GSFC and the National Imagery and Mapping Agency (NIMA) geopotential model EGM96. NASA Tech. Publ. TP-1998-206861. 575 pp.

Moritz, H., 1980. *Advanced Physical Geodesy*. Karlsruhe: Herbert Wichmann.

Paul, M. K., 1978. Recurrence relations for integrals of Associated Legendre functions. *Bulletin Géodésique*, **52**, 177–190.

Pavlis, N. K., 1988. *Modeling and estimation of a low degree geopotential model from terrestrial gravity data*. Report 386, Department of Geodetic Science and Surveying, Ohio State University, Columbus.

Pavlis, N. K., and Rapp, R. H., 1990. The development of an isostatic gravitational model to degree 360 and its use in global gravity modelling. *Geophysical Journal International*, **100**, 369–378.

Pavlis, N. K., Holmes, S. A., Kenyon, S. C., and Factor, J. K., 2008. An earth gravitational model to degree 2160: EGM2008. In *Presented at the 2008 General Assembly of the European Geosciences Union*, April 13–18, Vienna, Austria.

Rapp, R. H., 1969. Analytical and numerical differences between two methods for the combination of gravimetric and satellite data. *Bollettino di Geofisica Teorica ed Applicata*, **XI**, 108–118.

Rapp, R. H., 1986. Global geopotential solutions. In Sünkel, H. (ed.), *Mathematical and Numerical Techniques in Physical Geodesy*. Berlin/Heidelberg: Springer. Lecture Notes in Earth Sciences, Vol. 7.

Rapp, R. H., and Cruz, J. Y., 1986a. *The representation of the Earth's gravitational potential in a spherical harmonic*

expansion to degree 250. Report 372, Department of Geodetic Science and Surveying, Ohio State University, Columbus.
Rapp, R. H., and Cruz, J. Y., 1986b. *Spherical harmonic expansions of the Earth's gravitational potential to degree 360 using 30′ mean anomalies*. Report 376, Department of Geodetic Science and Surveying, Ohio State University, Columbus.
Rapp, R. H., and Pavlis, N. K., 1990. The development and analysis of geopotential coefficient models to spherical harmonic degree 360. *Journal of Geophysical Research*, **95**(B13), 21885–21911.
Rapp, R. H., Wang, Y. M., Pavlis, N. K., 1991. *The Ohio State 1991 geopotential and sea surface topography harmonic coefficient models*. Report 410, Department of Geodetic Science and Surveying, Ohio State University, Columbus.
Schuh, W. D., 1996. *Tailored numerical solution strategies for the global determination of the Earth's gravity field*. Technical Report, Institute of Theoretical Geodesy, Technical University of Graz, Austria.
Sneeuw, N., 1994. Global spherical harmonic analysis by least-squares and numerical quadrature methods in historical perspective. *Geophysical Journal International*, **118**, 707–716.
Tscherning, C. C., 2001. Computation of spherical harmonic coefficients and their error estimates using least-squares collocation. *Journal of Geodesy*, **75**, 12–18.
Wang, Y. M., 1987. Numerical aspects of the solution of Molodensky's problem by analytical continuation. *Manuscripta Geodaetica*, **12**, 290–295.
Wang, Y. M., 1988. *Downward continuation of the free-air gravity anomalies to the ellipsoid using the gradient solution, Poisson's integral and terrain correction – numerical comparison and the computations*. Report 393, Department of Geodetic Science and Surveying, Ohio State University, Columbus.
Wenzel, H.-G., 1985. *Hochauflösende Kugelfunktionsmodelle für das Gravitationspotential der Erde. Wiss. Arb.* 137, Fachrichtung Vermess. der Univ. Hannover, Hannover, Federal Republic of Germany.
Wenzel, G., 1998. Ultra high degree geopotential models GPM98A, B and C to degree 1800. http://www.gik.uni-karlsruhe.de/~wenzel/gpm98abc/gpm98abc.htm.
Wenzel, G., 1999. Schwerefeldmodellierung durch ultrahochauflösende Kugelfunktionsmodelle. *Zeitschrift für Vermessungswesen*, **124**(5), 144–154.

Cross-references

Geoid
Gravity Method, Satellite
Gravity, Global Models
Gravity, Gradiometry
Magnetic, Global Anomaly Map
Numerical Methods, Boundary Element

STATISTICAL SEISMOLOGY

David A. Rhoades
GNS Science, Lower Hutt, New Zealand

Synonyms

Earthquake statistics; Statistical methods in seismology

Definition

Stochastic process. A process evolving in time governed partly by probability distributions.

Point process. Stochastic process of point events in time or space or both.

Renewal process. Point process in which the probability distribution of the time to the next event depends only on the time since the last event.

Stationary Poisson process. Point process in which the numbers of events in distinct time intervals of the same length are independent and have the same distribution.

Branching process. Process involving ancestors and offspring.

Introduction

Statistical seismology is an emerging field of research at the interface of statistical and physical modeling of earthquake occurrence. It is concerned with quantifying and understanding the distribution of earthquakes in time, magnitude, and location. It includes empirical analysis of earthquake catalogues, stochastic modeling of earthquake occurrence, estimation of the probability of earthquake occurrence and of earthquake-induced ground shaking, and testing the forecasting power of physical and statistical models of seismicity (Vere-Jones et al., 2005). It provides input to the management of seismic hazard and risk. Although efforts have so far focused mainly on the study of earthquake catalogues, there is scope to incorporate other geophysical measurements into the analyses.

Empirical relations

One of the classical empirical relations of statistical seismology is the Gutenberg and Richter (1944) law, according to which the distribution of earthquake magnitudes M within a large time-space volume has a probability density proportional to 10^{-bM}, where the parameter b is close to 1. Physical upper limits to earthquake size necessitate tapering of the density at high magnitudes (Kagan, 1999). For small space-time volumes and in certain tectonic settings, the b-value can differ markedly from 1. Temporal variation in b-value has been proposed as an earthquake precursor. Since the radiated seismic energy E_S of an earthquake is approximately proportional to $10^{1.5M}$, the Gutenberg–Richter law implies that the distribution of radiated seismic energy has density proportional to $E_S^{-(\beta+1)}$, where $\beta = 2/3b$. Thus, the Gutenberg–Richter law can be viewed as a power law. The existence of such power laws is often taken as evidence of fractality and self-organized criticality in the earthquake process (Bak and Tang, 1989).

Earthquakes tend to occur in clusters. A common type of cluster is the main shock – aftershock sequence. The Omori–Utsu law, first noted by Fusakichi Omori in 1894, describes how an aftershock sequence decays over time (Utsu et al., 1995). According to this relation, which is also a power law, the rate of aftershock occurrence is proportional to $(t + c)^{-p}$, where t is the elapsed time following the occurrence of the main shock, and c and p are adjustable parameters, typically with $c << 1$ day, and $p \approx 1$. In some large aftershock sequences, the relation

can be seen to hold for decades after the occurrence of the main shock.

The area occupied by an aftershock sequence is approximately 10^{M-4} km^2, where M is the main shock magnitude (Utsu, 1961). This is consistent with scaling relations derived for earthquake source dimensions – fault length, width, and displacement – all of which are approximately proportional to $10^{0.5M}$ (Wells and Coppersmith, 1994). However, sometimes major earthquakes appear to trigger aftershocks at long distances from the earthquake source. These can be accommodated by a power law in which the aftershock density is proportional to $(d^2 + r^2)^{-q}$, where r is the distance from the source, q is an adjustable parameter, typically about 1.5, and d is a function proportional to$10^{0.5M}$.

According to Båth's law (Båth, 1965), the largest aftershock is typically about 1.2 magnitude units smaller than the main shock. This indicates that the main shock and its aftershocks do not by themselves conform to the Gutenberg–Richter relation, the main shock being too large relative to the other earthquakes in the set. The existence of other types of earthquake clusters than main shock – aftershock sequences, such as swarms (which have no main shock) and multiplets (which may have several main shocks) is a confounding element in attempts to systematically examine Båth's relation.

Rules for defining and classifying earthquake clusters are necessarily somewhat arbitrary. Sometimes a stochastic model of earthquake clustering is used to calculate the probability that a given earthquake belongs to a cluster, and then the analysis considers many different possible groupings of the earthquakes into clusters (Zhuang et al., 2002).

Data quality is an ever-present issue affecting statistical analysis of earthquake catalogues. It is necessary to establish the quality of a catalogue at the outset, because an analysis can be seriously compromised by changes in the quality of the catalogue over time, such as temporal and spatial variation of the magnitude threshold of completeness. On the positive side, catalogue data are generally improving in both quality and quantity as better seismograph networks and earthquake location techniques are employed, and this is creating opportunities for more detailed analyses.

Precursors

Studies aimed at identifying precursors of large earthquake are faced with a dual challenge: Whereas large earthquakes are relatively rare, the number of degrees of freedom available when attempting to identify precursors is large. Therefore, a careful analysis is required before the conclusion is drawn that a proposed precursory phenomenon is real. A retrospective analysis can be used to define a possible precursory phenomenon, but prospective testing is necessary to confirm it.

Many major earthquakes are preceded in the long term by an increase in the rate of occurrence and magnitude of minor earthquakes, in an area not much larger than the major earthquake source. This is called the precursory scale increase (Evison and Rhoades, 2004). Sometimes, it consists of a sequence of swarms. The magnitude of the largest precursor is typically about one unit smaller than that of the main shock, and the precursor time T_P (the time from the onset of the increase to the main shock) and the area in which the precursors, major earthquake, and aftershocks all take place are both approximately proportional to $10^{0.5M}$. Earthquake precursors for which T_P is proportional to $10^{0.5M}$ are known as precursors of the first kind (Rikitake, 1982), and include anomalies in tilt, strain and land deformation, and changes in the relative velocities of seismic P and S waves. A distinguishing feature of the precursory scale increase is that the magnitude of the largest precursor(s) can be used to predict the time of occurrence, magnitude, and source area of the major earthquake.

Models of earthquakes as a critical phenomenon have been invoked to suggest that an accelerating moment release (AMR) should occur in an area much larger than the earthquake source in the approach of criticality, i.e., an accelerating occurrence of minor earthquakes leading up to a major earthquake. Another much-studied phenomenon is precursory seismic quiescence (PSQ) – a proposed reduction in the rate of occurrence of minor earthquakes shortly before a major event. There is apparent empirical support for both AMR and PSQ, but its statistical significance has been questioned. A problem in resolving such matters is that these phenomena, like many other proposed precursors, have only been defined anecdotally and never identified with an explicit stochastic model of earthquake occurrence.

Stochastic models of earthquake occurrence

Early examples of stochastic modeling of earthquake sequences were given by Vere-Jones (1970). Modeling of earthquake occurrence has benefited from the development of the theory of stochastic point processes (Daley and Vere-Jones, 2003).

The idea of successive episodes of stress accumulation and release, as suggested by the elastic rebound theory of Henry Fielding Reid in 1910, is the basis for most modeling of recurrence of major earthquakes on a fault or fault segment. The time sequence of earthquakes is often modeled as a renewal process, assuming a characteristic magnitude for events on the fault. The recurrence-time distribution is variously taken as exponential (for constant hazard), Weibull, lognormal, or Brownian passage-time. The latter three distributions all imply some degree of regularity in the time intervals between earthquakes, and physical arguments have been advanced in support of each. The limited data so far available on earthquake recurrence on fault segments do not support a clear preference for a particular distribution or afford a real opportunity to test how informative renewal models are in this context. It is important to account for data and parameter uncertainties when applying such models to the estimation of earthquake hazard.

The renewal process model does not allow for interactions between neighboring faults or for variation in the size of successive events. Models involving time and magnitude include the time-predictable and slip-predictable models, in which either the time to the next earthquake depends on the magnitude of the last earthquake or the magnitude of the next earthquake depends on the elapsed time. These models are sometimes applied to the earthquakes in a region. In a stress-release model of Zheng and Vere-Jones (1994), also applicable to the earthquakes in a region, the hazard rate steadily increases over time, but drops upon the occurrence of an earthquake by an amount that depends on its magnitude. A coupled version of this model, allowing for stress transfer between discrete regions, has also been developed. Further tests are needed of how well these models perform.

The epidemic-type aftershock (ETAS) model of Ogata (1989) is a nonstationary Poisson process model developed originally as a model of temporal clustering of earthquakes in a discrete region. In this model, each earthquake has its own aftershock sequence, which decays over time according to the Omori-Utsu law. The number of aftershocks is governed by a magnitude-dependent productivity function. Some earthquakes occur independently according to a stationary Poisson process, and the magnitudes of all earthquakes follow the Gutenberg–Richter relation. The ETAS model fits earthquake data much better than a stationary Poisson process, and better than a model in which only the larger events have aftershocks. It is a type of branching process model. A double branching process model has been found to better describe the time-variation of earthquake occurrence (Marzocchi and Lombardi, 2008).

Stochastic models involving location, as well as time and magnitude, allow for a more realistic representation of the earthquake process, including clustering and long-range interactions between earthquakes.

The ETAS model was extended to include the spatial variable (Ogata, 1998) by introducing a distribution for aftershock location, usually a power law of distance similar to that described above. In this form, it can be used to describe the space-time clustering of earthquakes and also as a diagnostic tool for physical changes affecting the earthquake process. As a short-term forecasting model, this model performs far better than smoothed seismicity models, which capture the space and magnitude distributions of past earthquakes but are time-invariant. It is successful at forecasting aftershocks, as well as the minor proportion of major earthquakes which are preceded by a foreshock sequence. For a given location, the rate density of earthquake occurrence in this model can fluctuate in a short time over several orders of magnitude. Changes in the parameters of the spatial ETAS model have been used as indirect indicators of stress-changes in regions of the crust.

The Every Earthquake a Precursor According to Scale (EEPAS) model (Rhoades and Evison, 2004) is a space-time-magnitude model designed for longer-term forecasting of the major earthquakes. In this model, every earthquake contributes to the future distribution of hazard in time, magnitude, and location on a scale determined by its magnitude, through the predictive scaling relations associated with the precursory scale increase phenomenon. Under this model, the rate density can fluctuate slowly in time over about 1.5 orders of magnitude. The EEPAS model outperforms smoothed seismicity models in well-catalogued inter-plate regions such as California, Japan, and New Zealand.

Testing of forecasting methods

Time-varying models of earthquake occurrence have not yet been widely adopted for practical purposes. The stationary Poisson model with characteristic earthquake magnitudes on faults and spatially distributed background seismicity conforming to the Gutenberg–Richter magnitude relation are the basis for most seismic hazard models in practical use. But an international effort by the Collaboratory for the Study of Earthquake Predictability (CSEP) to provide transparent, verifiable prospective tests of time-varying earthquake occurrence models is in progress. Regional earthquake forecast testing centers have been established in California, New Zealand, Europe and Japan, and others are planned. To be testable, a model must provide estimates of the expected number of earthquakes for future time windows within grid cells finely delimited by location and magnitude. Several different time steps are used; for example, 24 h, 3 months, and 5 years. An overview of models submitted for testing in California was given by Field (2007) and first results were presented by Schorlemmer et al. (2010).

The performance of a probabilistic model of earthquake occurrence on a target set of earthquakes in a catalogue independent of the one to which it was fitted, is conveniently measured by the increase in the log likelihood of the earthquake catalogue under the model compared with that under a reference model, such as a stationary Poisson model with spatial smoothing of the locations of past earthquakes.

The CSEP testing centers use additional likelihood-based tests, which aim to identify significant differences between the total number of earthquakes expected and the number observed, and the distributions in time, magnitude, and location of cell expectations and those of the targeted earthquakes (Schorlemmer et al., 2007).

Some earthquake forecasting methods, such as the M8 algorithm, are alarm-based rather than probabilistic. For alarm-based methods, the error diagram (Molchan, 1990), in which the proportion of time or space-time for which a certain level of alarm is exceeded is plotted against the proportion of unpredicted earthquakes, is a useful assessment tool.

Analysis of synthetic earthquake catalogues

Physics-based numerical models of earthquake occurrence (Ben Zion, 2008) can be used to generate synthetic earthquake catalogues incorporating accepted physical

elements such as stress transfer between faults by means of elastic, viscoelastic, and other interactions; frictional or other material instabilities acting on fault surfaces; increasing stress accumulation due to tectonic forces; and the detailed geometry of the fault surfaces. For synthetic catalogues, important physical variables, which cannot be measured in the real earth, such as the distribution of stress, can be tracked in detail. Statistical analysis of synthetic earthquake catalogues can reveal under what physical conditions a particular statistical model is likely to be appropriate and contribute to physical understanding of empirical relations derived from earthquake catalogues.

Conclusion

Statistical seismology is concerned with modeling the empirical observations of earthquake occurrence and testing models against observations while trying to throw light on the physics of earthquake generation. A rapid improvement in seismological databases has been a stimulus to recent developments in the field, and this state of affairs is likely to continue. Improved earthquake catalogues and other systematically collected data, such as earth deformation data derived from Global Positioning System (GPS) networks offer opportunities for enhanced physical understanding of the earthquake process and more informative stochastic models. At the same time, increased computing power allows for increased detail and complexity in physics-based modeling. The major challenge statistical seismology faces is to use the new data, together with physical insights from empirical studies and detailed physical modeling, to develop ever more informative stochastic models of earthquake occurrence that can be applied to mitigating earthquake hazard.

Bibliography

Bak, P., and Tang, C., 1989. Earthquakes as a self-organized critical phenomenon. *Journal of Geophysical Research*, **94**, 15635–15637.

Båth, M., 1965. Lateral inhomogeneities of the upper mantle. *Tectonophysics*, **2**, 483.

Ben-Zion, Y., 2008. Collective behavior of earthquakes and faults: continuum-discrete transitions, progressive evolutionary changes and different dynamic regimes. *Reviews of Geophysics*, **46**(4), RG4006.

Daley, D. J., and Vere-Jones, D., 2003. *An Introduction to the Theory of Point Processes*, 2nd edn. New York: Springer, Vol. 1.

Evison, F. F., and Rhoades, D. A., 2004. Demarcation and scaling of long-term seismogenesis. *Pure and Applied Geophysics*, **161**, 21–45.

Field, E. H., 2007. Overview of working group for the development of regional earthquake likelihood models (RELM). *Seismological Research Letters*, **78**, 7–16.

Gutenberg, B., and Richter, C. F., 1944. Frequency of earthquakes in California. *Bulletin of the Seismological Society of America*, **34**, 185–188.

Kagan, Y. Y., 1999. Universality of the seismic moment-frequency relation. *Pure and Applied Geophysics*, **155**, 537–573.

Marzocchi, W., and Lombardi, A. M., 2008. A double branching model for earthquake occurrence. *Journal of Geophysical Research*, **113**, B08317, doi:10.1029/2007JB005472.

Molchan, G. M., 1990. Strategies in strong earthquake prediction. *Physics of the Earth and Planetary Interiors*, **61**, 84–98.

Ogata, Y., 1989. Statistical models for standard seismicity and detection of anomalies by residual analysis. *Tectonophysics*, **169**, 159–174.

Ogata, Y., 1998. Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, **50**(2), 379–402.

Rhoades, D. A., and Evison, F. F., 2004. Long-range earthquake forecasting with every earthquake a precursor according to scale. *Pure and Applied Geophysics*, **161**, 47–71.

Rikitake, T., 1982. *Earthquake Forecasting and Warning*. Center for Academic Publications, Tokyo/Reidel, Dordrecht, 402 p.

Schorlemmer, D., Gerstenberger, M. C., Wiemer, S., Jackson, D. D., and Rhoades, D. A., 2007. Earthquake likelihood model testing. *Seismological Research Letters*, **78**, 17–29.

Schorlemmer, D., Zechar, J. D., Werner, M., Jordan, T. H., and the RELM Working Group, 2010. First results of the regional earthquake likelihood models experiment. *Pure and Applied Geophysics*, **167**, 859–876.

Utsu, T., 1961. A statistical study on the occurrence of aftershocks. *Geophysical Magazine*, **30**, 521–605.

Utsu, T., Ogata, Y., and Matsu'ura, R. S., 1995. The centenary of the Omori formula for a decay law of aftershock activity. *Journal of Physics of the Earth*, **43**, 1–33.

Vere-Jones, D., 1970. Stochastic models for earthquake occurrence. *Journal of the Royal Statistical Society*, **B 32**, 1–62.

Vere-Jones, D., Ben-Zion, Y., and Zuniga, R. (eds.), 2005. Statistical seismology, Pageoph Topical Volumes. *Pure and Applied Geophysics*, **162**(6/7), 1023–1396.

Wells, D. L., and Coppersmith, K. J., 1994. New empirical relations among magnitude, rupture length, rupture width, rupture area, and surface displacement. *Bulletin of the Seismological Society of America*, **84**, 974–1002.

Zheng, X., and Vere-Jones, D., 1994. Further applications of stress release models to historical earthquake data. *Tectonophysics*, **229**, 101–121.

Zhuang, J., Ogata, Y., and Vere-Jones, D., 2002. Stochastic declustering of space-time earthquake occurrence. *Journal of the American Statistical Association*, **97**, 369–380.

Cross-references

Artificial Water Reservoir Triggered Earthquakes
Characteristic Earthquakes and Seismic Gaps
Earthquake Precursors and Prediction
Earthquake Prediction, M8 Algorithm
Earthquake, Magnitude
Earthquakes and Crustal Deformation
Earthquakes, Early and Strong Motion Warning
Earthquakes, Energy
Seismic Data Acquisition and Processing
Seismic Quiescence and Activation

SUBDUCTION ZONES

Geoffrey A. Abers
Lamont-Doherty Earth Observatory of Columbia University, Palisades, NY, USA

Definition

Subduction zones are convergent plate boundaries involving at least one oceanic plate. The oceanic plate descends

beneath the other into the deep mantle, forming a deep-sea trench, often a volcanic arc subparallel to it, an inclined Wadati–Benioff seismic zone, and other subparallel deforming belts.

Introduction

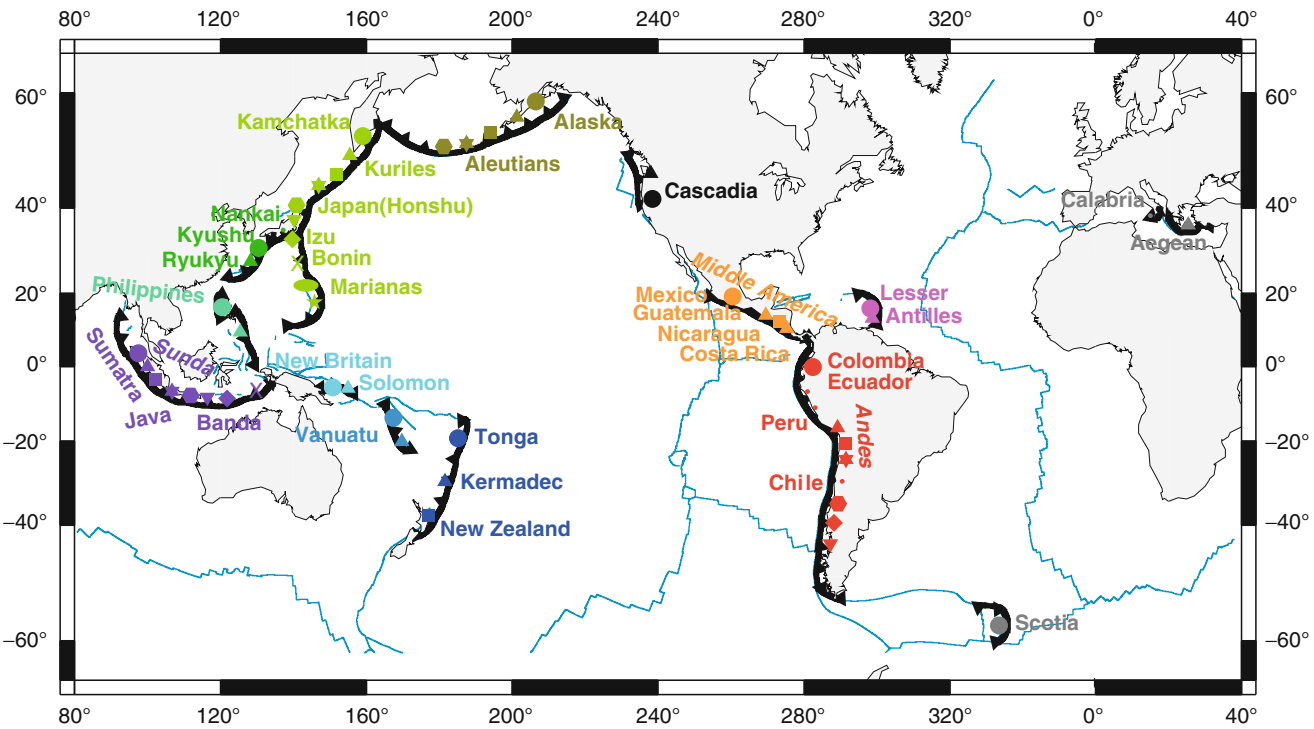
Subduction zones are a fundamental manifestation of planetary convection at the Earth's surface, and can be viewed as regulating two critical systems:

1. They control the primary material flux from the surface to the Earth's interior. Material at the seafloor subducts to depth, some of which emerges in volcanic arcs. These processes control the long-term budgets of volatiles (H_2O, CO_2, etc.), and the major accretion of material to continental crust.
2. Forces at subduction zones result in much of the planet's deformation, including the generation of the largest earthquakes and explosive volcanic eruptions over time scales of seconds, to the formation of the deepest seafloor and the growth of continents over millions of years. The negative buoyancy of subducting lithosphere probably represents the largest driver of global plate motions.

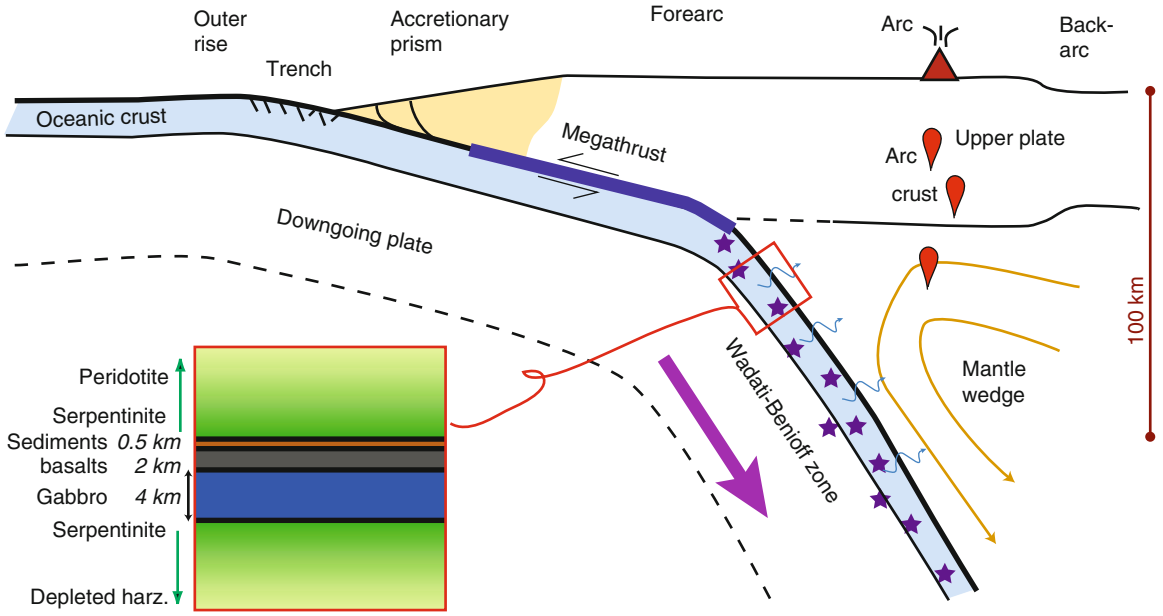
All of these processes ultimately result from the Earth releasing heat, and the deformation and thermal structure that results. At the largest scale, subduction regulates the mixing of material throughout the mantle of the lifetime of the planet, and the onset and termination of subduction represent profound geologic events. This article follows subduction zones from trench to depth, emphasizing the two themes of material flux and of large-scale deformation. It focuses on active systems from outer rise to sub-arc depths.

Morphology and nomenclature

The global subduction oceanic system is over 40,000 km long (Stern, 2002), largely but not entirely subducting oceanic plates of the Pacific basin (Figure 1). For the most part, any part of this subduction zone exhibits a common morphology (Figure 2): a downgoing, oceanic plate flexes into a deep-sea trench, beneath a submarine forearc that sometimes includes an accretionary prism. As the plate interface reaches 10–50 km depth, plate motion is accommodated on a shallow-dipping thrust zone, which in many places can generate great earthquakes. Landward of the thrust zone lies a volcanic arc, typically with a sharp front that overlies the region where the Wadati–Benioff zone seismicity reaches 75–180 km depth, with occasional volcanism further in the back-arc. The upper plate beneath and behind the arc varies between those that generate oceanic back-arc basins, exemplified by the Mariana-Izu-Bonin and Tonga systems, and those that undergo long-term convergence and compression, exemplified by the Andes, with a spectrum of behaviors in between (e.g., Lallemand et al., 2005). This characteristic morphology is a consequence of a regular, repeatable balance between forces that drive plate motions, buoyancy of



Subduction Zones, Figure 1 The global distribution of subduction zones. (Modified from Syracuse et al., 2010.)



Subduction Zones, Figure 2 Common features of subduction zones.

oceanic lithosphere, metamorphic devolatilization, and its rheological/magmatic consequences.

Outer rise, trench, and forearc

The shallow part of subduction systems, where temperatures are relatively cold, are controlled by the mechanics of elastic and brittle deformation. The trench occurs because the downgoing plate is flexed downward, and the shape of the plate seaward of the trench resembles closely that expected for flexure of a thin elastic surface (e.g., Watts and Talwani, 1974). The physics of elastic flexure also explains the outer rise – a region of slightly (100–500 m) elevated seafloor that lies seaward of the trench, typically 50–300 km distant (Parsons and Molnar, 1976). Strong bending leads to faulting in the outer rise, typically manifest as a series of normal faults with <1,000 m offset, seen on the seafloor and generating earthquakes (Chapple and Forsyth, 1979). These normal faults are found seaward of many if not most trenches, and are the most seaward indication of deformation within the subduction zone. They may play an important role in the material cycle as well, since the basins formed by them serve as traps for sediment that can be subducted to great depths (Hilde, 1983), and the faults themselves may act as conduits for water to reach into the mantle of the subducting plate, altering the mineralogy (e.g., Ranero et al., 2003). Such alteration of seafloor in the open ocean plays an important role in preconditioning the downgoing plate prior to subduction: the later metamorphic devolatilization of hydrous sediments and altered oceanic lithosphere releases fluids that lubricate the plate boundary, weaken the overlying material, and ultimately drive arc volcanism.

The sediment supply to the downgoing plate varies enormously between trenches near major continental river systems and those remote from such sources, leading to several variations in subduction zone behavior and characteristics. In the most extreme cases sediment supply overwhelms the rate at which material is subducted and accreted, so the bathymetric trench becomes a sediment-filled basin. Examples include the Lesser Antilles, Makran, and Cascadia subduction zones. A first-order distinction exists between subduction zones considered "accretionary," where a substantial amount of sediment has been offscraped to form an active accretionary prism, and "non-accretionary" subduction zones where all sediment is being subducted bypassing the forearc, and generally sediment supply is low (von Huene and Scholl, 1991). The accretionary prisms typically deform internally to form a critical-taper wedge, a geometry in which gravitational stresses created by the bathymetric slope are balanced by internal friction, similar to wedges in front of bulldozers or snowplows (Davis et al., 1983). Accreted material includes deep-sea sediment, but also fragments of bathymetric features such as seamounts on the incoming plate; subduction of relatively rigid seamounts can have a profound effect on the local morphology of the forearc, and may contribute to heterogeneity along the thrust zone at greater depth.

Thrust zone

As of mid-2010, the nine largest recorded earthquakes, and 22 of the largest 25, occurred on subduction thrust zones or "megathrusts," including all earthquakes with magnitudes of 9.0 and larger (U.S. Geol. Surv. Web site, http://earthquake.usgs.gov/). These great earthquakes also

generate most of the largest tsunamis, because they can displace large portions of the seafloor (see chapter on *Tsunami*). Great earthquakes occur here because the plate-boundary fault system can sustain large rupture areas, and earthquake size scales with rupture area. There are two reasons for the large rupture areas. First, at subduction zones cold material advects downward quickly compared with rates of heat conduction, as discussed below, so temperatures along the fault zone can remain below that of the brittle-ductile transition (or downdip limit of unstable sliding; Scholz, 1998) to greater depths than other environments, typically 30–50 km (Tichelaar and Ruff, 1993). Second, subduction zone thrusts dip at 5–25°, compared with steeper dips for faults in many other tectonic settings, so a greater fault area exists above the brittle-ductile transition. Both of these reasons allow for faults that can exceed 100 km in the downdip direction (e.g., for the Gulf of Alaska megathrust). Along strike, individual ruptures can extend for more than 1,000 km in some cases, most notably Chile (1960; Mw = 9.5) and Sumatra (2004; Mw = 9.2). Magnitudes here (Mw) are moment magnitudes (see chapter on *Earthquake, Magnitude*).

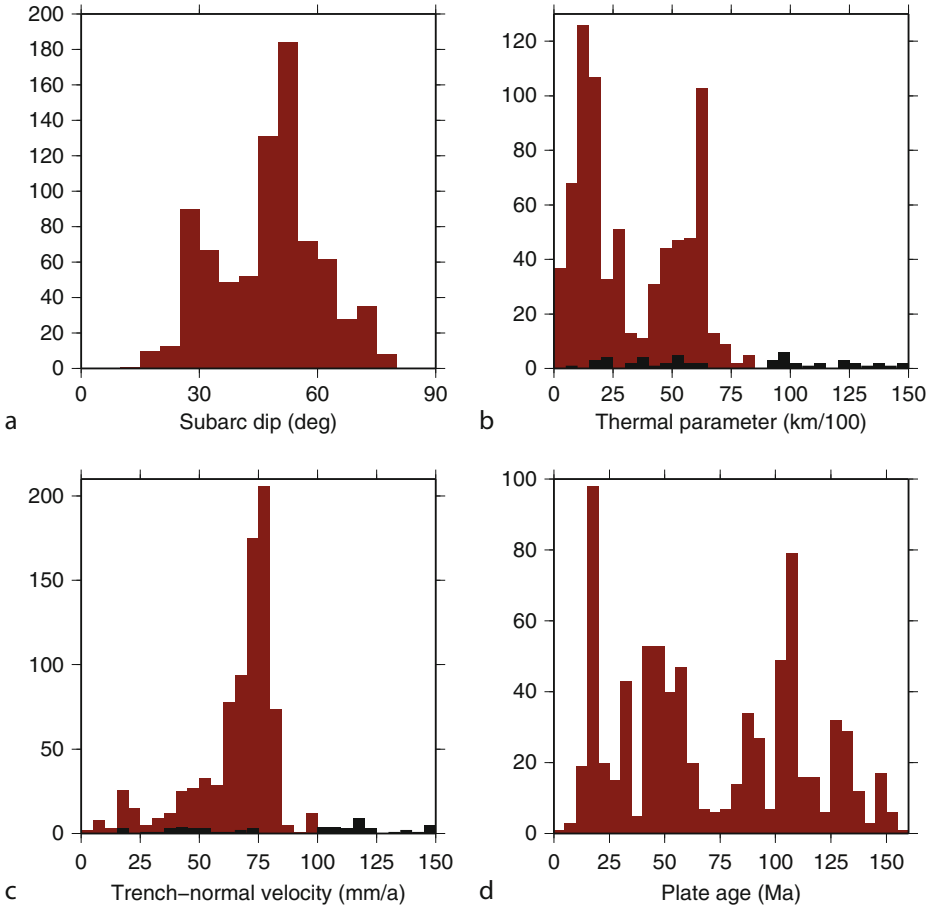
The fault area, and hence the ultimate size of earthquakes, is controlled by updip and downdip changes in material properties. At its updip end, the thrust zone undergoes a transition from aseismic slip on a decollement (subhorizontal plate-boundary fault) to stick–slip behavior or earthquake rupture, at depths of a few kilometers. At these depths the plate interface likely lies at temperatures of 75–150°C. The causes of this transition likely have to do with compaction, lithification, and low-temperature mineral phase transitions that change the frictional properties of this fault surface (Saffer and Marone, 2003). Similarly, the downdip limit can be in many places approximated by a threshold temperature, in this case near 350–500°C, with notable exceptions (e.g., Hyndman et al., 1997). Both transitions must be a consequence of change in frictional properties at this depth, with aseismic creep taking place at greater depths (e.g., Scholz, 1998). In between, the megathrust fault zone shows a variety of slip behaviors from purely creeping or "decoupled" to fully seismic or "coupled." In this usage, "coupling" refers to the ratio of fault slip inferred from summing the seismic moment of earthquakes over some time interval, to the total plate motion expected across that boundary over the same time interval. Coupled or "locked" thrust zones can be identified with geodesy, which shows the elastic strain accumulation building up between large earthquakes.

Recent discoveries of slow, silent creep events indicate that the downdip transition (and perhaps the updip one) may be much more complicated (e.g., Rogers and Dragert, 2003; Ito et al., 2007; Obara, 2002). Continuous geodetic observations have shown that patches of the thrust zone can slip over periods of days to months, in some cases accounting for a large fraction of the plate motion within the patch that slips (usually, adjacent to the locked zone). In some cases these events repeat quasi-periodically at time scales of months to years, although periodicity is not yet established for many regions of slow slip (Schwartz and Rokosky, 2007). Often, the slow slip is accompanied by a variety of exotic seismic phenomena, including nonvolcanic tremor, low-frequency earthquakes, very-low-frequency earthquakes, and creep (Ide et al., 2007). In most cases, the tremor seems to come from the plate boundary or just above it and shows motion consistent with thrusting on the plate boundary; these observations are subject of much current research (Gomberg, 2010). The physics of these quasi-periodic slip transients remain poorly understood; excess fluid pressure is suspected to play a role.

Kinematics

Subduction zones can be divided between those with little upper-plate deformation, with upper-plate compression, and exhibiting back-arc opening. The highest subduction rates, and indeed the highest relative plate motion rates anywhere, reach 240 mm/year at the northern end of the Tonga subduction system (Bevis et al., 1995). Major-plate motion between Australia and the Pacific Plate is only ~80 mm/year, the remainder being accommodated by spreading in the Lau back-arc system. Subduction can in principle be arbitrarily slow, although the lowest rates associated with an active arc probably occur in the Lesser Antilles (20–25 mm/year convergence; DeMets, 2001). A wide range of convergence rates lie in between these end-members with a sharp global peak near 70 mm/year such that half of all subduction zones have arc-normal convergence rates of 53–76 mm/year (Figure 3).

While spreading at mid-ocean ridges is orthogonal to their axes, at subduction zones convergence commonly trends obliquely to trench or arc (e.g., McCaffrey, 1992; Fitch, 1972). The obliquity, ϕ, is the angle between the normal to trench axis and the plate convergence direction. At low obliquities, convergence occurs in the direction of plate motion but commonly at higher obliquities strain is partitioned between a trench-normal component, represented by thrust faulting on the megathrust, and an along-strike component taken up by internal deformation. The along-strike component may be taken up within the forearc, as arc-parallel strike-slip systems, as sequences of strike-slip faults oblique to the trench axis, or by other means. At many subduction zones the obliquity changes along strike as both plate orientation and location relative to Euler pole change; for example, convergence is trench-normal ($\phi \sim 0$) in the Alaska Peninsula segment of the Aleutian subduction zone, $\phi \sim 45°$ from the trench in the central Aleutians, and plate motion is essentially parallel to the trench ($\phi \sim 90°$) in the far western Aleutians (DeMets et al., 1990). The latter case implies that material no longer descends into the mantle in the far western Aleutians, but translates along strike. A primary consequence of obliquity is that the trench-normal component of convergence can be significantly less than plate motion rates.



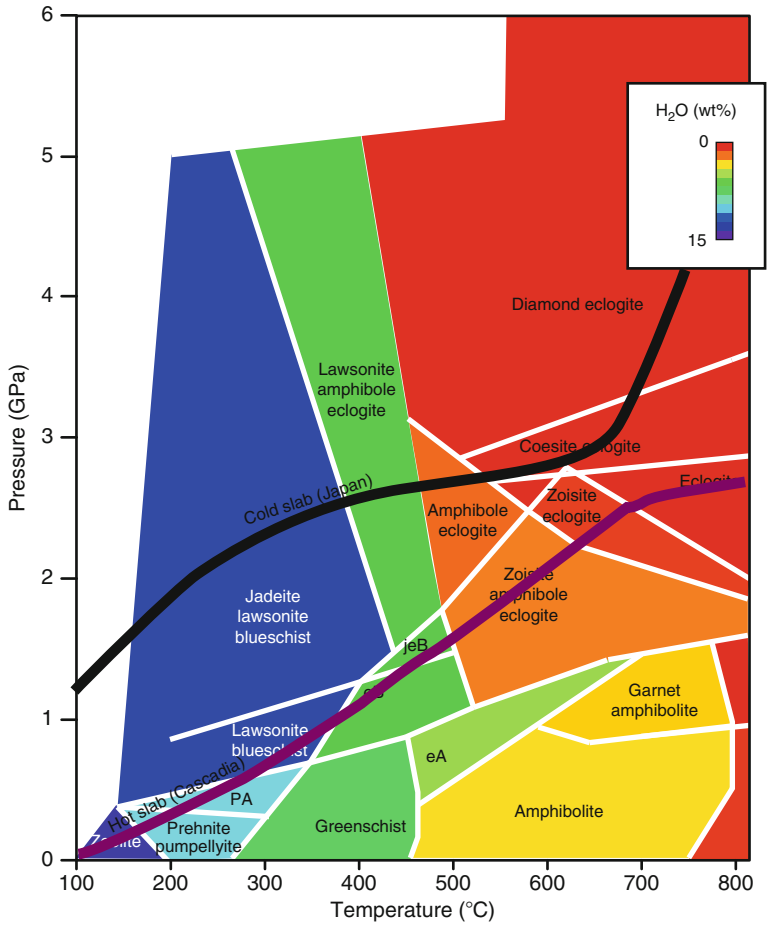
Subduction Zones, Figure 3 Histograms showing common parameters at subduction zones. Each is calculated at arc volcanoes or the nearest point on the plate boundary to them. (Based on data from Syracuse and Abers 2006.)

The kinematics of plates at asthenospheric depths may be more complex. While it is convenient to draw flow lines of subducting plates parallel to their dips, assuming steady-state slab geometry, it is not obvious that is the case. In many geodynamic models (e.g., Billen, 2008) slabs sink vertically faster than they would if they followed their trajectories, in other words, slabs roll back. Such rollback is probably required to accommodate back-arc opening in many places, although few if any constraints exist on these kinematics. Flow kinematics are complicated by inferences of large along-strike flow inferred from seismic anisotropy (e.g., Russo and Silver, 1994; Hoernle et al., 2008; see chapter on *Seismic Anisotropy*), not easily explained by simple models of wedge flow.

Thermal structure, plate buoyancy, mantle flow

At subduction zones, cold oceanic lithosphere advects downward relatively quickly compared with the rate of heat conduction, leading to some of the largest lateral temperature gradients in the planet. Old, fast-subducting slabs can retain temperatures $<600°C$ in their cores to depths exceeding 100 km, where ambient mantle temperatures are otherwise $\sim$1,400°C (e.g., Syracuse et al., 2010; Wada and Wang, 2009). The low temperatures lead to relatively large density differences, giving oceanic lithosphere significant negative buoyancy and probably supplying one of the major forces that drive plate motions (e.g., Forsyth and Uyeda, 1975). Because temperatures are low, mineral phases that hold significant H_2O are stable to much greater depths than elsewhere, making subduction zones major conveyors of volatiles to the deep earth. Still, volcanism occurs in nearly all oceanic subduction zones, indicating that parts of the system remain warm enough to produce melting. As a slab descends, it induces corner flow in the overlying mantle wedge, which draws warm material from the distal back-arc toward the slab corner (Figure 4). This advection of warm material ultimately heats the top of the slab, leading to very strong temperature gradients near the slab surface, and driving critical material-transport processes there.

Where the slab surface lies at depths less than 50 km, plate-boundary motion is mostly localized along the megathrust that defines the plate boundary, so the



Subduction Zones, Figure 4 Pressure-temperature trajectories for the top of slabs in hot (Cascadia) and cold (Japan) subduction zones, and metamorphic facies predicted for hydrated metabasalts, color coded by H_2O content. (Facies from Hacker et al. 2003; trajectories from Syracuse et al., 2010.)

upper-plate deformation accommodates little of the plate motion (at least the downdip component; oblique slip partitions strike-parallel motion into the upper plate). As a consequence, heat transport in this region can be approximated as the balance between conduction through an overlying, stationary upper plate, and downward advection of cold material along the thrust zone (Molnar and England, 1990). Shear heating along the thrust zone may elevate temperatures but its effect seems to be minor in the few places where abundant heat-flow measurements have been made (Cascadia, Japan, Tonga), equivalent to that generated by a fault with shear stress of a few tens of MPa at most (Hyndman and Wang, 1995; von Herzen et al., 2001). Thus, interplate thrust faults are relatively weak, perhaps due to high fluid pressures. One consequence is that metamorphism of subducted material follows a low-temperature high-pressure trajectory through this region, to blueschist facies (Figure 4).

At greater depth, the overlying material flows and transports heat from the back-arc mantle into the wedge. As this hot material is entrained downward by viscous drag, it rapidly heats the top of the downgoing plate, resulting in dehydration of the top of the downgoing plate. Such heating must happen under or trenchward of the volcanic arc, since many arcs erupt basaltic lavas with chemistry that requires temperatures in excess of 1,200°C beneath the arc (Kelemen et al., 2003a). The temperatures predicted for the top of the slab depend upon the assumed rheology of the mantle wedge, but for power-law flows typical of olivine dislocation creep, a narrow (<25–50 km wide) thermal boundary layer should rapidly develop at the top of the slab, leading to slab-surface temperatures of 700–1,000°C directly beneath the volcanic front (Kelemen et al., 2003a; van Keken et al., 2002; Syracuse et al., 2010). This heating appears to be sufficient to drive off most mineral-bound volatiles in the altered oceanic crust and sediment. More complicated modeling that allows secondary flow of buoyant slab-surface material into the wedge (e.g., Gerya and Yuen, 2003) might lead to lower temperatures, if such flows are sustainable.

Heat supplied by the mantle wedge continues to conduct into the slab, gradually warming its interior. Simple

thermal models (e.g., McKenzie, 1969; Molnar et al., 1979) predict that the maximum depth of isotherms should scale with $\Phi = AV\sin\delta$, where A is the age of the incoming lithosphere, V is its downdip velocity, and δ is the slab dip. The quantity Φ, sometimes called the "thermal parameter" (e.g., Kirby et al., 1996), describes in a simple way the overall thermal structure of slabs at great depth, and correlates well with the maximum depth of earthquakes over some depth ranges (Gorbatov and Kostoglodov, 1997), indicating a thermal control for the processes that cause deep earthquakes. Ultimately, the negative buoyancy of cold slabs and their ability to descend into the lower mantle must depend upon Φ as well. The negative thermal buoyancy of the slab counterbalances the positive buoyancy of subducting crust, since gabbro is much less dense than peridotite, such that the net buoyancy is a function of plate age and thickness of subducting crust (Cloos, 1993). For normal (5–8 km) thickness of oceanic crust, lithosphere attains negative buoyancy if more than 10–30 Ma old, while crust thicker than 15–25 km will always resist subduction for any observed age of oceanic crust.

The transition from localized slip along a plate boundary to distributed flow appears to correspond to where the slab reaches 70–90 km depth in many settings (Wada and Wang, 2009; Abers et al., 2006; Syracuse et al., 2010). Evidence for this transition comes from heat-flow observations at a small number of arcs (northern Japan and Cascadia), which show a step from very low forearc heat flow to relatively high sub-arc and back-arc heat flow, at this point. Also, seismic attenuation shows a sharp step at this depth, in a manner consistent with a cold "nose" of mantle material overlying the shallower portion of the slab, transitioning rapidly to hot mantle beneath the arc. The cause of this transition likely has to do with the maximum depth at which plate motion can be accommodated by a localized shear zone, and hence rheology of that shear zone; the location of arcs relative to the slab indicates a transition at temperatures about 550°C (Conder, 2005; Syracuse et al., 2010).

Fate of the downgoing plate: devolatilization and metamorphism

The subducting oceanic lithosphere can be divided into several layers (Figure 2, inset):

(a) Deep-sea sediment, of widely variable thickness and composition,
(b) Variably altered oceanic crust (AOC) – largely the upper, basaltic part where hydrothermal circulation has the largest effect,
(c) Less altered gabbroic lower crust, and
(d) Subducting mantle lithosphere some of which may be hydrated to varying degrees.

Although some of this package will be scraped off in accretionary margins, much probably survives to subduct to sub-arc depths. Arc geochemistry includes signatures of subducted sediment, with lava chemistry that varies systematically in some elements in ways that correlate with variations in sediment chemistry (Plank and Langmuir, 1998). This observation requires that most of the sediment, especially that deposited away from the trench, subduct to sub-arc depths. For example, lavas at several arcs include trace amounts of ^{10}Be, a radionuclide with ~1.6 Ma half-life formed in the atmosphere and deposited with the shallowest sediment (Tera et al., 1986), indicating that the shallowest deep-marine sediment subducts intact and the material reaches the earth's surface within a few Ma (a couple of half-lives). Finally, seismological evidence also supports the deep subduction of oceanic crust relatively intact, to generate images seen in receiver functions (e.g., Rondenay et al., 2008; Kawakatsu and Watada, 2007), high-frequency guided waves (Abers, 2000), and other mode conversions (e.g., Helffrich, 1996).

In its entirety, this package subducts on the order 2×10^9 Tg/Ma of water at the trench (Hacker, 2008), about half of which enters as pore water that is expelled by shallow compaction, and the remainder bound in minerals. Numerous drilling and sampling studies constrain well the H_2O content (and CO_2) of the subducting sediment and AOC, but lower sections are relatively less well known. Perhaps the largest uncertainty in these estimates lies in the water content of subducted mantle lithosphere, a potentially vast reservoir. Peridotite will hydrate to serpentinite and serpentine can hold ~14 wt% H_2O, giving this very thick layer potential to dominate the global water budget. However, it is not yet known how much H_2O reaches mantle depths; probably mantle hydration at mid-ocean ridges is minor because near-ridge mantle temperatures there should exceed that of serpentine stability (~600°C) but later faulting such as at fracture zones and the outer rise (see Thermal Structure, Plate Buoyancy, Mantle Flow) could be major conduits. Less is known about concentrations of CO_2 and other volatiles degassed by metamorphism. Depending upon serpentinite contributions, the volatile outputs of volcanic arcs either roughly match the inputs, implying near-complete volatile recycling, or are significantly less, implying a long-term flux of H_2O from the surface ocean to deep mantle.

Anhydrous oceanic crust can be expected to undergo a series of phase transitions from gabbro to eclogite. These result in a major density increase; gabbro has a density ~10% lower than that of peridotite at similar temperatures and pressures, so unmetamorphosed oceanic crust has buoyancy that resists subduction, while eclogite (dominated by garnet and pyroxenes) is significantly denser than peridotite. The density increase, in transforming from gabbro to eclogite, may contribute significantly to the negative buoyancy of slabs that drives plate tectonics. However, reaction rates for garnet growth (a major step in eclogite formation) within coarse-grained, anhydrous gabbro may be relatively slow, such that gabbro may persist metastably to considerable depth within subduction zones (Ahrens and Schubert, 1975). The presence of water should substantially enhance that reaction rate (e.g., Hacker, 1996).

If hydrated, basalt/gabbro (and metasediment) are expected to undergo a lengthy sequence of metamorphic

reactions that, generally, release H_2O (Figure 4). As a consequence, H_2O should be released over a wide range of depths from the thrust zone to sub-arc mantle, hydrating the overlying mantle wedge (Schmidt and Poli, 1998; Hacker, 2008). At typical to cold subduction conditions, oceanic crust can be expected to transition through greenschist and blueschist facies, capable of holding up to ~5 wt% H_2O, with major dehydration occurring in the transformation to eclogite during high-pressure heating (Hacker et al., 2003). In warmer subduction zones, defined as those with fairly young incoming plates and slow subduction rates (e.g., Cascadia, southern Japan), descending crust should follow a higher-temperature path perhaps through amphibolite facies, dehydrating as amphibole breaks down at ~2 GPa. Most thermal models with realistic mantle-wedge viscosity structure predict that subducting crust should dehydrate by depths of 100–200 km (e.g., Syracuse et al., 2010), at least in its shallow parts.

The mantle within the subducting plate potentially holds hydrous minerals as well, primarily serpentine but also chlorite, as might entrained material lying above the downgoing plate. These layers should also dehydrate, at temperatures of 600–700°C for serpentines at <200 km depth (e.g., Schmidt and Poli, 1998) with chlorite breaking down at temperatures about 200° higher. Serpentine dehydration temperatures are probably reached in all subduction zones on the slab surface and, hence, above it, although it is possible that a weakly hydrated zone of chlorite stability may persist past slab depths (e.g., Grove et al., 2009). Subducting mantle lithosphere in most subduction zones should be cold enough for serpentine to be stable past the arc (Syracuse et al., 2010), and may dehydrate gradually past that. Water persisting to depths past 200 km has the potential to remain within the mantle to considerable depth, in the "alphabet" phases of dense hydrous magnesium silicates, which have broad stability fields at near-slab conditions (Thompson, 1992). Water reaching the transition zone (410–660 km depth) may remain for a long time, as the dominant minerals ringwoodite and wadsleyite can retain wt% water as defects in their nominally anhydrous structure (Kohlstedt et al., 1996). This reservoir has potential to hold several times more water than the current surface ocean; thus, the ability of serpentinized mantle lithosphere to transport H_2O to great depths may be a dominant control on the Earth's long-term water budget.

As water is released, it ascends into the overlying mantle wedge, crust, back along the thrust zone, and through the forearc (Peacock, 1990), potentially hydrating it (Hyndman and Peacock, 2003). These pathways provide major controls on rheological properties in all these regions and on the behavior and composition of arc volcanoes. Also, it is often suggested that the dehydration process generates earthquakes at intermediate depths (70–150 km) in subduction zones (Green and Houston, 1995; Kirby et al., 1996; Hacker et al., 2003). These processes are discussed extensively in the chapter on seismicity-subduction zones.

Subduction channel

In many collisional mountain belts and accreted terranes, high-pressure, low-temperature rocks have reached the Earth's surface after having descended to great depths. In some "ultra-high pressure" (UHP) terranes, mineral assemblages or textures in these rocks record pressures in excess of 3–4 GPa (depths > 100 km), indicating that these rocks descended deeply in subduction zones before returning to the surface. In part to explain these observations, a "subduction channel" has been often proposed (Cloos and Shreve, 1988). The subduction of weak, hydrated metasediments and altered oceanic crust, and their subsequent shearing and interaction with metamorphic fluids, suggests the presence of a heterogeneous weak zone at the top of the downgoing plate. Initially, material in this channel will be dragged downward by coupling with the subducting plate, but it should be buoyant relative to overlying mantle, and in the right conditions a return flow could develop delivering high-grade metamorphic rocks to the surface. A weak zone is also indicated by the presence of a "cold nose" in many subduction zones in the forearc mantle wedge, a feature that probably requires a weak decoupling layer to depths of 70–90 km on the top of the downgoing plate (e.g., Wada and Wang, 2009). The presence of UHP rocks at the earth's surface indicates that the return-flow process or something similar probably happens in some instances, but it remains unclear if the subduction channel is a common feature of subduction zones.

Melting and the volcanic arc

Reviews elsewhere describe the geochemical evidence for the nature of melting beneath volcanic arcs and the generation of arc crust (e.g., Stern, 2002; Kelemen et al., 2003b), which briefly summarized here for their relevance to geophysical observations. Magmas emplaced and erupted at subduction zones span a wide range of compositions. Basalts are found at most subduction zones and indicate that primary melting occurs within the mantle beneath the volcanic arc, since basalts primarily form from melting of peridotites (e.g., Gill, 1981). These basalts often differ from those found at mid-ocean ridges or hot spots in several important chemical trends; in particular, many elements that are strongly abundant in subducting sediment and altered crust are found in abundance in many volcanic arc rocks in a manner that correlates with sediment chemistry (Plank and Langmuir, 1998). In some arcs there is strong evidence for andesites forming under high pressure, perhaps requiring melting of downgoing crust (e.g., Kelemen et al., 2003b).

Magmatic differentiation within the crust and mixing with preexisting arc basement result in a wide spectrum of volcanic output. Typically, seismic velocities within arcs show an increase with depth consistent with more silicic, felsic rocks at shallow depths (Vp = 6.0–6.4 km/s) underlain by more gabbroic rocks in the middle-to-lower crust (Vp = 7.0–7.4 km/s) (e.g., Suyehiro et al., 1996;

Shillington et al., 2004). Since they are petrologically fractionated from more basaltic parents, the andesitic or granitic rocks observed at mid-to-upper crustal levels should be balanced by more mafic silica-poor residuals, perhaps in the lower crust. However, such lower-crustal cumulates are not very abundant, as interpreted from seismic velocities, exposed arc sections, and within continents formed by arc accretion. In general, there is good reason to think that continents grew from accretion of arcs since they share many geochemical characteristics (Rudnick, 1995), but overall continents are significantly more silicic than primary arc basalts. This discrepancy probably requires a significant mass loss in the complementary, mafic cumulates that should form in arc lower crust. Where and how this mass loss occurs remains unclear; it may be that the lower crust in arcs occasionally founders or delaminates back into the mantle, since its density may be higher than that of peridotite, but many other possibilities exist. Geophysical observations of the lower crust and upper mantle have potential to place constraints on the formation and evolution of arc crust, and hence the origins of continental material (e.g., Kodaira et al., 2007; Tatsumi et al., 2008).

Primary magmas emerging from the mantle contain several wt% H_2O and elevated CO_2, several times higher than in mid-ocean ridge basalts (e.g., Hirschmann, 2006). Mantle melting beneath arcs often occurs in the presence of H_2O (Grove et al., 2006), leading to the frequent speculation that dehydration reactions within the subducting slab trigger melting in the overlying mantle wedge, and hence the location of dehydration within the slab controls the location of arc volcanism (e.g., Tatsumi, 1986). However, experiments on hydrous metabasalts show that dehydration occurs over a wide range of depths beginning much shallower than the sub-arc mantle wedge (e.g., Schmidt and Poli, 1998), so that H_2O should flux the mantle at many depths, not just under the volcanic arc. Also, primary, dry basalts have been found in some arcs, so H_2O is not always required for melting. It may be that the return flow (the incoming limb of wedge corner flow) sufficiently decompresses hot asthenosphere to trigger melting. The relative importance of wet (flux) melting and dry (decompression) melting within arcs remains poorly understood, as do the pathways that magmas take from their locus of formation to volcanic arcs. In particular, it is not obvious why the arc front is a narrow, well-defined feature while the conditions for melting should be more widely spread. The depth to slabs beneath the arc front averages 110 km but varies coherently by a factor of 2 from arc to arc, suggesting some other controlling process probably related to thermal structure (England et al., 2004; Syracuse and Abers, 2006).

Seismic imaging of the subduction zone

Although subduction zones exhibit a characteristic surface morphology of trench, forearc, volcanic arc, etc., the evidence that these associations represent subduction of deep material comes primarily from geophysical imaging of deep structure. Early observations of seismic wave propagation in Tonga showed that the Wadati–Benioff zone lies in a cold, low-attenuation structure unlike mantle elsewhere on the planet, and provided some of the first evidence that these were sites of descending cold lithosphere (Oliver and Isacks, 1967). Since then, methods of seismic tomography have improved greatly (see *Seismic Tomography*) as has the global data set of earthquake travel times, leading to a clear picture of P-wave high-velocity anomalies that show cold slabs descending through the upper mantle and in some cases into the lower mantle (e.g., van der Hilst et al., 1997). These data show slabs as seismic anomalies 4–10% faster than surrounding mantle (see compilation by Lay, 1997), consistent with thermal models that predict temperatures 400–800°C colder than surroundings. These images, based on global earthquake travel time data sets, show that in some cases slabs reach the mantle transition zone and flatten out, as if reaching a partial barrier to flow (e.g., Japan) while in other cases they pass through the transition zone delivering subducted material to the lower mantle.

Seismic recordings from networks and arrays in subduction zones provide strong constraints on the subduction zone structure of a scale relevant to slab metamorphism, arc petrogenesis, and wedge flow. For example, as dehydration occurs, densities increase as do seismic velocities; basalts are 10–15% slower than surrounding peridotites while eclogites have similar velocities. Furthermore, serpentinized peridotite can be tens of percent slower than peridotite particularly for shear waves, depending upon the extent of serpentinization (hydration). As a result, seismic imaging has potential to show where and how metamorphism within slabs occurs. For example, with migration of P-to-S scattered waves (see *Seismic, Receiver Function Technique*) subducting crust has been observed to ~45 km depth in Cascadia and depths exceeding 100 km in Alaska and Japan, consistent with expected depths of basalt dehydration and differences in thermal structure (Rondenay et al., 2008). Studies of earthquake travel times and signal attenuation also provide information on structure of the mantle wedge where melting initiates, and have shown a clear hot mantle wedge, with a variety of structures suggesting melt transport pathways. Probably the densest long-lived monitoring effort exists in Japan, although a variety of portable array experiments have illuminated subduction zones in parts of the Andes, Central America, Tonga, the Marianas, Alaska, Cascadia, Greece, and other locales.

One of the more complex suites of seismic observations comes from the anisotropic fabric within the mantle of subduction zones, which results in "splitting" of shear waves into orthogonally polarized fast and slow pulses (see *Seismic Anisotropy*). Mantle flow in the dislocation creep regime leads to lattice-preferred orientation of olivine, and olivine is moderately (~10%) anisotropic, leading to an association between seismic anisotropy and mantle flow. However, observations in many subduction zones lie counter to expectation (e.g., Russo and Silver, 1994).

Simple models of mantle flow in subduction zones predict "corner flow," in which mantle on either side of the descending plate is entrained via viscous drag (Figure 2). To fill the potential void left by the downward advection, mantle far from the subducting slab flows inward toward it, eventually turning and flowing downward. While details of this flow depend upon the viscosity structure and its sensitivity to temperature, stress, water, and other variables, the overall pattern of flow is one in which the maximum strain or stretch should be perpendicular to trench and arc, in the direction of flow. In many subduction zones, anisotropic fabric does not show this pattern, and instead the fast direction for shear wave propagation is more often than not parallel to the trench, at least below and behind the volcanic arc (Fischer et al., 1998). More complicated fabrics are also observed, with sharp 90° changes in fabric between sub-arc mantle and the shallower part of the forearc (e.g., in North Japan and Alaska), and the pattern varies between different subduction zones. Explanations for these unexpected patterns remain a matter of much debate, and include strong along-strike flow, secondary flows within the wedge, influence of melt on fabric, activation of unusual olivine slip systems at some subduction zone conditions, abundance of hydrous, foliated minerals in the forearc, extensive fabric within the downgoing slab, and unexpectedly strong upper-plate fabrics.

Subduction zone hazards

The forces present in subduction zones also produce some of the most violent natural hazards known. The interplate thrust zone generates all of the largest earthquakes, which produce strong shaking but also submarine landslides, meter-scale uplift and subsidence, and most of the largest tsunamis. The 2004 Sumatra earthquake represents the primary example of the last 50 years; a magnitude 9.2 earthquake ruptured over 1,000 km of plate boundary, producing an ocean-crossing tsunami that directly resulted in over 150,000 deaths and huge devastation to property. These earthquakes are devastating because subduction zone thrust faults are very large, because they are submarine but can break the sea surface, and because they occur in places with often oversteepened bathymetry. Subduction zones are also site to many of the most violent volcanic eruptions. Large volatile (especially water) contents of subduction zone magmas, compared with most other settings, lead to magmas rich in gas that exsolve at shallow depths. Magmatic differentiation leads to relatively viscous, felsic magmas that rise slowly and do not degas easily. The combination leads to explosive and sometimes very large eruptions; in the last half century, Mt St Helens (Cascadia), Pinatubo (Philippines), and Montserrat (Lesser Antilles) all produced violent and well-documented eruptions in subduction zones, among many others. Worldwide, about 500 volcanoes have known historic eruptions, roughly two-thirds in subduction zones.

Summary

Subduction zones represent the primary return of material from the earth's surface to its interior. In the process of doing so, great earthquakes occur, violent volcanoes erupt, and the largest lateral temperature gradients in the Earth's interior are generated. At shallow depths, elastic and brittle processes control deformation including most of the planet's large earthquakes, while at greater depths the subduction system deforms by large-scale ductile flow. Along with oceanic crust, sediment and volatiles enter the trench, lubricating the plate boundary at shallow depths, and then are metamorphosed where slabs descend below hot mantle, releasing H_2O and other volatiles into the overlying mantle wedge. While some of this material may mechanically ascend back to the surface, in the form of high-pressure metamorphic rocks, much feeds explosive volcanism in the volcanic arc or descends into the deep mantle. Melting at sub-arc depths is controlled by a combination of advection via flow of hot mantle from the back-arc, and fluxing by slab-derived fluids, a combination of which produces a sharp volcanic front. Magmas ascend to form arc crust and differentiate, some fraction of which persists to form the basis of continental crust. This transfer of energy and material leads to a rich set of phenomena that play a major role in creating the geologic record.

Bibliography

Abers, G. A., 2000. Hydrated subducted crust at 100–250 km depth. *Earth and Planetary Science Letters*, **176**, 323–330.

Abers, G. A., van Keken, P. E., Kneller, E. A., Ferris, A., and Stachnik, J. C., 2006. The thermal structure of subduction zones constrained by seismic imaging: implications for slab dehydration and wedge flow. *Earth and Planetary Science Letters*, **241**, 387–397.

Ahrens, T. J., and Schubert, G., 1975. Gabbro-eclogite reaction rate and its geophysical significance. *Reviews of Geophysics and Space Physics*, **13**, 383–400.

Bevis, M., et al., 1995. Geodetic observations of very rapid convergence and back-arc extension at the Tonga arc. *Nature*, **374**, 249–251.

Billen, M. I., 2008. Modeling the dynamics of subducting slabs. *Annual Review of Earth and Planetary Sciences*, **36**, 325–356.

Chapple, W. M., and Forsyth, D. W., 1979. Earthquakes and bending of plates at trenches. *Journal of Geophysical Research*, **84**, 6729–6749.

Cloos, M., 1993. Lithospheric buoyancy and collisional orogenesis; subduction of oceanic plateaus, continental margins, island arcs, spreading ridges, and seamounts. *Geological Society of America Bulletin*, **105**, 715–737.

Cloos, M., and Shreve, R. L., 1988. Subduction-channel model of prism accretion, melange formation, sediment subduction, and subduction erosion at convergent plate margins: 1. Background and description. *Pure and Applied Geophysics*, **128**, 455–500.

Conder, J. A., 2005. A case for hot slab surface temperatures in numerical viscous flow models of subduction zones with an improved fault zone parameterization. *Physics of the Earth and Planetary Interiors*, **149**, 155–164.

Davis, D., Suppe, J., and Dahlen, F. A., 1983. Mechanics of fold-and-thrust belts and accretionary wedges. *Journal of Geophysical Research*, **88**, 1153–1172.

DeMets, C., 2001. A new estimate for present-day Cocos-Caribbean plate motion; implications for slip along the central American volcanic arc. *Geophysical Research Letters*, **28**, 4043–4046.
DeMets, C., Gordon, R., Argus, D., and Stein, S., 1990. Current plate motions. *Geophysical Journal International*, **101**, 425–478.
England, P., Engdahl, E. R., and Thatcher, W., 2004. Systematic variations in the depths of slabs beneath arc volcanoes. *Geophysical Journal International*, **156**, 377–408.
Fischer, K. M., Fouch, M. J., Wiens, D. A., and Boettcher, M. S., 1998. Anisotropy and flow in Pacific subduction zone back-arcs. *Pure and Applied Geophysics*, **151**, 463–475.
Fitch, T. J., 1972. Plate convergence, transcurrent faults and internal deformation adjacent to southeast Asia and the western Pacific. *Journal of Geophysical Research*, **77**, 4432–4460.
Forsyth, D. W., and Uyeda, S., 1975. Relative importance of driving forces of plate motion. *Geophysical Journal of the Royal Astronomical Society*, **43**, 163–200.
Gerya, T. V., and Yuen, D. A., 2003. Rayleigh-Taylor instabilities from hydration and melting propel 'cold plumes' at subduction zones. *Earth and Planetary Science Letters*, **212**, 47–62.
Gill, J. B., 1981. *Orogenic andesites and plate tectonics*. New York: Springer, p. 390.
Gomberg, J., 2010. Slow-slip phenomena in Cascadia from 2007 and beyond: a review. *Geological Society of America Bulletin*, **122**, 963–978.
Gorbatov, A., and Kostoglodov, V., 1997. Maximum depth of seismicity and thermal parameter of the subducting slab: general empirical relation and its application. *Tectonophysics*, **277**, 165–187.
Green, H. W., and Houston, H., 1995. The mechanics of deep earthquakes. *Annual Review of Earth and Planetary Sciences*, **23**, 169–213.
Grove, T. L., Chatterjee, N., Parman, S. W., and Médard, E., 2006. The influence of H_2O on mantle wedge melting. *Earth and Planetary Science Letters*, **249**, 74–89.
Grove, T. L., Till, C. B., Lev, E., Chatterjee, N., and Medard, E., 2009. Kinematic variables and water transport control the formation and location of arc volcanoes. *Nature*, **459**, 694–697.
Hacker, B. R., 1996. Eclogite formation and the rheology, buoyancy, seismicity, and H_2O content of oceanic crust. In Bebout, G. E., Scholl, D., Kirby, S., and Platt, J. P. (eds.), *Subduction: Top to Bottom*. Washington, DC: American Geophysical Union. AGU Monograph Series, Vol. 96, pp. 337–346.
Hacker, B. R., 2008. H_2O subduction beyond arcs. *Geochemistry Geophysics Geosystems* **9**: art. No. Q03001.
Hacker, B. R., Abers, G. A., and Peacock S. M., 2003. Subduction factory 1: theoretical mineralogy, density, seismic wavespeeds, and H_2O content. Journal of Geophysical Research, **108**: art. No. 2029.
Helffrich, G., 1996. Subducted lithospheric slab velocity structure: observations and mineralogical inferences. In Bebout, G. E., Scholl, D., Kirby, S., and Platt, J. P. (eds.), *Subduction: Top to Bottom*. Washington, DC: American Geophysical Union. AGU Monograph Series, Vol. 96, pp. 215–222.
Hilde, T. W. C., 1983. Sediment subduction versus accretion around the Pacific: convergence and subduction. *Tectonophysics*, **99**, 381–397.
Hirschmann, M. M., 2006. Water, melting, and the deep earth H_2O cycle. *Annual Review of Earth and Planetary Sciences*, **34**, 629–653.
Hoernle, K., Abt, D. L., Fischer, K. M., Nichols, H., Hauff, F., Abers, G. A., van den Bogaard, P., Heydolph, K., Alvarado, G., Protti, M., and Strauch, W., 2008. Arc-parallel flow in the mantle wedge beneath Costa Rica and Nicaragua. *Nature*, **451**, 1094–1098.
Hyndman, R. D., and Peacock, S. M., 2003. Serpentinization of the forearc mantle. *Earth and Planetary Science Letters*, **212**, 417–432.
Hyndman, R. D., and Wang, K., 1995. The rupture zone of Cascadia great earthquakes from current deformation and the thermal regime. *Journal of Geophysical Research*, **100**, 22133–22154.
Hyndman, R. D., Yamano, M., and Oleskevich, D. A., 1997. The seismogenic zone of subduction thrust faults. *Island Arc*, **6**, 244–260.
Ide, S., Beroza, G., Shelly, D. R., and Uchide, T., 2007. A scaling law for slow earthquakes. *Nature*, **447**, 76–79.
Ito, Y., Obara, K., Shiomi, K., Sekine, S., and Hirose, H., 2007. Slow earthquakes coincident with episodic tremors and slow slip events. *Science*, **315**, 503–506.
Kawakatsu, H., and Watada, S., 2007. Seismic evidence for deep-water transportation in the mantle. *Science*, **316**, 1468–1471.
Kelemen, P. B., Rilling, J. L., Parmentier, E. M., Mehl, L., and Hacker, B. R., 2003a. Thermal structure due to solid-state flow in the mantle wedge beneath arcs. In Eiler, J. M. (ed.), *Inside the Subduction Factory*. Washington, DC: AGU. Geophysical Monograph, Vol. 138, pp. 293–311.
Kelemen, P. B., Hanghøj, K., and Greene, A. R., 2003b. One view of the geochemistry of subduction related magmatic arcs, with emphasis on primitive andesite and lower crust. In Rudnick, R. L. (ed.), *The Crust*. Oxford: Elsevier. Treatise on Geochemistry, Vol. 3.
Kirby, S., Engdahl, E. R., and Denlinger, R., 1996. Intermediate-depth intraslab earthquakes and arc volcanism as physical expressions of crustal and uppermost mantle metamorphism in subducting slabs. In Bebout, G. E., Scholl, D., and Kirby, S. (eds.), *Subduction: Top to Bottom*. Washington, DC: Amererican Geophysical Union. Geophysical Monograph, Vol. 96, pp. 195–214.
Kodaira, S., Sato, T., Takahashi, N., Miura, S., Tamura, Y., Tatsumi, Y., and Kaneda, Y., 2007. New seismological constraints on growth of continental crust in the Izu-Bonin intra-oceanic arc. *Geology*, **35**, 1031–1034.
Kohlstedt, D. L., Keppler, H., and Rubie, D. C., 1996. Solubility of water in the alpha, beta and gamma phases of (Mg, Fe)(2)SiO_4. *Contributions to Mineralogy and Petrology*, **123**, 345–357.
Lallemand, S., Hueuret, A., and Boutelier, D., 2005. On the relationships between slab dip, back-arc stress, upper plate absolute motion, and crustal nature in subduction zones. *Geochemistry Geophysics Geosystems*, **6**: art. no. Q090006.
Lay, T., 1997. *Structure and Fate of Subducting Slabs*. New York: Academic, p. 185.
McKenzie, D. P., 1969. Speculations on the consequences and causes of plate motions. *Geophysical Journal of the Royal Astronomical Society*, **18**, 1–32.
McCaffrey, R., 1992. Oblique plate convergence, slip vectors, and forearc deformation. *Journal of Geophysical Research*, **97**, 8905–8915.
Molnar, P., and England, P. C., 1990. Temperatures, heat flux, and frictional stress near major thrust faults. *Journal of Geophysical Research*, **95**, 4833–4856.
Molnar, P., Freedman, D., and Shih, J. S. F., 1979. Lengths of intermediate and deep seismic zones and temperatures in downgoing slabs of lithosphere. *Geophysical Journal of the Royal Astronomical Society*, **56**, 41–54.
Obara, K., 2002. Nonvolcanic deep tremor associated with subduction in southwest Japan. *Science*, **296**, 1679–1681.
Oliver, J., and Isacks, B., 1967. Deep earthquake zones, anomalous structures in the upper mantle, and the lithosphere. *Journal of Geophysical Research*, **72**, 4259–4275.
Parsons, B., and Molnar, P., 1976. The origin of outer topographic rises associated with trenches. *Geophysical Journal of the Royal Astronomical Society*, **45**, 707–712.
Peacock, S. M., 1990. Fluid processes in subduction zones. *Science*, **248**, 329–337.
Plank, T., and Langmuir, C. H., 1998. The chemical composition of subducting sediment and its consequences for the crust and mantle. *Chemical Geology*, **145**, 325–394.

Ranero, C., Morgan, J., McIntosh, K., and Reichert, C., 2003. Bending-related faulting and mantle serpentinization at the middle America trench. *Nature*, **425**, 367–373.
Rogers, G., and Dragert, H., 2003. Episodic tremor and slip on the Cascadia subduction zone: the chatter of silent slip. *Science*, **300**, 1942–1943.
Rondenay, S., Abers, G. A., and van Keken, P. E., 2008. Seismic imaging of subduction zone metamorphism. *Geology*, **36**, 275–278.
Rudnick, R. L., 1995. Making continental crust. *Nature*, **378**, 571–578.
Rudnick, R. L., and Fountain, D. M., 1995. Nature and composition of the continental crust; a lower crustal perspective. *Reviews of Geophysics*, **33**, 267–309.
Russo, R. M., and Silver, P. G., 1994. Trench-parallel flow beneath the Nazca plate from seismic anisotropy. *Science*, **263**, 1105–1111.
Saffer, D. M., and Marone, C., 2003. Comparison of smectite- and illite-rich gouge frictional properties: application to the updip limit of the seismogenic zone along subduction megathrusts. *Earth and Planetary Science Letters*, **215**, 219–235.
Schmidt, M. W., and Poli, S., 1998. Experimentally based water budgets for dehydrating slabs and consequences for arc magma generation. *Earth and Planetary Science Letters*, **163**, 361–379.
Scholz, C. H., 1998. Earthquakes and friction laws. *Nature*, **391**, 37–42.
Schwartz, S. Y., and Rokosky, J. M., 2007. Slow slip events and seismic tremor at circum-pacific subduction zones. *Reviews of Geophysics*, **45**: art. no. RG3004.
Shillington, D., Van Avendonk, H. J. A, Holbrook, W. S., Kelemen, P. B., Hornbach, M. J., 2004. Composition and structure of the central Aleutian island arc from arc-parallel wide-angle seismic data. *Geochemistry Geophysics Geosystems*, **5**: Art no. Q10006.
Stern, R.J., 2002. Subduction zones. *Reviews of Geophysics*, **40**: art. no. 1012.
Suyehiro, K., Takahashi, N., Ariie, Y., Yokoi, Y., Hino, R., Shinohara, M., Kanazawa, T., Hirata, N., Tokuyama, H., and Taira, A., 1996. Continental crust, crustal underplating, and low-Q upper mantle beneath an oceanic island arc. *Science*, **272**, 390–392.
Syracuse, E. M., and Abers, G. A., 2006. Global compilation of variations in slab depth beneath arc volcanoes and implications. *Geochemistry Geophysics Geosystems*, **7**: art. no. Q05017.
Syracuse, E. M., van Keken P. E., and Abers G. A., 2010. The global range of subduction zone thermal models. *Physics of the Earth and Planetary Interiors*, **183**, 73–90.
Tatsumi, Y., 1986. Formation of the volcanic front in subduction zones. *Geophysical Research Letters*, **13**, 717–720.
Tatsumi, Y., Shukuno H., Tani K., Takahashi N., Kodaira S., and Kogiso, T., 2008. Structure and growth off the Izu-Bonin-Mariana arc crust: 2. Role of crust-mantle transformation and the transparent Moho in arc crust evolution. *Journal of Geophysical Research.***113**: art. no. B02203.
Tera, F., Brown, L., Morris, J., Sacks, I. S., Klein, J., and Middleton, R., 1986. Sediment incorporation in island-arc magmas: inferences from ^{10}Be. *Geochimica et Cosmochimica Acta*, **50**, 535–550.
Thompson, A. B., 1992. Water in the Earth's upper mantle. *Nature*, **358**, 295–302.
Tichelaar, B. W., and Ruff, L. J., 1993. Depth of seismic coupling along subduction zones. *Journal of Geophysical Research*, **98**, 2017–2037.
Van der Hilst, R. D., Widyantoro, S., and Engdahl, E. R., 1997. Evidence for deep mantle circulation from global tomography. *Nature*, **386**, 578–584.
van Keken, P. E., Kiefer, B., and Peacock, S. M., 2002. High-resolution models of subduction zones: implications for mineral dehydration reactions and the transport of water into the deep mantle. *Geochemistry Geophysics Geosystems*, **3**: art. no. 1056.
Von Herzen, R., Ruppel, C., Molnar, P., Nettles, M., Nagihara, S., and Ekstrom, G., 2001. A constraint on the shear stress at the Pacific-Australia plate boundary from heat flow and seismicity at the Kermadec forearc. *Journal of Geophysical Research*, **106**, 6817–6833.
von Huene, R., and Scholl, D. W., 1991. Observations at convergent margins concerning sediment subduction, subduction erosion, and the growth of continental crust. *Reviews of Geophysics*, **29**, 279–316.
Wada, I., and Wang, K., 2009. Common depth of slab-mantle decoupling: reconciling diversity and uniformity of subduction zones. *Geochemistry Geophysics Geosystems*, **10**: Q10009
Watts, A. B., and Talwani, M., 1974. Gravity anomalies seaward of deep-sea trenches and their tectonic implications. *Geophysical Journal of the Royal Astronomical Society*, **36**, 57–90.

Cross-references

Earthquake, Magnitude
Seismic Anisotropy
Seismic Tomography
Seismic, Receiver Function Technique
Tsunami

SURFACE WAVES

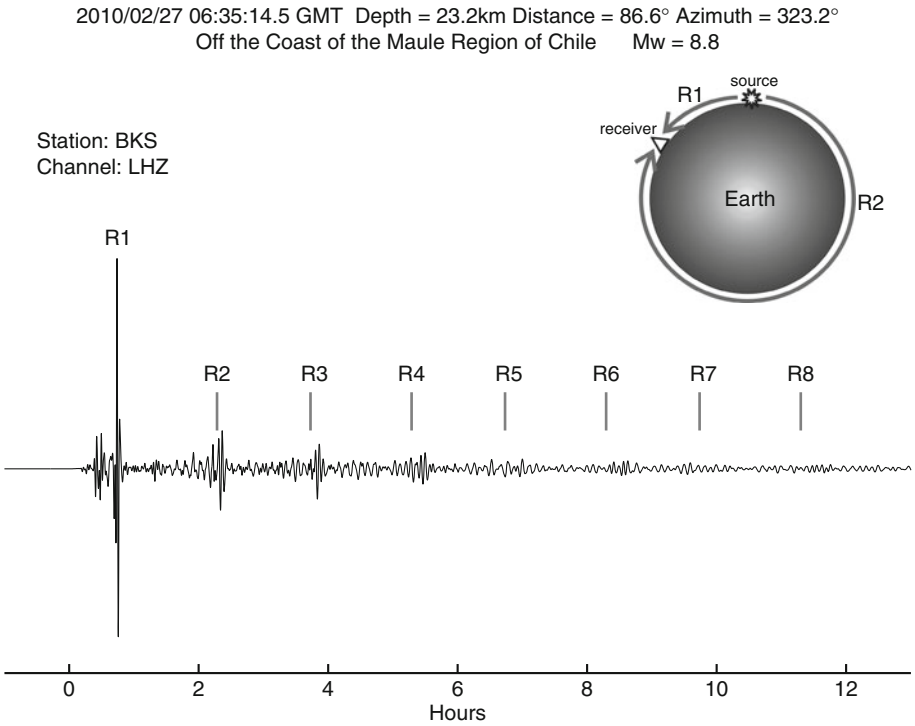
Barbara Romanowicz
Berkeley Seismological Laboratory, Berkeley, CA, USA

Definition

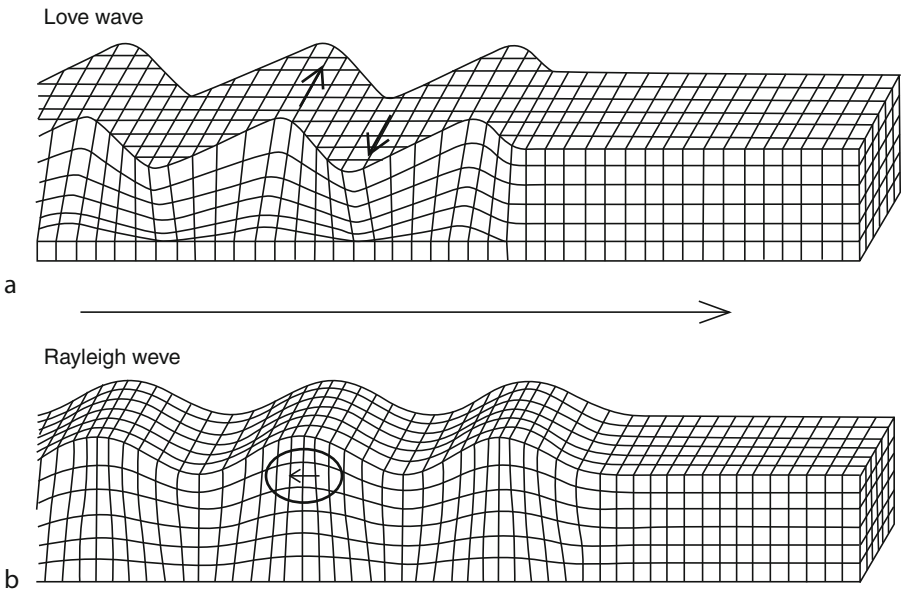
Surface waves are elastic waves which propagate along the surface of the earth and whose energy decays exponentially with depth.

Surface waves contain most of the long period energy (periods greater than 20 s) generated by earthquakes and recorded at teleseismic distances. Most prominent on records of moderate ($M \geq 5.5$) earthquakes are the fundamental mode wave trains that have propagated along the direct great circle path between the epicenter and the station. The dispersive and attenuative properties of these wave trains have been used extensively, since the 1950s, to infer crust and upper mantle structure at the regional scale. For earthquakes of magnitude 7 or larger, successive, earth-circling surface wave trains can be followed for many hours (Figure 1) and are either analyzed individually or, at the longest periods ($T > 250$ s) they are combined over time lengths of tens of hours or days to produce a spectrum of Earth's free oscillations (see *Free Oscillations of the Earth*).

In an isotropic medium, there are two types of surface waves which propagate independently (Figure 2). Rayleigh waves, polarized in the vertical plane containing the source and the receiver (P-SV energy), have a retrograde elliptical particle motion. These waves have generally good signal-to-noise ratio on the vertical component records and have been studied the most. On the other hand, the analysis of Love waves, polarized



Surface Waves, Figure 1 Example of vertical component record showing earth-circling Rayleigh wave trains (marked R1...R8) following the M8.8 Maule Chile earthquake of February 27, 2010. This record is from station BKS of the Berkeley Digital Seismic Network (BDSN). The successive wavepackets are well separated from each other at this epicentral distance (86.6°). The long period surface waves are often referred to as "mantle waves." (Courtesy of Shan Dou.)



Surface Waves, Figure 2 Particle motion for Rayleigh and Love waves. Rayleigh waves have retrograde particle motion confined to the vertical plane of propagation. Love waves have purely transverse motion in the horizontal plane. (After Bolt, B. A., 1976. *Nuclear Explosions and Earthquakes*. W. H. Freeman.)

horizontally in a direction perpendicular to the propagation direction (SH energy), suffers from the more complex data processing required, and from higher levels of long period background noise on the horizontal component records, due primarily to the effect of atmospheric pressure variations, inducing ground tilts. Love waves, unlike Rayleigh waves, can exist only in a heterogeneous medium (i.e., a layered medium).

The most striking property of surface waves is their dispersive character, with, in general, waves of longer periods traveling faster, directly reflecting the increase of elastic velocity with depth in the earth's crust and mantle. Indeed, studies of crustal and upper mantle structure progressed rapidly in the 1950s and early 1960s, as the tools developed to measure group and phase velocity dispersion and interpret them in terms of layered mantle and crust models (e.g., Ewing et al., 1957).

In a spherical earth, the theory of generation and propagation of surface waves is intimately linked to that of Earth's free oscillations, leading to the distinction of various modes. The lowest mode is termed fundamental mode, followed by higher modes, or overtones. Fundamental mode Love and Rayleigh waves are generally well separated from other seismic phases on the seismograms, and are well excited by shallow, crustal earthquakes, while higher modes (i.e., overtones) travel at higher group velocities and appear as compact packets of mixed overtones. They are better excited by deeper earthquakes (Figure 3).

Surface waves recorded at teleseismic distances contain information about both the characteristics of the earthquake source and the structure of the earth's crust and mantle along the source-station path. Separating the two effects has been one of the long-standing challenges faced by seismologists.

Group and phase velocity

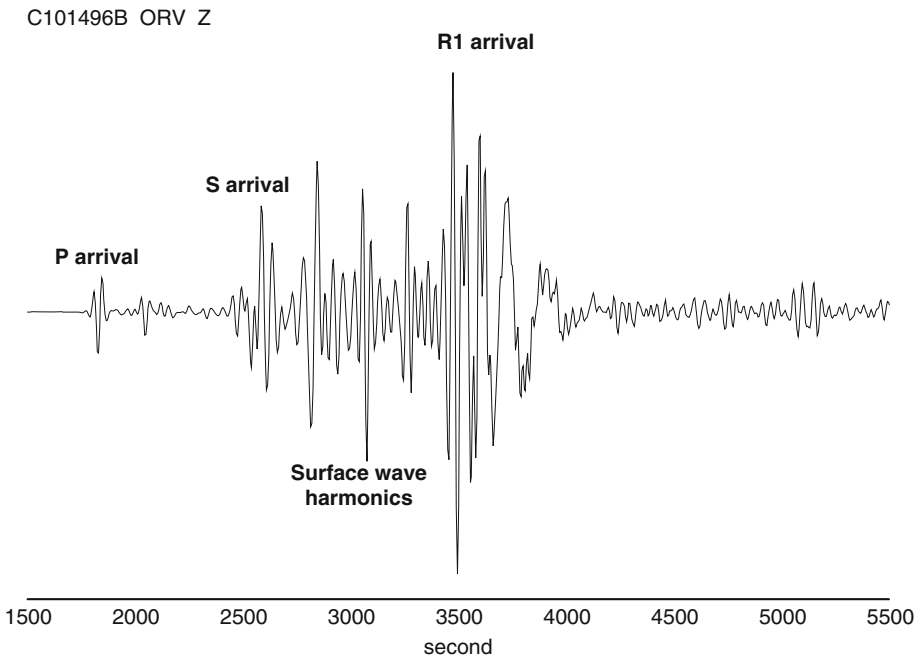
The velocity of propagation of surface waves can be described in terms of either group velocity or phase velocity. Phase velocity is the velocity of a specific frequency component of the wave train. Phase velocity derives from the expression of a plane wave of wavenumber k, and angular frequency ω, propagating in the direction x, for example, in one dimension:

$$f(x,t) \sim \exp i(\omega t - kx) \tag{1}$$

It is defined as the velocity of the phase of the component of frequency ω:

$$C(\omega) = \omega/k \tag{2}$$

where ω is the angular frequency and k is the wavenumber. Frequency is related to period of the wave, T, by $\omega = 2\pi/T$, while $k = 1/\lambda$ where λ is wavelength.



Surface Waves, Figure 3 Example of vertical component record for the October 14, 1996 Solomon Islands earthquake (depth = 19 km, *Mw* = 6.7) at BDSN station ORV showing the arrivals of multiply reflected body wave phases forming a higher-mode Rayleigh wave train in front of the fundamental mode (R1). The Airy phase, corresponding to the group velocity minimum around 230 s, is well visible at the end of the R1 train. The seismogram has been bandpass filtered with cutoff frequencies at 35 and 400 s. Courtesy of Yuancheng Gung.

Group velocity is the velocity with which the energy travels. The group velocity is derived from considering a wave train containing multiple frequencies:

$$u(\omega, t) = \int A(\omega) \exp i(\omega t - kx) d\omega \quad (3)$$

and finding the point (x,t) where the phase is stationary:

$$d(\omega t - kx)/d\omega = 0 \quad \text{hence} \quad t = x dk/d\omega \quad (4)$$

This leads to the definition of group velocity U as:

$$U(\omega) = d\omega/dx \quad (5)$$

In general, at any given period, the group velocity of surface waves in the earth will be less than the phase velocity. The minimum in the group velocity curve is termed Airy phase, a major feature of many surface wave records. Note that waves with periods shorter than the period of the Airy phase will exhibit reverse group velocity dispersion, with the shorter periods traveling faster. Phase velocity, on the other hand, always increases with period.

Surface wave generation and propagation

The theoretical formulation for the excitation of surface waves and normal modes of the earth, an eigenvalue problem, was developed in the 1960s (e.g., Haskell, 1964), much stimulated by the occurrence of the great Chilean earthquake of May 22, 1960 (M 9.5). A computational method, following the original theoretical approach of Saito (1967) based on Runge-Kutta matrix integration method, has long been the main reference for the practical calculation of surface waves and normal modes in laterally homogeneous, elastic, flat or spherical earth models. Today, a widely used approach for spherical geometry, efficient to relatively short periods (10 s), is based on a propagator matrix method in which minors of sets of solutions are used effectively (Woodhouse, 1988). The association of a normal mode formalism to compute dispersion and excitation of surface waves (and complete seismograms), with a moment tensor formalism to describe the earthquake source (e.g., Backus and Mulcahy, 1976) led to the rapid development of source studies based on surface waves in the 1980s.

To obtain the frequency spectrum of a single mode surface wave train from the expression of a seismogram obtained by summation of normal modes in a spherically symmetric earth, one uses Poisson's formula (e.g., Aki and Richards, 1980), which decomposes the modes into infinite trains of propagating surface waves traveling in opposite directions around the earth. In this process, a high frequency approximation is used, in which the phase velocity (see below) of a surface wave is related to the corresponding normal mode frequency by Jeans' formula:

$$C(\omega) = \frac{a\omega_l}{l + 1/2} \quad (6)$$

where l is the angular order of the mode and ω_l its eigenfrequency, and a is the radius of the earth. In this high-frequency approximation, surface waves propagate along the great circle path between the epicenter and the station and are sensitive to structure only along this great circle.

The spectrum of a single mode propagating surface wave at distance Δ, azimuth θ, and angular frequency ω can be expanded as follows:

$$U(\Delta, \theta, \omega) = U_s(\theta, \omega) \mathrm{S}(\Delta) U_p(\Delta, \theta, \omega) \mathrm{F}(\omega, \theta_0) \mathrm{D}(\omega) \mathrm{I}(\omega) \quad (7)$$

where U_s is the source spectrum, U_p contains propagation effects, I is the instrument response, $\mathrm{S}(\Delta)$ is the geometrical spreading term, and F and D describe the source process.

The propagation term U_p can be expressed as (e.g., Romanowicz, 2002):

$$U_p(\Delta, \theta, \omega) = \frac{1}{(\sin \Delta)^{1/2}} \exp(i\pi/4) \exp(im\pi/2) \times \exp[-i\omega\Delta/C(\omega, \theta)] \exp[-\eta(\omega, \theta)\Delta] \quad (8)$$

where m denotes the number of polar passages and $C(\omega,\theta)$, $\eta(\omega,\theta)$ are, respectively, the average phase velocity and attenuation coefficient along the source-station path.

On the other hand, the source term $U_s(\theta,\omega)$ can be expressed as a linear combination of the moment tensor elements M_{ij} of the source.

Structure studies using surface waves

Fundamental mode studies (dispersion)

Fundamental mode surface waves are well suited for the study of crust and upper mantle elastic structure, which can be deduced from their group and/or phase dispersion properties. They allow the sampling of vast areas of the globe that are otherwise devoid of seismic stations, such as the oceans.

Many early studies documented the correlation of seismic velocity variations with surface tectonic features, using regional measurements of phase and group velocities of fundamental mode Love and Rayleigh waves in the period range 20–100 s (e.g., Knopoff, 1972) or at longer periods, reaching deeper into the mantle (e.g., Toksöz and Anderson, 1966). Since the early 1970s, the computation of group velocity has relied on the multiple filtering technique introduced by Dziewonski et al. (1969). This approach, later perfected as the FTAN method (e.g., Lander, 1989), involves two steps. In the first step, an "energy diagram" is formed by windowing the seismogram over group arrival time and filtering over an array of specified center frequencies. The resulting amplitudes are then plotted as a function of time and frequency and their values contoured. The group velocity curve is

obtained by tracing the loci of maximum amplitude as a function of period. In a second step, the time domain seismogram is filtered using multiple filters centered on the group velocity curve (e.g., Figure 4).

Phase velocity is obtained from the phase Φ of the Fourier spectrum of a dispersed wave train that has been corrected for the contribution of the source and the instrument:

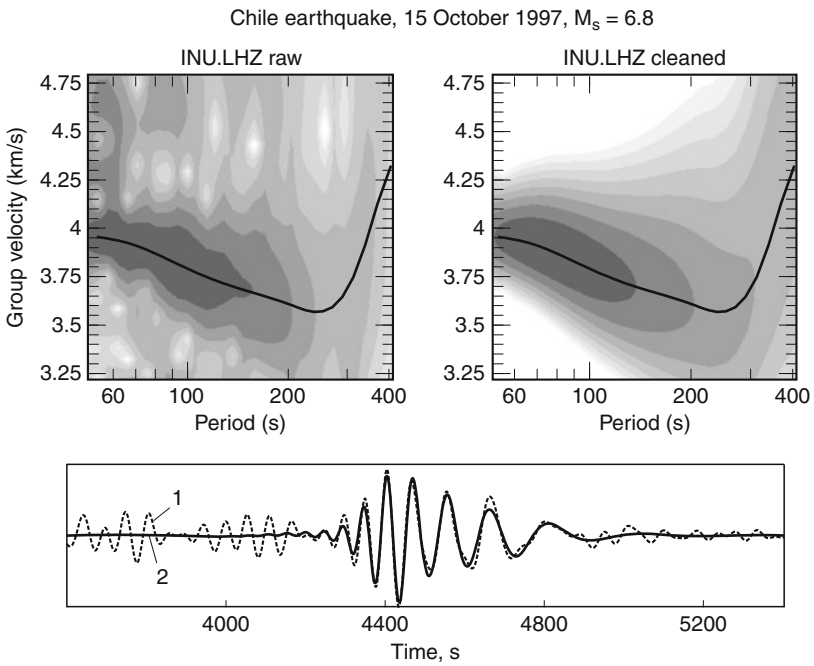
$$C(\omega) = \frac{X}{t_0 - [\Phi - N - (m/4) - 1/8]/\omega} \qquad (9)$$

where t_0 is the start time of the Fourier window with respect to the event's origin time, m is the number of polar passages, and N is an integer arising from the 2π indeterminacy of the phase. This integer is determined first at long periods to obtain reasonable values of phase velocity compatible with well-constrained global earth models. The phase velocity curve is then successively continued to shorter periods at fixed N, so as to obtain a smooth curve. This can become a problem at periods shorter than 30 s, where small variations in phase velocity correspond to rapid cycling of the phase.

The contribution of the earthquake source to the phase needs to be known accurately for the computation of phase velocity, less so for group velocity. This can be circumvented by making measurements using the "two-station method," in which dispersion is measured between two stations approximately aligned with the epicenter on a great circle, thus eliminating the common source phase. If, in addition, the two stations are located within a relatively homogeneous geological province, such measurements are called "pure path" and lead directly to the determination of elastic velocity structure beneath that province. This method has been extended to "many stations" to infer structure beneath an array of stations spanning a geologically homogeneous region.

Noise cross-correlation methodology

Recently, a radically different methodology has been introduced in seismology. The medium's "Green's function" between two seismic stations can be extracted from cross-correlations of records at both stations over long time periods, without the need for using earthquake data (see *Seismic, Ambient Noise Correlation*). These "noise cross-correlations" have been particularly effective for the measurement of the fundamental mode dispersion between two stations in and around the microseismic frequency band (5–30 s), where the background seismic noise is the most energetic. It has been shown that this approach leads to results that are as good as those obtained using standard methods, in regions where seismicity is strong and provides good illumination of the structure using "standard" approaches (e.g., Shapiro et al., 2005).



Surface Waves, Figure 4 Example of group velocity dispersion diagrams obtained for the fundamental Rayleigh wave mode on the vertical record of the Chile Ms 6.8 earthquake of October 15, 1997, at Geoscope station INU, using the FTAN multiple filtering approach. (**a**) Dispersion diagram before filtering; (**b**) same after filtering. The group velocity dispersion curve obtained follows the maxima of energy as delineated by the gray scale contours. (**c**) Corresponding time domain seismogram before (*broken line*) and after (*solid line*) variable filtering. The bandpass is 50–400 s. (Reprinted from Romanowicz 2002.)

This has opened up new horizons for the study of regional crustal structure, especially in seismically quiet regions. There remain some theoretical issues to be resolved, in particular regarding biases introduced by the nonuniform distribution of noise sources around the study region, and the possibility of extracting the overtone part of the Green's function. This is because most of the noise sources that contribute to the reconstruction are located near the earth's surface (see Sect. The earth's background noise spectrum – hum and microseisms) and thus preferentially excite the fundamental mode. Nevertheless, the methodology is developing at a rapid pace.

Global studies of structure using surface waves

With the advent of digital recording in the mid-1970s and the expansion of global digital long-period and later broadband networks (see *Seismological Networks*), the processing of the relatively long time series needed to measure surface waves became much easier and opened the way, in the 1980s to large-scale and global studies of upper mantle structure.

Global studies first proceeded according to a regionalization scheme, in which it was assumed that the depth variation of seismic velocities is the same across each tectonic province. These studies confirmed and extended to longer periods (and hence larger depths) early results on the age dependence of structure in the oceans (e.g., Romanowicz, 2002; see also *Earth's Structure, Upper Mantle*). The constraint of regionalization was soon relaxed, replaced by expansion of lateral heterogeneity into a global basis of spherical harmonics. This revealed, in particular, that the correlation of structure with tectonics disappears below depths of 200–300 km (e.g., Ekström et al., 1997). Dispersion measurements have focused primarily on the fundamental mode, generally in the period range 35–150 s, and thus are generally limited in resolution to depths shallower than ~300–400 km. This approach has led to several generations of global phase velocity maps at discrete frequencies, with increasing spatial resolution (e.g., Figure 5). In these studies, lateral variations of dispersive properties are now well resolved down to wavelengths of ~1,000–2,000 km. At the long period end, the period range is limited by the difficulty of separating consecutive wave trains, and at the short period end, by the increased complexity of surface wave propagation in the strongly heterogeneous crust and uppermost mantle, resulting in lateral refractions and multipathing. The latter are not taken into account in the simple high frequency, great circle propagation assumptions underlying the construction of dispersion maps.

Inversion of surface wave dispersion data

Inversion of surface wave dispersion data generally involves two steps. By combining measurements at a given frequency for many different paths crossing a particular region, or, at longer periods ($T >\sim 40$ s) over the entire globe (e.g., Figure 5), dispersion maps can be obtained by a standard linear inversion procedure, yielding point by point dispersion curves over the region considered.

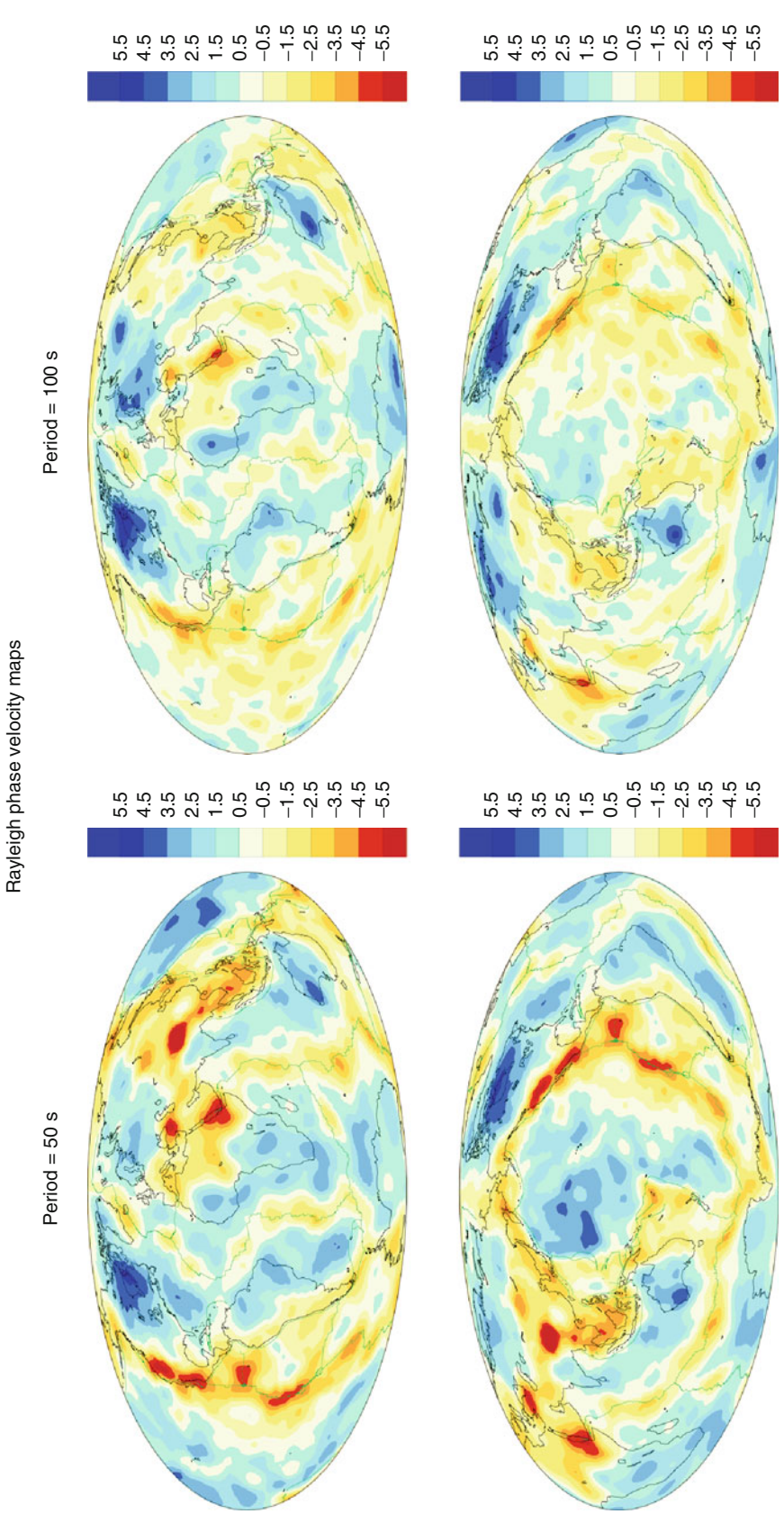
These dispersion curves can then be inverted to obtain the local variation of elastic parameters with depth, by matching observed dispersion curves to theoretical curves calculated for layered earth models, using inverse theory (e.g., Thurber and Ritsema, 2007; see also *Inverse Theory, Linear*). The models thus obtained suffer from some degree of nonuniqueness, as the problem is under-determined. In particular, surface waves cannot by themselves resolve strong gradients of velocity with depth, so that, in general, some a priori constraints are applied, such as fixing the depth of the Moho, or limiting the extent of low velocity layers. A major advance was the theoretical contribution of Backus and Gilbert (1970), which made it possible to calculate the resolving kernels and errors associated with the various least squares solutions of the inverse problem. Surface wave dispersion is most sensitive to shear velocities, so that, in general, a priori conversion factors between shear velocity and compressional velocity, on the one hand, and density, on the other, are assumed.

Crustal corrections

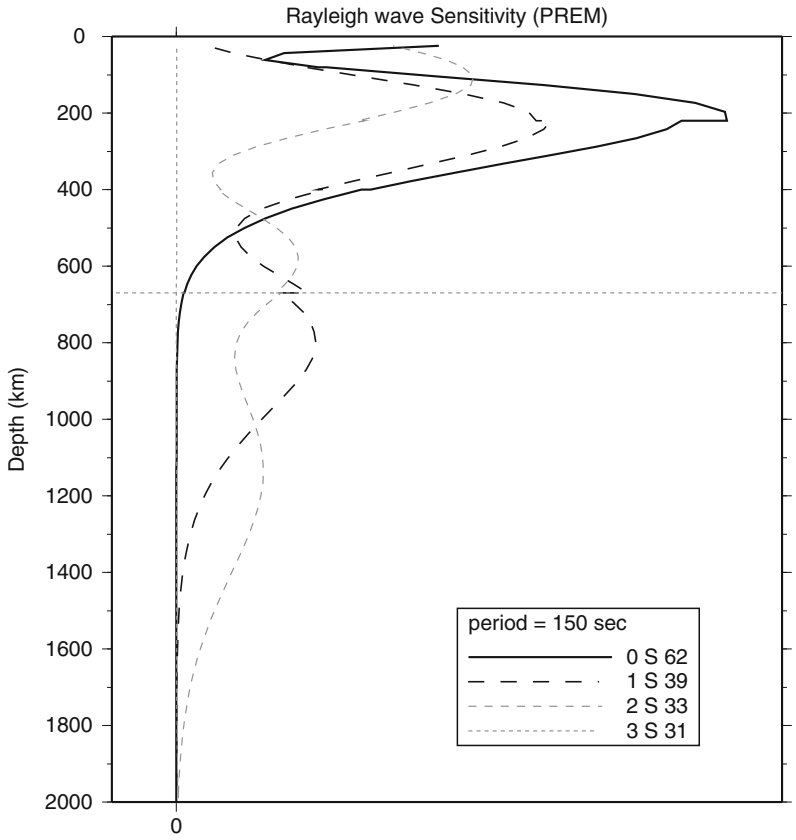
A significant issue regarding the inversion of long period fundamental mode surface waves (40–200 s) for upper mantle structure is that of crustal corrections. Indeed, surface waves are sensitive to shallow structure even at long periods, but not sensitive enough for inversion, unless more difficult to measure short periods (down to ~10 s) are included. Until recently, most studies simply performed crustal corrections in the framework of linear perturbation theory. However, Montagner and Jobert (1988) showed that the effect of the large variations in crustal thickness is nonlinear, and proposed a more accurate correction procedure based on a tectonic regionalization. Developing accurate crustal models worldwide (see *Earth's Structure, Continental Crust*) remains a challenge for large-scale surface wave inversions for structure of the upper mantle, especially in the era of numerical computation of waveforms.

Inversion using overtones

While they are well separated in the time domain from other mode branches, and therefore well suited for frequency domain single mode analysis techniques, fundamental mode surface waves have several shortcomings: at intermediate periods (~20–150 s) their sensitivity to structure below about 200 km is poor, whereas longer period mantle waves, which reach down to the top of the upper mantle transition zone, have poor spatial resolution. In any case, resolving structure in the transition zone, which is also poorly sampled by body waves, requires the inclusion of higher-mode surface waves, whose sensitivity is larger at these depths (e.g., Figure 6). They are also a powerful tool for investigating structures where low velocity zones may be present (e.g., Kovach



Surface Waves, Figure 5 Examples of global Rayleigh wave phase velocity maps at periods of 50 s (*left*) and 100 s (*right*), highlighting the diminishing lateral variations of structure with period (i.e., with depth). These maps were obtained using the dataset of Ekström et al. (1997). (Courtesy of Guy Masters.)



Surface Waves, Figure 6 Sensitivity with depth of Rayleigh wave fundamental mode (*solid line*) and first two overtones (*broken lines*) at a period of 100 s, computed in the PREM model (Dziewonski and Anderson, 1981). The fundamental mode sensitivity peaks around 200 km whereas the overtones have significant sensitivity down to the top of the lower mantle. Courtesy of Yuancheng Gung.

and Anderson, 1964). In general, however, higher-mode surface waves overlap in the time-frequency domain, and single mode dispersion methods therefore cannot be applied.

In the 1970s, array methods were developed independently by Nolet (1975) and Cara (1978) to extract information from overtones, and applied in the period range 20–100 s to paths across Eurasia and the Pacific Ocean, respectively. These methods require a linear regional array of stations approximately aligned with the epicenter (and not in a nodal direction of the source radiation pattern), in order to separate modes in the (ω,k) domain, where k is the wavenumber. They are limited in application to a few regions of the world with relatively dense, linear arrays. A waveform-based method involving the comparison of observed and synthetic seismograms at a single station was proposed by Lerner-Lam and Jordan (1983). In this approach, branch cross-correlation functions (bccf's) are formed between a particular single-mode synthetic and the observed seismograms and used to invert for an average structure along each source-station path. This approach lends itself to the derivation of secondary observables such as group or phase velocity. The main drawback of this methodology is the contamination of the single mode objective function by interference from other mode branches. More recently, van Heijst and Woodhouse (1997) proposed a "mode-branch stripping" method based on a bccf approach combined with a frequency-stepping procedure, yielding single mode dispersion and amplitudes. A global multimode dispersion dataset assembled in this fashion has been used in the construction of several generations of global shear velocity models (e.g., Ritsema et al., 2004). See also *Earth's Structure, Upper Mantle*.

Full waveform inversions

Extracting single mode dispersion information directly from spectra of windowed wave trains is only effective for the fundamental mode, which is well separated on the seismogram, and even in this case, overlapping with overtone wave trains or, at long periods, other orbits of the fundamental mode, can be a problem. A waveform matching approach is therefore desirable, where observed seismograms are compared to synthetics computed for a given reference earth model, and the model is perturbed to improve the fits of the observed and synthetics in the target time window of observation, which can be

optimized. In particular, such an approach is suitable both for fundamental mode and overtone surface waves.

For many years, the computation of long period synthetic seismograms relied on asymptotic approximations to normal mode first order perturbation theory, as developed in the 1970s. The most widely used approximation is the "path average approximation," introduced for waveform modeling by Woodhouse and Dziewonski (1984). The key steps that led to the various formalisms routinely used today are described in Romanowicz et al. (2008). See also *Seismic, Waveform Modeling and Tomography*.

Working under the assumption of PAVA, Nolet (1990) introduced "the partitioned waveform inversion" (PWI), in which inversion for elastic structure proceeds in two steps, more closely related to the two-step inversion using dispersion data. Path integral parameters are defined and retrieved by nonlinear waveform fitting over each path, and in a second step, inverted linearly for elastic structure. The advantage of this approach is that it reduces the number of parameters to be fit in the nonlinear part of the inversion. The PWI also includes progressive filtering of data starting at the longest periods (see *Seismic Tomography*).

Because it includes only coupling of modes along single dispersion branches, the PAVA is rigorously only suitable for single-mode wave trains such as the fundamental mode. The corresponding sensitivity kernels to structure are 1D kernels, that is, they only depend on the laterally averaged structure in the vertical plane between the source and the receiver. When applied to overtones and body waves, coupling across mode branches needs to be included. For a review of these methodologies, see Romanowicz et al. (2008) and Thurber and Ritsema (2007).

Upper mantle anisotropy from surface wave studies

Early studies of surface wave dispersion showed that, in many regions, it is not possible to find a single isotropic model that satisfies both fundamental mode Love and Rayleigh wave dispersion simultaneously. This provided some of the earliest evidence for the presence of anisotropy in the crust and upper mantle (e.g., Anderson, 1961). This discrepancy can be explained by introducing a transversely isotropic medium with a vertical symmetry axis (radial anisotropy), down to at least 200 km depth. It is in this context that the still widely used Preliminary Reference Earth Model (PREM) was constructed (Dziewonski and Anderson, 1981). Gung et al. (2003) showed that radial anisotropy is also present at depths greater than 200 km beneath the lithospheric roots of stable continents and cratons, and explains discrepancies in lithospheric thickness obtained using surface waves of different polarization, thus reconciling lithospheric thickness obtained from seismology with estimates from other geophysical and xenolith data.

Surface wave dispersion also varies with azimuth of propagation, an indication of "azimuthal anisotropy" (e.g., Forsyth, 1975). The global azimuthal variations of Rayleigh and Love wave dispersion at long periods (100–250 s) were first mapped by Tanimoto and Anderson (1985), who showed that the fast direction appears to correlate with flow directions in the mantle. Montagner and Tanimoto (1991) developed the first global model of radial and azimuthal anisotropy in the upper mantle based on fundamental mode surface wave data. While some questions remain about trade-offs between lateral heterogeneity and anisotropy in this type of inversion, Montagner and collaborators have shown that they can explain their datasets with fewer parameters when azimuthal anisotropy is considered than when it is ignored.

To first order, asymptotically, the azimuthal variation of phase velocity (Love or Rayleigh waves) is of the form (Smith and Dahlen, 1973):

$$C(\omega,\theta) = A_0 + A_1\cos(2\theta) + A_2\sin(2\theta) + A_3\cos(4\theta) + A_4\sin(4\theta) \tag{10}$$

where θ is the azimuth of the wavenumber vector defined clockwise from north. The coefficients $A_i(\omega)$ depend linearly on the elastic tensor at every point in the medium. Theoretical expressions and how they can be applied for crust and upper mantle studies are reviewed in Romanowicz (2002) and Montagner (2007). Because Rayleigh waves are sensitive both to shallow crustal and deeper mantle anisotropy, it is important to use a wide frequency range to resolve the depth dependence of anisotropy using surface waves. Montagner and Nataf (1988) introduced "vectorial tomography." By combining radial and azimuthal dispersion terms, under the assumption that the material possesses a symmetry axis (orthotropic medium), they showed that the 3D model can be described using seven elastic parameters (plus density): the five parameters describing transverse isotropy, and two angles describing the orientation in space of the axis of symmetry. Combining long period waveforms and SKS splitting measurements, Yuan and Romanowicz (2010) showed that azimuthal anisotropy provides a powerful tool to detect layering in the continental lithosphere.

Using different seismological tools, including surface wave overtones (e.g., Trampert and van Heijst, 2002), it has also been suggested that azimuthal anisotropy is present in the transition zone. To date, these results are not consistent with each other and still controversial.

Other diagnostic effects of anisotropy in long period surface waves are waveform anomalies caused by Rayleigh-Love coupling, which generates "quasi-Love" waves on vertical components and "quasi-Rayleigh" waves on transverse (e.g., Park and Maupin, 2007). See also *Seismic Anisotropy*.

Effects of scattering and the Born approximation

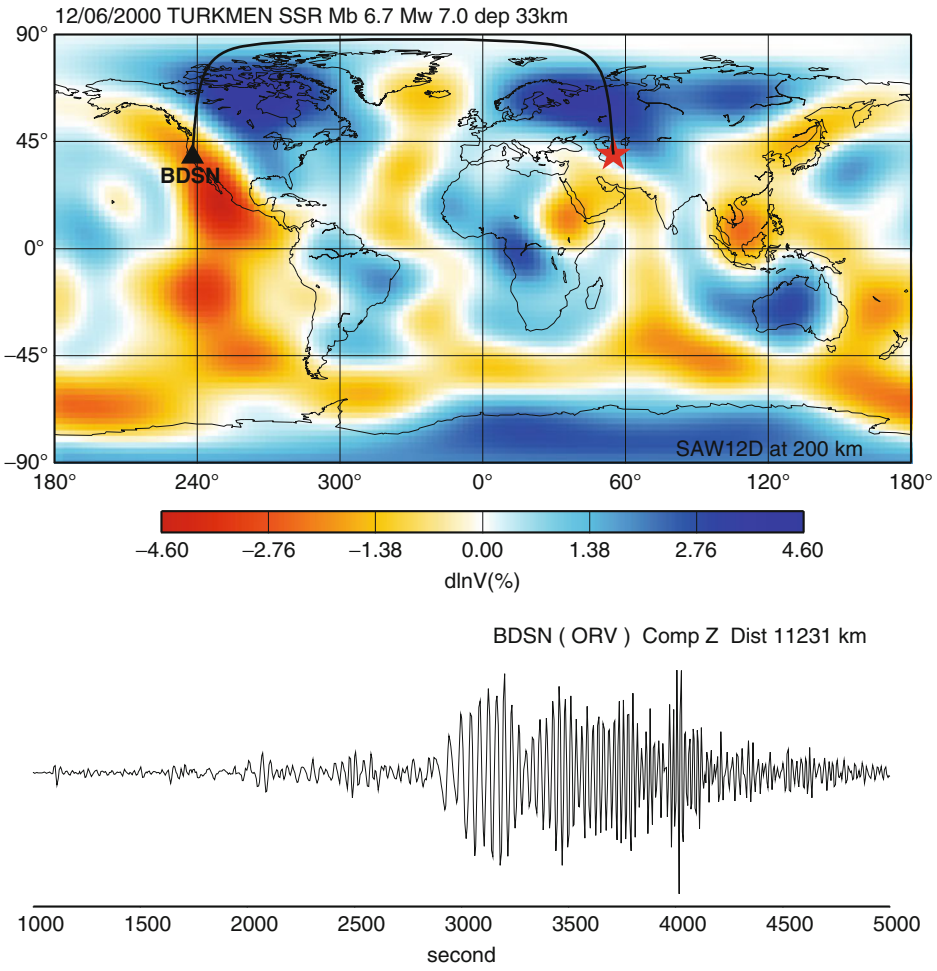
Until recently, most regional and global models of upper mantle structure derived from surface waves were based

on the standard "path average" approximation (PAVA). This is valid only if the wavelength of lateral variations of structure is long with respect to that of the surface waves considered. And it is not strictly appropriate for overtones.

Many observations indicate that lateral heterogeneity is strong enough to cause departures from PAVA. For example, 20 s surface waves sensitive to shallow crustal structure consistently show multipathing (e.g., Capon, 1970 or Figure 7). At longer periods ($T > 100$ s), later arriving trains often show larger amplitudes than the ones preceding them which cannot be explained by lateral variations of attenuation.

Several approaches have been developed to account for these effects and exploit them to obtain better constraints on lateral variations of structure. Woodhouse (1974) introduced the concept of local modes. These are the surface wave modes of a laterally homogeneous model, which locally has the depth distribution of the laterally varying model. If the medium is laterally smooth, the local mode branches propagate as independent wave trains, and their dispersion is modified according to the evolution of the local modes. However, if lateral variations are sharp (e.g., in the presence of a structural discontinuity such as an ocean-continent boundary), the coupling of the local modes cannot be neglected and its strength depends on the width of the structural transition zone (Kennett, 1972). For a recent review see Romanowicz (2002). Kennett (1992) derived a formalism for mode-coupling in 2D slowly laterally varying structures, which was extended to the 3D scattering case, in Cartesian coordinates (e.g., Bostock and Kennett, 1992).

In the framework of scattering theory, the use of tools based on the single-scattering Born approximation has gained popularity in the last 5 years. Born scattering is well suited for inversion since the scattered wavefield



Surface Waves, Figure 7 Example of vertical component record at BDSN station ORV for an *Mw* 7.0 earthquake in Turkmenistan showing multipathing of 20–30 s surface waves. The record has been bandpass filtered between 0.001 and 0.03 Hz. The great circle path is shown on the top plot with a shear wave tomographic model in the background. It arrives to the station from the north, following the coast and along major structural boundaries in the crust that are likely the cause of the multipathing. (Courtesy of Yuancheng Gung.)

depends linearly on structural perturbations (e.g., Snieder, 1988). The use of Born kernels is widely advocated to take into account the effects on dispersion of off-great circle sensitivity of surface waves (e.g., Zhou et al., 2006). However, the relative merits of single-scattering Born sensitivity kernels versus simpler averaging kernels, such as "fat rays" is still debated (e.g., Boschi et al., 2006).

Born is also a poor approximation in the case of strong heterogeneity, or when the region of scattering is large (e.g., Friederich et al., 1993).

Most recently, the introduction to global seismology of the Spectral Element Method (Komatitsch and Vilotte, 1998), which provides an accurate representation of the teleseismic wavefield in arbitrary 3D media up to the second surface wave orbit, has provided a new powerful tool for the study of Earth structure, and is progressively being implemented for waveform modeling of crust and upper mantle structure at the local (e.g., Tape et al., 2010), regional (Fichtner et al., 2010), and global scale (Lekic and Romanowicz, 2010), focusing, so far, on the relatively long period surface waves. These studies show that the use of a more accurate wave propagation theory results in the improved retrieval of amplitudes of lateral variations of shear wave velocity, and in particular, significantly lower values in low velocity zones. For applications at the global scale, this promising new tool presents several challenges: a considerable computational cost, and the necessity of accurate representation of the 3D structure in the crust, a topic that undoubtedly will be addressed in the coming years.

Surface wave attenuation

Surface wave measurements provide the primary constraints on attenuation structure in the crust and uppermost mantle. In the absence of perturbing effects due to scattering and focusing, the amplitude spectrum of a single mode wave train can be written as:

$$A(\omega) = A_0(\omega)\exp(-\eta(\omega)X) \tag{11}$$

where X is the epicentral distance in kilometers and $A_0(\omega)$ represents the amplitude at the source. The attenuation coefficient $\eta(\omega)$ is related to the quality factor Q through (e.g., Aki and Richards, 1980):

$$\eta(\omega) = \frac{\omega}{2C(\omega)Q(\omega)}$$

where C is the phase velocity.

Regional studies of amplitudes of fundamental mode surface waves in the period range 5–100 s and of Lg waves have long established the presence of large variations of Q correlated with tectonic provinces and in particular with the age of the oceans, and with time elapsed since the latest tectonic activity on continents. For recent reviews, see Mitchell (1995) and Romanowicz and Mitchell (2007). Recently, attenuation structure in the vicinity of the East Pacific Rise and in the back-arcs and wedges of subduction zones obtained from a combination of surface wave and body wave data have revealed highly attenuating zones underlying volcanic arcs, down to the slab surface. See also *Seismic, Viscoelastic Attenuation*.

Lateral variations in attenuation at long periods can be an order of magnitude larger than those in elastic velocity. Progress in constraining global 3D anelastic structure of the upper mantle has been slow because of the inherent difficulty of measuring attenuation in the presence of focusing and scattering effects that can be as large as anelastic ones and depend strongly on the short-wavelength details of the elastic structure. The long wavelength 3D attenuation models obtained thus far indicate that lateral variations in attenuation in the first 200–250 km of the mantle are correlated with tectonics. In the transition zone, the long wavelength pattern shifts to one correlated with hotspots and with the structure at the base of the mantle, with high attenuation overlying the two large low shear velocity provinces under Africa and in the Pacific (e.g., Romanowicz and Mitchell, 2007). Progress in resolving finer details of attenuation structure awaits the routine implementation of numerical methods for the computation of long period seismograms to fully account for focusing effects.

Source studies using surface waves

Fundamental mode surface wave spectra contain information about the source moment tensor, source depth (centroid), source process time, and, under favorable circumstances for very large earthquakes, source directivity. Tsai and Aki (1971) first showed that the amplitude spectrum of Rayleigh waves contained the signature of source depth, in the form of a "hole" in the spectrum in the period range 10–100 s, which appears at a period depending on depth, and which also depends on the source mechanism.

In order to correct for propagation, different approaches need to be taken depending on the size of the earthquake and the period range considered. At very long periods ($T > 180$ s) and for large earthquakes ($M > 6.5$), propagation effects can be accounted for approximately using a spherically symmetric reference earth model and ellipticity corrections. At shorter periods, and for smaller earthquakes, corrections on individual source-station paths need to be known much more accurately (for a review, see Romanowicz, 2002). In the last 15 years, the availability of increasingly accurate global 3D tomographic models of the upper mantle has made it possible to extend the period range to shorter periods (down to ~120 s) and to efficiently make use of methodologies, based on time-domain waveform inversion, that are no longer restricted to the fundamental mode, thus providing more accurate estimation of the source depth. Such a waveform approach was first introduced by Dziewonski et al. (1981), who combined waveforms of mantle waves at periods greater than 120 s with overtone waveforms at periods greater than 80 s. This forms the basis of a now routine procedure that serves to construct the widely used Harvard centroid moment tensor (CMT) catalog (Now "global CMT catalog").

The earth's background noise spectrum: hum and microseisms

The presence of background peaks in the frequency range 2–7 mHz, corresponding to the fundamental spheroidal free oscillations of the earth, in the absence of large earthquakes, was discovered recently by Japanese scientists (e.g., Kobayashi and Nishida, 1998). It was later determined that the vertical component long period seismic noise spectrum was made up of Rayleigh waves (Nishida et al., 2002) and that the main source of the "hum" was located in the oceans (Rhie and Romanowicz, 2004), resulting from nonlinear interactions between the atmosphere (oceanic storms), the oceans, and the solid earth. The "hum" has also been detected on the horizontal components, corresponding to Love waves.

The largest background seismic noise is found in the period range 2–20 s and corresponds to the so-called microseismic peak. Microseisms were studied extensively in the 1950s and determined to be surface waves generated in the oceans. There are two types of microseisms: "primary" and "secondary." Secondary microseisms are the strongest. Also called "double-frequency microseisms," they are generated at frequency f by a nonlinear mechanism involving two opposing ocean waves of frequency $f/2$ (Longuet-Higgins, 1950). In the last decade, there has been renewed interest in the study of microseisms, in the context of the use of noise cross-correlations for the study of Earth structure, as seen above. While it has been proposed that the earth's low frequency hum generation mechanism is related to that of microseisms, no common explanation has yet been found for the vertical and horizontal hum generation. See also *Seismic Noise*.

Summary

Surface waves propagate along the surface of the earth with exponentially decreasing energy with depth, with waves of longer period reaching greater depths. They contain most of the elastic energy generated by earthquakes at periods greater than 20 s and are dispersive. There are two types of surface waves, Rayleigh and Love waves, with different polarization properties. The fundamental mode surface waves are well separated from other energy arrivals in the time domain and provide the best constraints to date on continental scale and global scale structure. They play a particularly important role in constraining the upper mantle structure in the ocean basins where few seismic stations have been installed. While the dispersion of fundamental mode surface waves is a classical tool for the investigation of upper mantle structure, higher modes, or overtones provide constraints at transition zone depths (400–700 km) and deeper. They require more sophisticated analysis tools because they cannot be distinguished from each other readily on the seismogram. Waveform inversion approaches are increasingly favored to handle both fundamental mode and overtone interpretation, and hold increasing promise as numerical methods for the computation of the seismic wavefield in arbitrary 3D structures are being implemented, and can account for complex scattering and focusing effects accurately. In addition to isotropic shear velocity, surface waves provide constraints on the distribution of polarization and azimuthal anisotropy in the upper mantle, as well as on anelastic attenuation. When path effects have been corrected for, surface waves provide robust constraints on the source depth and radiation pattern, as expressed by its moment tensor.

The earth's continuous background noise consists primarily of surface waves, at least at periods longer than 5 s. Notable is the microseismic noise peak around 6–7 s. Most of this energy is generated in the oceans by nonlinear interactions involving wind-driven ocean waves, and the seafloor. Recently, a method based on the cross-correlation of noise records has gained popularity and has been applied successfully to resolve crust and uppermost mantle structure. It is particularly useful in seismically quiet regions.

Bibliography

Aki, K., and Richards, P. G., 1980. *Quantitative Seismology, Theory and Methods*. San Francisco: W.H. Freeman.

Anderson, D. L., 1961. Elastic wave propagation in layered anisotropic media. *Journal of Geophysical Research*, **66**, 2953–2963.

Backus, G., and Gilbert, F., 1968. The resolving power of gross earth data. *Geophysical Journal of the Royal Astronomical Society*, **16**, 169–2050.

Backus, G., and Mulcahy, M., 1976. Moment tensors and other phenomenological descriptions of seismic sources I- continuous displacements. *Geophysical Journal of the Royal Astronomical Society*, **46**, 341–371.

Boschi, L., Becker, T. W., Soldati, G., and Dziewonski, A. M., 2006. On the relevance of Born theory in global seismic tomography. *Geophysical Research Letters*, **33**, L06302, doi:10.1029/2005GL025063.

Capon, J., 1970. Analysis of Rayleigh-wave multipath propagation at LASA. *Bulletin of the Seismological Society of America*, **60**, 1701–1731.

Cara, M., 1978. Regional variations of higher Rayleigh-mode phase velocities: a spatial filtering method. *Geophysical Journal of the Royal Astronomical Society*, **54**, 439–460.

Dziewonski, A. M., and Anderson, D. L., 1981. Preliminary reference Earth model. *Physics of the Earth and Planetary Interiors*, **25**, 297–356.

Dziewonski, A. M., Bloch, S., and Landisman, M., 1969. A new technique for the analysis of transient seismic signals. *Bulletin of the Seismological Society of America*, **59**, 427–444.

Dziewonski, A. M., Chou, A. T., and Woodhouse, J. H., 1981. Determination of earthquake source parameters from waveform data for studies of global and regional seismicity. *Journal of Geophysical Research*, **86**, 2825–2852.

Ekström, G., Tromp, J., and Larson, E. W. F., 1997. Measurements and global models of surface wave propagation. *Journal of Geophysical Research*, **102**, 8137–8157.

Ewing, W. M., Jardetsky, W. S., and Press, F., 1957. *Elastic Waves in Layered Media*. New York: McGraw Hill.

Fichtner, A., Kennett, B., Igel, H., and Bunge, H.-P., 2010. Full waveform tomography for radially anisotropic structure: new insights into present and past states of the Australasian upper mantle. *Earth and Planetary Science Letters*, **290**, 270–280.

Forsyth, D. W., 1975. The early structural evolution and anisotropy of the oceanic upper mantle. *Geophysical Journal of the Royal Astronomical Society*, **43**, 103–162.

Friederich, W., Wielandt, E., and Stange, S., 1993. Multiple forward scattering of surface waves: comparison with an exact solution and Born single-scattering methods. *Geophysical Journal International*, **112**, 264–275.

Gung, Y. C., Panning, M., and Romanowicz, B., 2003. Anisotropy and thickness of the lithosphere. *Nature*, **422**, 707–711.

Haskell, B., 1964. Radiation pattern of surface waves from point sources un a multi-layered medium. *Bulletin of the Seismological Society of America*, **54**, 377.

Kennett, B. L. N., 1972. Seismic waves in laterally heterogeneous media. *Geophysical Journal of the Royal Astronomical Society*, **27**, 301–325.

Knopoff, L., 1972. Observation and inversion of surface-wave dispersion. *Tectonophysics*, **13**, 497–519.

Kobayashi, N., and Nishida, K., 1998. Continuous excitation of planetary free oscillations by atmospheric disturbances. *Nature*, **395**, 357–360.

Komatitsch, D., and Vilotte, J.-P., 1998. The spectral-element method: an efficient tool to simulate the seismic response of 2D and 3D geological structures. *Bulletin of the Seismological Society of America*, **88**, 368–392.

Kovach, R. L., and Anderson, D. L., 1964. Higher mode surface waves and their bearing on the structure of the Earth's mantle. *Bulletin of the Seismological Society of America*, **54**, 161–182.

Lander, A. V., 1989. Frequency-time analysis. In Keilis-Borok, V. I. (ed.), *Seismic Surface Waves in a Laterally Inhomogeneous Earth*. Dordrecht: Kluwer Academic, pp. 153–163.

Lekic, V., and Romanowicz, B., 2010. Inferring upper mantle structure by full waveform tomography with the Spectral Element Method. *Geophysical Journal International*, in revision.

Lerner-Lam, A., and Jordan, T. H., 1983. Earth structure from fundamental and higher-mode waveform analysis. *Geophysical Journal of the Royal Astronomical Society*, **75**, 759–797.

Longuet-Higgins, M., 1950. A theory of the origin of microseisms. *Philosophical Transactions of Royal Society London*, **243**, 1–35.

Mitchell, B. J., 1995. Anelastic structure and evolution of the continental crust and upper mantle from seismic surface wave attenuation. *Reviews of Geophysics*, **33**, 441–462.

Montagner, J. P., 2007. Upper mantle structure: global isotropic and anisotropic elastic tomography. In Romanowicz, B., and Dziewonski, A. M. (eds.), *Treatise on Geophysics.* Oxford: Elsevier, Vol. 1, pp. 559–590.

Montagner, J. P., and Jobert, N., 1983. Variation with age of the deep structure of the Pacific Ocean inferred from very long-period Rayleigh wave dispersion. *Geophysical Research Letters*, **10**, 273–276.

Montagner, J. P., and Jobert, N., 1988. Vectorial tomography II. Application to the Indian Ocean. *Geophysical Journal International*, **94**, 309–344.

Montagner, J. P., and Nataf, H. C., 1988. Vectorial tomography – I. Theory. *Geophysical Journal International*, **94**, 295–307.

Montagner, J. P., and Tanimoto, T. 1991. Global upper mantle tomography of seismic velocities and anisotropy. *Journal of Geophysical Research*, **96**(20), 20337–20351.

Nishida, K., Kobayashi, N., and Fukao, Y., 2002. Origin of Earth's ground noise from 2 to 20 mHz. *Geophysical Research Letters*, **29**, 1413, doi:10.1029/2001GL013862.

Nolet, G., 1975. Higher-Rayleigh modes in western Europe. *Geophysical Research Letters*, **2**, 60–62.

Nolet, G., 1990. Partitioned waveform inversion and two-dimensional structure under the network of autonomously recording seismographs. *Journal of Geophysical Research*, **95**, 8499–8512.

Park, J., and Maupin, V., 2007. Theory and observations – wave propagation in anisotropic media. In Romanowicz, B., and Dziewonski, A. M. (eds.), *Treatise on Geophysics*. Oxford: Elsevier, Vol. 1, pp. 323–360.

Rhie, J., and Romanowicz, B., 2004. Excitation of Earth's free oscillations by atmosphere-ocean-seafloor coupling. *Nature*, 431, 552–555.

Ritsema, J., van Heijst, H. J., and Woodhouse, J. H., 2004. Global transition zone tomography. *Journal of Geophysical Research*, **109**, B02302, doi:10.1029/2003JB002610.

Romanowicz, B., 2002. Inversion of surface waves: a review. In Lee, W. H. K. (ed.), *Handbook of Earthquake and Engineering Seismology*, *Part A*. IASPEI, pp. 149–174.

Romanowicz, B., and Mitchell, B. J., 2007. Q in the Earth from crust to core. In Romanowicz, B., and Dziewonski, A. M. (eds.), *Treatise of Geophysics*. Oxford: Elsevier, Vol. 1, pp. 731–774.

Romanowicz, B., Panning, M., Gung, Y., and Capdeville, Y., 2008. On the computation of long period seismograms in a 3D earth using normal mode based approximations. *Geophysical Journal International*, **175**, 520–536.

Saito, M., 1967. Excitation of free oscillations and surface waves by a point source in a vertically heterogeneous Earth. *Journal of Geophysical Research*, **72**, 3689.

Shapiro, N. M., Campillo, M., Stehly, L., Ritzwoller, M. H., 2005. High-resolution surface wave tomography from ambient seismic noise. *Science*, **307**, 1615–1618.

Snieder, R., 1988. Large-scale waveform inversions of surface waves for lateral heterogeneity, 1. Theory and numerical examples. *Journal of Geophysical Research*, **93**, 12055–12065.

Tanimoto, T., and Anderson, D. L., 1985. Lateral heterogeneity and azimuthal anisotropy of the upper mantle: Love and Rayleigh waves 100-250s. *Journal of Geophysical Research*, **90**, 1842–1858.

Tape, C., Liu, J., Maggi, A., and Tromp, J., 2010. Seismic tomography of the Southern California crust based on the spectral-element and adjoint methods. *Geophysical Journal International*, **180**, 433–462.

Thurber, C., and Ritsema, J., 2007. Seismic tomography and inverse methods. In Romanowicz, B., and Dziewonski, A. M. (eds.), *Treatise on Geophysics*. Oxford: Elsevier, Vol. 1, pp. 1323–1360.

Toksöz, M. N., and Anderson, D. L., 1966. Phase velocities of long period surface waves and structure of the upper mantle, 1, Great circle Love and Rayleigh wave data. *Journal of Geophysical Research*, **71**, 1649–1658.

Trampert, J., and van Heijst, H., 2006. Global azimuthal anisotropy in the transition zone. *Science*, **296**, 1297–1299.

Tsai, Y. B., and Aki, K., 1971. Amplitude spectra of surface waves from small earthquakes and underground nuclear explosions. *Journal of Geophysical Research*, **75**, 5729.

van Heist, H., and Woodhouse, J., 1997. Measuring surface-wave overtone phase velocities using a mode-branch stripping technique. *Geophysical Journal International*, **131**, 209–230.

Woodhouse, J. H., 1974. Surface waves in laterally varying structure. *Geophysical Journal of the Royal Astronomical Society*, **37**, 461–490.

Woodhouse, J. H., 1988. The calculation of eigenfrequencies and eigenfunctions of the free oscillations of the earth and the sun. In Doornbos, D. J. (ed.), *Seismological Algorithms*. San Diego: Academic Press, pp. 321–370.

Woodhouse, J. H., and Dziewonski, A. M., 1984. Mapping the upper mantle: three dimensional modelling of the earth structure by inversion of seismic waveforms. *Journal of Geophysical Research*, **89**, 5953–5986.

Yang, Y., and Forsyth, D. W., 2006. Regional tomographic inversion of the amplitude and phase of Rayleigh waves with 2-D sensitivity kernels. *Geophysical Journal International*, **166**, 1148–1160.

Yu, Y., and Park, J., 1994. Hunting for azimuthal anisotropy beneath the Pacific Ocean region. *Journal of Geophysical Research*, **99**, 15399–15421.
Yuan, H., and Romanowicz, B., 2010. Lithospheric layering in the north American craton. *Nature*, **466**, 1063–1069.
Zhou, Y., Nolet, G., Dahlen, F., and Laske, G., 2006. Global upper-mantle structure from finite-frequency surface-wave tomography. *Journal of Geophysical Research*, **111**, B04,304, doi:10.1029/2005JB003677.

Cross-references

Earth's Structure, Continental Crust
Earth's Structure, Upper Mantle
Earthquakes, Source Theory
Free Oscillations of the Earth
Inverse Theory, Linear
Lithosphere, Continental
Propagation of Elastic Waves: Fundamentals
Seismic Anisotropy
Seismic Noise
Seismic Tomography
Seismic, Ambient Noise Correlation
Seismic, Viscoelastic Attenuation
Seismic, Waveform Modeling and Tomography
Seismological Networks

T

T WAVES

Emile A. Okal
Department of Earth & Planetary Sciences, Northwestern University, Evanston, IL, USA

Definition and introduction

T phases are defined as seismic recordings of signals having traveled an extended path as acoustic waves in the water body of the oceans. This is made possible by the "Sound Fixing and Ranging" (SOFAR) channel, a layer of minimum sound velocity acting as a wave guide at average depths of 1,000 m (Okal, 2007). It allows the efficient propagation of extremely small signals over extremely long distances, in practice limited only by the finite size of the ocean basins. The existence of the SOFAR channel results from the dependence of the velocity of sound in water on temperature, pressure and salinity. As a result, the detailed structure of the channel (including the value of the minimum velocity) varies both geographically (mainly with latitude) and seasonally, but at first order, the SOFAR channel can be regarded as a quasi-universal feature of the world's oceans.

At the shoreline of a continent or island, the acoustic wave is converted into a seismic wave which can be recorded by a seismometer or, occasionally, felt by humans. Efficient propagation inside the waveguide requires short wavelengths, in practice frequencies above 3 Hz, and thus T phases generally attenuate fast once converted, although in exceptional geometries they have been detected up to 1,000 km inside stable continental shields. Even though records of T waves were reported as early as 1927, they were not identified as hydroacoustic signals until the late 1940s, when results from antisubmarine warfare during World War II were declassified.

Sources of T waves

T waves can be generated by a diversity of underwater sources, as long as their energy becomes trapped in the SOFAR channel. Most earthquakes in oceanic margins generate strong T phases, since in the presence of a sloping ocean floor, multiple downslope reflections can lead to efficient penetration of the SOFAR (Johnson et al., 1963). However, their generation by earthquakes in abyssal plains deprived of large-scale bathymetry defies the laws of geometrical optics, and remains to this day a major theoretical challenge; it is generally thought that abyssal T phases result from scattering along small random heterogeneities of the ocean floor (Yang and Forsyth, 2003). Also, T waves are occasionally generated by non-shallow earthquakes, even the deepest ones, even though their source is de facto distant from the ocean column. This excitation is made possible by travel through the subducting oceanic slabs which, being cold, propagate high frequencies with limited attenuation and can deliver them to the water interface for conversion to acoustic waves (Okal, 2001a).

Underwater landslides are also T-wave generators. As compared with earthquakes, such sources, which move generally smaller segments of ocean floor over much larger distances, have a considerably longer duration, a lower-frequency spectrum, and thus result in T waves of longer duration but smaller amplitudes (Okal, 2003).

By contrast, explosions at sea, which are intrinsically high-frequency, are particularly efficient sources of T waves; in most instances, they involve a shallow source located inside the SOFAR channel, thus providing optimal feeding of acoustic energy into the channel. In particular, seismic reflection experiments are routinely recorded thousands of kilometers away.

Underwater volcanic eruptions are also regular generators of T waves. Because of the diversity of processes involved during a volcanic sequence, these signals may

Harsh K. Gupta (ed.), *Encyclopedia of Solid Earth Geophysics*, DOI 10.1007/978-90-481-8702-7,

feature a wide variety of characteristics. The opening of cracks during the formation of magmatic plumbing systems is essentially in the nature of an earthquake, while magmatophreatic explosions at the contact between a lava conduit and the oceanic column are reminiscent of chemical explosions. The steady output of lava into the ocean takes the form of a prolonged noise of low amplitude which may be compared to underwater slumping (Caplan-Auerbach et al., 2001). Finally, the oscillation of magma inside the plumbing system, known to cause volcanic tremor long recorded seismically on subaerial volcanoes, generates harmonic T-wave signals. Supercritical hydrothermal venting during volcanic events has also been documented as the source of monochromatic hydroacoustic signals (Talandier and Okal, 1996).

Among the most remarkable T waves ever reported, signals from the eruption of Macdonald Volcano (29°S; 140°W) on 29 May 1967 led to the eventual discovery of this unchartered underwater seamount, and its interpretation as the active member of the Cook-Austral hotspot chain (Johnson, 1970).

The collision of large icebergs, during which their masses can rub against each other, has also been documented as a source of T waves, which can exhibit a harmonic character during episodes of stick-and-slip friction between tabular icebergs (MacAyeal et al., 2008). Hydroacoustic waves are also generated during the disintegration of large icebergs, and thus could be used to monitor and assess any possible change in climatic conditions (Li and Gavrilov, 2006).

Finally, marine mammals can talk into the SOFAR channel and correspond across hundreds of kilometers, using signals in the range of a few tens of hertz, but specific to each species.

Human perception

T waves of sufficient amplitude can be felt by shoreline populations. For example, the underwater nuclear test WIGWAM on 14 May 1955 off the coast of Southern California was felt in Hawaii and even in Japan, 8,000 km away. Large events such as the Alaskan earthquake of 10 July 1958 and the deep Bolivian shock of 9 June 1994 were felt in Hawaii, and T waves from the 2004 Sumatra earthquake were felt in the Maldives and possibly on Zanzibar; they could have provided a warning of the impending tsunami in these otherwise aseismic areas, but their populations were unprepared.

Use in comprehensive nuclear-test ban treaty monitoring

Because hydroacoustic waves can detect small sources (especially intra-oceanic ones) at great distances, they have been included as one of the four technologies used by the International Monitoring System (IMS) of the Comprehensive Nuclear-Test Ban Treaty. Their spectacular efficiency in propagation allows a full coverage of the all the Earth's oceans using only 11 receiver sites, six of which are hydrophones deployed within the SOFAR channels, and five special seismic stations close to shorelines ("T-phase stations"), instrumented for high-frequency seismic recording in the 50−100 Hz range (Okal, 2001b).

Acoustic thermometry of ocean climate

This project consisted of using acoustic sources in the 57 Hz range fired in the vicinity of Heard Island, South Indian Ocean, a location allowing the illumination of all three major oceans (Munk et al., 1994). Detection at extreme ranges of up to 17,000 km allowed precise measurement of acoustic velocities, and it was envisioned to repeat these experiments over several decades, to monitor any evolution of sound velocities, as a proxy to changes in temperature, possibly in the context of global warming. Unfortunately, the project had to be halted in the late 1990s under pressure from the environmental lobby.

Quantification

Like other seismic phases, quantitative measurements of T waves can provide information on source parameters. However, absolute measurements (especially of amplitudes) at seismic receivers are difficult to interpret due to the complexity and site specificity of the conversion processes. In addition, the high-frequency nature of the phase results in amplitude saturation at relatively low magnitudes. In this context, quantification of T waves has emphasized duration of wavetrains over amplitude, and relative measurements under common or comparable receiver conditions. *Okal et al*. (2003) have introduced the concept of *T*−phase energy flux, *TPEF*, reminiscent of the evaluation of seismic energy from body waves, and which should scale with seismic moment M_0 for regular seismic sources. Any deviation in the ratio $\Gamma = TPEF/M_0$ identifies anomalous source properties; in particular, slow earthquakes exhibit deficient Γ, a property with some potential in tsunami warning.

The comparison of duration and amplitude of T wavetrains can efficiently discriminate between earthquakes and explosions at sea, based on the strong differences in the scaling of their source processes (Talandier and Okal, 2001), although small earthquakes located inside steep volcanic edifices remain a challenge in this respect.

A few geophysical applications

As most of their path is hydroacoustic, T phases propagate at a velocity close to that of sound in water, 1.5 km/s; that gives them late arrivals on seismic records, hence their name "T" for "third". Such slow velocities allow a remarkable resolving power for source locations, after accounting for small geographic and seasonal variations in SOFAR velocities, and for the contribution of the converted [seismic] segments of the phase.

In this context, the superior detection capabilities of T phases have been used to improve considerably our coverage of low-magnitude seismicity in the oceanic

environment, and especially our understanding of small-scale processes at mid-oceanic ridges, notably in conjunction with swarms of volcanic activity (Dziak et al., 1995).

The slowness of T waves also allows to very precisely beam a receiving array (e.g., triangles of hydrophones at sites of the International Monitoring System [IMS]) to back-track the source of T phases, or even of individual fragments of their wavetrains; this procedure was used to obtain a detailed source tomography of the great 2004 Sumatra-Andaman event (e.g., Tolstoy and Bohnenstiehl, 2005).

Conclusion

Because of their high-frequency character, T waves provide a valuable complement to more traditional phases in the field of body-wave seismology. Their incorporation into the IMS allows their more systematic recording and opens up a number of new research opportunities.

Bibliography

Caplan-Auerbach, J., Fox, C. G., and Duennebier, F. K., 2001. Hydroacoustic detection of submarine landslides on Kilauea volcano. *Geophysical Research Letters*, **28**, 1811–1813.

Dziak, R. P., Fox, C. G., and Schreiner, A. E., 1995. The June-July 1993 seismo-acoustic event at CoAxial segment, Juan de Fuca Ridge: Evidence for a lateral dike injection. *Geophysical Research Letters*, **22**, 135–138.

Johnson, R. H., 1970. Active submarine volcanism in the Austral Islands. *Science*, **167**, 977–979.

Johnson, R. H., Northrop, J., and Eppley, R., 1963. Sources of Pacific T phases. *Journal of Geophysical Research*, **68**, 4251–4260.

Li, B., and Gavrilov, A. N., 2006. Hydroacoustic observation of Antarctic ice disintegration events in the Indian Ocean, Proc. First Australasian Acoustical Societies Conference, pp. 479−484, Christchurch.

MacAyeal, D. R., Okal, E. A., Aster, R. C., and Bassis, J. N., 2008. Seismic and hydroacoustic tremor generated by colliding icebergs. *Journal of Geophysical Research*, **113**(F03), 1–10. F03011.

Munk, W. H., Spindel, R. C., Baggeroer, A., and Birdsall, T. G., 1994. The Heard Island feasibility test. *Journal of the Acoustical Society of America*, **96**, 2330–2342.

Okal, E. A., 2001a. "Detached" deep earthquakes: Are they really? *Physics of the Earth and Planetary Interiors*, **127**, 109–143.

Okal, E. A., 2001b. T − phase stations for the international monitoring system of the comprehensive nuclear-test ban treaty: A global perspective. *Seismological Research Letters*, **72**, 186–196.

Okal, E. A., 2003. T waves from the 1998 Papua New Guinea earthquake and its aftershocks: Timing the tsunamigenic slump. *Pure and Applied Geophysics*, **160**, 1843–1863.

Okal, E. A., 2007. The generation of T waves by earthquakes. *Advances in Geophysics*, **49**, 1–65.

Okal, E. A., Alasset, P.-J., Hyvernaud, O., and Schindelé, F., 2003. The deficient T waves of tsunami earthquakes. *Geophysical Journal International*, **152**, 416–432.

Talandier, J., and Okal, E. A., 1996. Monochromatic T waves from underwater volcanoes in the Pacific Ocean: ringing witnesses to geyser processes? *Bulletin of the Seismological Society of America*, **86**, 1529–1544.

Talandier, J., and Okal, E. A., 2001. Identification criteria for sources of T waves recorded in French Polynesia. *Pure and Applied Geophysics*, **158**, 567–603.

Tolstoy, M., and Bohnenstiehl, D. R., 2005. Hydroacoustic constraints on the rupture duration, length, and speed of the great Sumatra-Andaman earthquake. *Seismological Research Letters*, **76**, 419–425.

Yang, Y., and Forsyth, D. W., 2003. Improving epicentral and magnitude estimation of earthquakes from T phases by considering the excitation function. *Bulletin Seismological Society of America*, **93**, 2106–2122.

Cross-references

Seismic Monitoring of Nuclear Explosions
Seismology, Monitoring of CTBT

THERMAL STORAGE AND TRANSPORT PROPERTIES OF ROCKS, I: HEAT CAPACITY AND LATENT HEAT

Christoph Clauser
Institute for Applied Geophysics and Geothermal Energy, E.ON Energy Research Center, RWTH Aachen University, Aachen, Germany

Synonyms

Specific heat capacity; Thermal capacity; Volumetric heat capacity

Definition

Specific heat capacity c. Physical property defining the amount of sensible heat that can be stored in or extracted from a unit mass of rock per unit temperature increase or decrease, respectively. Isobaric and isochoric specific heat capacities are defined at constant pressure and volume, respectively; dimension: J kg^{-1} K^{-1}.

Thermal capacity (*also: volumetric heat capacity*) ρ c. The product of isobaric specific heat capacity and density. Physical property defining the amount of sensible heat that can be stored in or extracted from a unit volume of rock per unit temperature increase or decrease, respectively; dimension: J m^{-3} K^{-1}.

Thermal storage properties

The thermal regime of the Earth is defined by its heat sources and sinks, the heat storage and transport processes, and their corresponding physical properties. The storage properties are discussed below. The transport properties, thermal conductivity and thermal diffusivity, are dealt with in this volume in the companion chapter "*Thermal Storage and Transport Properties of Rocks, II: Thermal conductivity and diffusivity*."

Heat can be stored as sensible heat and enthalpy of transformation. In the Earth, sensible heat is defined by the heat capacity of rocks, and the enthalpy of transformation by their latent heat of fusion.

Heat is transmitted within the Earth mainly by diffusion (in the steady state: conduction), advection, and radiation. Generally, heat diffusion dominates heat radiation within the lithosphere of the Earth, at temperatures below

about 1 000 K. For sufficiently large flow rates, convection-driven heat advection, however, can be equally or even more efficient than diffusion. The heat diffusion–advection equation for a porous medium is:

$$\frac{\partial(\rho c_P T)}{\partial t} = \nabla \cdot \left(\underbrace{\lambda \nabla T}_{q_{dif}} - \underbrace{(\rho c_P)_f T \mathbf{v}}_{q_{adv}} \right) + \underbrace{A}_{\text{generation}}, \quad (1)$$

where T is temperature in K, t time in s, ρ density in kg m^{-3}, c isobaric specific heat capacity in J kg^{-1} K^{-1}, λ thermal conductivity in W m^{-1} K^{-1}, **v** specific fluid discharge (volume flow rate per unit cross section) in m s^{-1}, A (radiogenic) heat generation rate in W m^{-3}; subscript "f" refers to fluid properties. Diffusive bulk storage and transport properties for heat in rocks governing this equation are thermal capacity ρ c_P, also referred to as volumetric heat capacity, and thermal conductivity λ. Advection is governed by the thermal capacity and specific discharge of the flowing fluid, $(\rho\ c_P)_f$ and **v**, respectively.

Heat advection does not require very large flows for becoming as or more efficient than heat diffusion. This is often the case in sedimentary basins (see, e.g., Clauser et al. (2002) for a literature review). But fluid-driven heat advection may be also important in crystalline rocks and on a crustal scale (e.g., Clauser, 1992). The non-dimensional Péclet and Nusselt numbers, Pe and Nu, are ratios of specific heat flows q_{adv} and q_{dif} (Equation 1) indicating the efficiency of (fluid flow driven) advective heat transport versus heat conduction for given flow geometries. For instance, assuming flow of magnitude v over a distance L and a temperature difference $T_1 - T_0$ one obtains:

$$\begin{aligned} Pe &= \frac{q_{adv}}{q_{dif}} = \frac{(\rho c)_f v (T_1 - T_0)}{\lambda (T_1 - T_0)/L} = \frac{(\rho c)_f v L}{\lambda}; \\ Nu &= \frac{q_{adv} + q_{dif}}{q_{dif}} = \frac{q_{adv}}{q_{dif}} + 1 = Pe + 1. \end{aligned} \quad (2)$$

Thus advection or diffusion (in the steady state: conduction) dominate for Pe > 1 or Pe < 1, respectively (in terms of the Nusselt number for Nu > 2 or Nu < 2, respectively). At temperatures above 1 000 K, heat is propagated increasingly as radiated electromagnetic waves, and heat radiation begins to dominate diffusion (see *Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity*, this volume).

Heat capacity

Heat can be stored and delivered as sensible heat or as latent heat required or liberated by phase changes. This and the next two paragraphs are concerned with sensible heat. Following this, latent heat will be discussed.

Sensible heat capacity C is defined as the ratio of heat ΔQ required to raise the temperature of a mass M of rock by ΔT. For each molecule of mass m, this temperature increase requires an energy of (f/2) k ΔT, where f is the number of degrees of freedom of the molecule and k = 1.380 650 4(24) × 10^{-23} J K^{-1} is Boltzmann's constant. For a body of mass M, a temperature increase of ΔT requires an energy of ΔQ = (M/m) (f/2) k ΔT. Thus the heat capacity of the body at constant volume is:

$$C_V = \frac{\Delta Q}{\Delta T} = \frac{f}{2} k \frac{M}{m} = \frac{f}{2} k N_A = \frac{f}{2} R \ (JK^{-1}), \quad (3)$$

where Avogadro's number N_A = 6.022 141 79(30) × 10^{23} mol^{-1} equals the number of molecules or atoms in an amount of substance of 1 mol and R = k N_A = 8 314 472(15) J mol^{-1} K^{-1} is the molar Gas constant (numerical values for N_A, R, k, and all other physical constants used in this chapter are from CODATA, 2006). For solids, f = 6, corresponding to the three degrees of freedom of potential and kinetic lattice vibration energy in each space direction. Accordingly, the heat capacity of one mole of substance, the molar heat capacity at constant volume is constant:

$$C_{V,mol} = 3kN_A = 3R = 24.94 \ (J\ mol^{-1} K^{-1}) \quad (4)$$

where N_A = 6.022 141 79(30) × 10^{23} mol^{-1} is Avogadro's number, the number of molecules in one mole of substance, and R = k N_A the molar gas constant. Isobaric heat capacity $C_{P,mol}$ is larger than isochoric heat capacity $C_{V,mol}$ because additional work is required for volume expansion. Both are related by:

$$C_{P,mol} = C_{V,mol} + R = \frac{f+2}{2} R \ (J\ mol^{-1} K^{-1}). \quad (5)$$

With $C_{V,mol}$ from Equation 4 and assuming, as above, f = 6 this yields:

$$\begin{aligned} C_{P,mol} &= 3R + R = 4R \\ &= 33.26 \ (J\ mol^{-1} K^{-1}), \text{or:} \\ C_{P,mol} &- C_{V,mol} = R. \end{aligned} \quad (6)$$

Equation 4, the *Dulong–Petit law*, is satisfied well for heavy elements. In contrast, molar heat capacities of lighter elements remain below this limiting value, the lower the temperature the smaller $C_{V,mol}$.

Below the Debye temperature, Θ_D, heat capacity varies with temperature. Θ_D tends to zero as T^3 as absolute temperature approaches zero. Θ_D falls in the range 85 K and 450 K for most substances and 200 K and 1 000 K for most minerals (Stacey and Davis, 2008; Table 1). Therefore, heat capacity in the Earth can be well explained by classical Debye theory, in particular in the mantle and except for a thin crustal layer near the Earth's surface. There are, however, exceptions such as beryllium (Θ_D = 1 440 K) and diamond ($\Theta_D \approx$ 1 800 K). These are caused by the so-called freezing of vibrational or rotational degrees of freedom, which cannot absorb heat any more at low temperature. Therefore heat capacity tends to zero close to absolute zero.

Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat, Table 1 Debye temperature Θ_D and mass number A of selected elements (Kittel, 2004)

Element	Mass number A	Θ_D (K)
Carbon	6	2 230
Aluminum	13	428
Silica	14	645
Titanium	22	420
Chromium	24	630
Manganese	25	410
Iron	26	470
Nickel	28	450
Copper	29	343
Zinc	30	327
Silver	47	215
Cadmium	48	209
Tin	50	200
Tantalum	73	240
Tungsten	74	400
Platinum	78	240
Gold	79	165
Lead	82	105

Isobaric and isochoric specific heat capacity

Specific heat capacity c of a substance is defined as heat capacity C related to unit mass:

$$c = \frac{\Delta Q}{M\,\Delta T} = \frac{f}{2}\frac{k}{m} = \frac{f}{2}\frac{k}{A_r\,m_u}\left(J\,kg^{-1}K^{-1}\right), \tag{7}$$

where $m_u = 1.660\,538\,782(83) \times 10^{-27}$ kg is the atomic mass constant, defined as $^1/_{12}$ of the atomic mass of the carbon isotope ^{12}C, and A_r the atomic mass of a substance relative to m_u. Isobaric specific heat capacity c_P is larger than isochoric specific heat capacity c_V because additional work is required for volume expansion. Their ratio, the adiabatic exponent, is:

$$c_P/c_V = (f+2)/f. \tag{8}$$

Alternatively, isobaric specific heat capacity c_P can be expressed by enthalpy H(T,P) = E + P V, a state function of temperature and pressure, where E, P, and V are internal energy, pressure, and volume, respectively. In a closed system, the change in internal energy dE is the sum of the change in heat dQ and the work dW delivered: dE = dQ + dW. If we only consider volume expansion work, dW = −P dV, the change in enthalpy dH becomes:

$$\begin{aligned} dH(T,P) &= dE + pdV + VdP = dQ + VdP \\ &= \left(\frac{\partial H}{T}\right)_P dT + \left(\frac{\partial H}{P}\right)_T dP. \end{aligned} \tag{9}$$

Comparing coefficients, we obtain:

$$\frac{dQ}{dT} = \left(\frac{\partial H}{\partial T}\right)_P \doteq c_P. \tag{10}$$

Thus, Equation 10 defines isobaric specific heat capacity c_P as the first derivative of enthalpy with respect to temperature. Comparison of Equations 7 and 10 shows that both expressions are equivalent for dQ = ΔQ/M, and the isobaric enthalpy change ΔH is equal to the specific heat content ΔQ/M.

Isobaric and isochoric specific heat capacity are related to compressibility β = ΔV/(VΔP) and its inverse, incompressibility or bulk modulus K = VΔP/ΔV, by $c_P/c_V = \beta_T/\beta_S = K_S/K_T$ (e.g., Stacey and Davis, 2008), where subscripts T and S refer to isothermal and adiabatic conditions, respectively. Inserting the thermodynamic relation $\beta_T = \beta_S + \alpha^2\,T/(\rho\,c_P)$ (e.g., Birch, 1966) between isothermal and adiabatic compressibility yields the relative difference between isobaric and isochoric specific heat capacity:

$$c_P/c_v = 1 + \alpha\gamma T, \tag{11}$$

where α = ΔV/(VΔT) is the volume expansion coefficient,

$$\gamma = \frac{\alpha K_S}{\rho c_P} = \frac{\alpha K_T}{\rho c_V}, \tag{12}$$

the dimensionless Grüneisen parameter, and ρ density. Inserting the expressions for α and K into Equation 12 yields:

$$\gamma = \frac{1}{\rho c_P}\frac{\Delta V}{V\,\Delta T}\frac{V\Delta P}{\Delta V} = \frac{\Delta P}{\rho c_P \Delta T}. \tag{13}$$

Thus the Grüneisen parameter γ is the relative pressure change in a material heated at constant volume.

For solids, i.e., f = 6, the absolute difference between isobaric and isochoric specific heat capacity follows from Equations 11 and 8:

$$c_P - c_V = \frac{K_T\alpha^2 T}{\rho} = \frac{3K_S\alpha^2 T}{4\rho}. \tag{14}$$

For crustal rocks (γ = 0.5; α = 20 μK^{-1}; T < 10^3 K; ρ = 2 600 kg m^{-3}; K_S < 75 GPa (Dziewonski and Anderson, 1981; Stacey and Davis, 2008)), the difference between isobaric and isochoric specific heat capacity is less than 9 J kg^{-1} K^{-1} or 1% according to Equations 14 and 11, respectively. Thus, the distinction between isobaric and isochoric specific heat capacity is negligible for crustal rocks at temperatures below 1 000 K. However, it need be made for mantle rocks. From here on, "specific heat capacity" will always refer to isobaric specific heat capacity, denoted simply by the letter c without the subscript "P."

This classical treatment of heat capacity is sufficient for temperatures above the Debye temperature. In the Earth, temperature exceeds the Debye temperature everywhere except in the crust (Stacey and Davis, 2008). Therefore, in experiments at room temperature and atmospheric pressure, we observe deviations from the values predicted by Equations 3–14, which are based on the classical Dulong–Petit theory. The lower the temperature, lighter the element, and stronger the lattice bonding become, the larger are these deviations. Clearly, interpretation of heat

capacity below the Debye temperature is beyond classical mechanics and requires quantum mechanical treatment. This is, however, beyond the scope of this text and interested readers are referred to standard physics textbooks (e.g., Tipler and Mosca, 2007). Therefore heat capacity at crustal temperatures should not be calculated from Equations 3 and 14 but rather be measured or calculated from appropriate, quantum mechanical equations.

Čermák and Rybach (1982) compiled data on isobaric specific heat capacity for different rock-forming minerals and different igneous, metamorphic, volcanic, and sedimentary rocks as well as the corresponding variations with temperature.

Measuring techniques

Specific heat capacity c can be measured directly or derived as the isobaric derivative of enthalpy H with respect to temperature. Specific heat capacity of rocks varies with temperature, pressure, porosity, and saturants. Accordingly, in situ values may deviate from laboratory data according to temperature, pressure, and type and content of pore fluid.

Numerous steady-state and transient calorimetric methods are available for measuring specific heat capacity. The most popular are mixing or drop calorimeters and heat flux differential scanning (DSC) calorimeters. The first method yields an absolute value; the second one is a comparative method. All of these methods and their details are discussed in the literature (e.g., Hemminger and Cammenga, 1989; Brown, 2001; Haines, 2002) to which interested readers are referred. The isobaric enthalpy change (or specific heat content) ΔH of solids may be determined by the method of mixtures using a Bunsen-type calorimeter, in which the unknown isobaric enthalpy change of a sample relative to a base temperature, e.g., 25 °C, is compared to the corresponding known isobaric enthalpy change of platinum (Kelley, 1960; Somerton, 1992).

Calculated heat capacity

When no direct measurements can be performed, the isobaric enthalpy change and specific heat capacity of rocks can be calculated according to Kopp's law, Equation 15, as the arithmetic mean of the individual mineralogical and fluid contributions weighted by the volume fractions n_i of the N individual phases relative to total rock volume:

$$\Delta H = \sum_{i=1}^{N} n_i \Delta H_i; \quad c = \sum_{i=1}^{N} n_i c_i; \quad 1 = \sum_{i=1}^{N} n_i. \qquad (15)$$

Based on data for various minerals (e.g., Kelley, 1960; Berman and Brown, 1985; Somerton, 1992; Waples and Waples, 2004), the isobaric enthalpy change ΔH or specific heat capacity c can be computed from Equation 15 for any rock consisting of an arbitrary number of minerals with given volume fractions.

Temperature dependence

Derived from measured variation of isobaric enthalpy change ΔH with temperature of various oxides, Kelley (1960) suggested a second-order polynomial for fitting ΔH from which $c_P = (\partial H/\partial T)_P$ can be easily calculated. Somerton (1992) and Clauser (2006) report ΔH and c values of various rock-forming oxides and pore fluids. An alternative approach is fitting heat capacity measured at different temperatures directly to polynomials of various degrees (e.g., Maier and Kelley, 1932; Berman and Brown, 1985; Fei and Saxena, 1987; Holland and Powell, 1996; Robertson and Hemingway, 1995). Waples and Waples (2004) provide a discussion of the various approaches. The polynomial proposed by Berman and Brown (1985),

$$C_{P,mol} = k_0 + k_1 T^{-0.5} + k_2 T^{-2} + k_3 T^{-3} (T \text{ in K}), \qquad (16)$$

works over a large temperature range and yields no values incompatible with the Dulong–Petit law for high temperatures. Table 2 lists values for the coefficients k_0–k_3 in Equation 16 determined from fits of heat capacity of selected minerals measured at different temperatures.

As an alternative, Waples and Waples (2004) propose a statistical approach describing the general temperature dependence of all rocks and minerals, which can be rescaled easily for individual rocks and minerals. To this end, measured specific heat capacity c_P was normalized by the corresponding normalizing value $c_{P,n}$ at 200 °C (473.15 K), a temperature at or near which data was available. The resulting polynomial regressions yielded much better coefficients of determination R^2 for data measured on nonporous rock ($R^2 = 0.93$) than for those measured on minerals ($R^2 = 0.62$) while the trends were similar. The regression on the combined data for minerals and nonporous rocks yields an expression for the normalized specific heat capacity of a mineral or nonporous rock at arbitrary temperature T with a coefficient of determination $R^2 = 0.65$:

$$c_{P,n}(T) = 0.716 + 1.72 \times 10^{-3} T - 2.13 \times 10^{-6} T^2 + 8.95 \times 10^{-10} T^3, (T \text{ in } ^\circ\text{C}). \qquad (17)$$

Equation 17 can be rescaled for any mineral or nonporous rock at any temperature T_2 provided a value $C_P(T_1)$ measured at temperature T_1 is available, for instance from any of the compilations of Berman and Brown (1985), Fei and Saxena (1987), Berman (1988), Holland and Powell (1996), or Robertson and Hemingway (1995):

$$c_P(T_2) = c_P(T_1) \frac{c_{P,n}(T_2)}{c_{P,n}(T_1)}. \qquad (18)$$

Additionally, Waples and Waples (2004) consider the variation of specific heat capacity with lithology, where interested readers find a specific discussion regarding coals of different carbon content or maturity.

Mottaghy et al. (2005) used a second-order polynomial in temperature to fit the variation of isobaric specific heat

Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat, Table 2 Coefficients for calculating isobaric molar heat capacity $C_{P,mol}$ (J mol^{-1} K^{-1}) (From Equation 16, Berman, 1988)

Mineral	Chemical composition	k_0 (J mol^{-1} K^{-1})	$k_1 \times 10^{-2}$ (J mol^{-1} K$^{-½}$)	$k_2 \times 10^{-5}$ (J mol^{-1} K^{1})	$k_3 \times 10^{-7}$ (J mol^{-1} K^{2})	T (K)
Albite	$NaAlSi_3O_8$	393.64	−24.155	−78.928	107.064	250–1 373
Almandine	$Fe_3Al_2Si_3O_{12}$	573.96	−14.831	−292.920	502.208	420–997
Andalusite	Al_2SiO_3	236.48	−11.029	−75.268	93.644	254–1 601
Anorthite	$CaAl_2Si_2O_8$	439.37	−37.341	0.0	31.702	292–1 373
Calcite	$CaCO_3$	178.19	−16.577	−4.827	16.660	257–1 200
Cordierite	$Mg_2Al_4Si_3O_{18}$	954.39	−79.623	−21.173	−37.021	256–1 652
Corundum	Al_2O_3	155.02	−8.284	−38.614	40.908	250–2 300
Dolomite	$CaMg(CO_3)_2$	328.48	−25.544	−46.885	79.038	250–650
Enstatite (ortho)	$MgSiO_3$	166.58	−12.066	−22.706	27.915	254–1 273
Fayalite	Fe_2SiO_4	248.93	−19.239	0.0	−13.910	255–1 370
Forsterite	Mg_2SiO_4	238.64	−20.013	0.0	−11.624	253–1 807
Hematite	Fe_2O_3	146.86	0.0	−55.768	52.563	258–1 757
Kaolinite	$Al_2Si_2O_5(OH)_4$	523.23	−44.267	−22.443	9.231	256–560
Kyanite	Al_2SiO_3	262.68	−20.014	−19.997	−6.318	252–1 503
Lime	CaO	58.79	−1.339	−11.471	10.298	250–1 176
Magnesite	$MgCO_3$	162.30	−11.093	−48.826	87.466	288–750
Magnetite	Fe_3O_4	207.93	0.0	−72.433	66.436	257–1 825
Muscovite	$KAl_3Si_3O_{10}(OH)_2$	651.49	−38.732	−185.232	274.247	257–967
Periclase	MgO	61.11	−2.962	−6.212	0.584	250–1 798
Potassium feldspar	$KAlSi_3O_8$	381.37	−19.411	−120.373	183.643	250–997
Quartz	SiO_2	80.01	−2.403	−35.467	49.157	250–1 676
Sillimanite	Al_2SiO_5	256.73	−18.827	29.774	25.096	253–1 496
Sphene	$CaTiSiO_5$	234.62	−10.403	−51.183	59.146	255–1 495
Spinel	$MgAl_2O_4$	235.90	−17.666	−17.104	4.062	256–1 805
Talc	$Mg_3Si_4O_{10}(OH)_2$	664.11	−51.872	−21.472	−32.737	250–639
Wollastonite	$CaSiO_3$	149.07	−6.903	−36.593	48.435	251–1 433

capacity with temperature measured on a suite of metasedimentary, volcanic, magmatic, and metamorphic rocks:

$$c_P(T) = A_0 + A_1 T^1 + A_2 T^2 (c_p \text{ in J kg}^{-1}\text{K}^{-1}, 1\,^\circ\text{C} \le T \le 100\,^\circ\text{C}). \quad (19)$$

The average values for the coefficients A_0–A_2 determined from a regression of $c_P(T)$ measured over a temperature range of 1 °C–100 °C on 26 samples from seven boreholes are: $\bar{A}_0 = 0.074717725$ J kg^{-1} K^{-1}; $\bar{A}_1 = 1.862585346$ J kg^{-1} K^{-2}; $\bar{A}_2 = -2510.632231$ J kg^{-1} K^{-3}.

Based on a composition of 30 % quartz, 60 % feldspar (albite), and 10 % phyllosilicates (5% phlogopite, and 5 % annite), Whittington et al. (2009) suggest average "bulk crustal" molar specific heat capacity equations based on end-member mineral data for two temperature ranges, separated by the transition at 846 K (~573 °C) between α- and β-quartz:

$$C_{P,mol}(T) = \begin{cases} 199.50 + 0.0857T - 5.0 \times 10^6 T^2; & T \le 846\text{K} \\ 229.32 + 0.0323T - 47.9 \times 10^{-6} T^2; & T > 846\text{K} \end{cases}; \quad (C_{P,mol} \text{ in J mol}^{-1}\text{K}^{-1}). \quad (20)$$

Assuming an average molar mass of 0.22178 kg mol^{-1}, this yields the variation of isobaric specific heat capacity c_P with temperature shown in Figure 1.

Volumetric heat capacity: thermal capacity

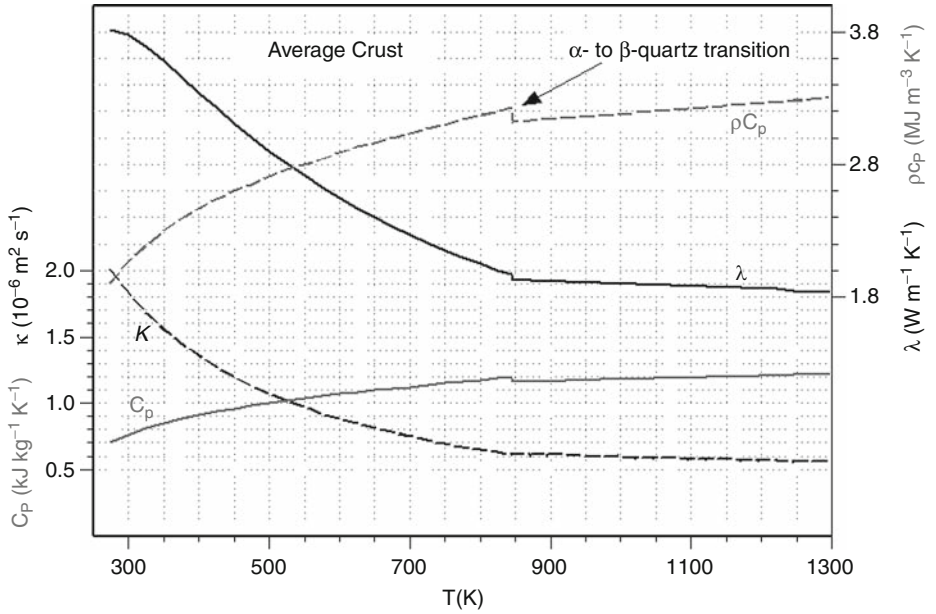
When heat capacity is related to unit volume rather than to unit mass or unit amount of substance, it is referred to as volumetric heat capacity or thermal capacity. It can be calculated as the product of specific heat capacity c and density ρ or as the ratio of thermal conductivity λ and thermal diffusivity κ by

$$\rho c = \lambda/\kappa. \quad (21)$$

Again, Kopp's law yields the rock's bulk thermal capacity $(\rho\, c)_b$ as:

$$(\rho\, c)_b = (1-\phi)\,(\rho\, c)_s + \phi \sum_{i=1}^{N} S_i\,(\rho c)_i, \quad (22)$$

where ϕ is porosity, $(\rho\, c)_s$ thermal capacity of the rock skeleton, S_i fractional saturation, and $(\rho\, c)_i$ thermal capacity of the i^{th} fluid phase in the pore space. The skeleton thermal capacity itself may be calculated again from Kopp's law for a given mineral assemblage and the corresponding volume fractions of the solid phase from Equation 15. Because of the low density of air and gas – about three orders of magnitude lower than that of water



Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat, Figure 1 Variation of specific heat capacity c_P, the phonon components of thermal diffusivity and thermal conductivity κ_p and λ_p, respectively, and thermal capacity $\rho c_P = \lambda_p/\kappa_p$ with temperature in an average crust according to Equation 20.

and rock – the contribution of the gas phase to thermal capacity can often be ignored. In this case, N = 2 for the fluid phases water and oil or N = 1 for water only. Expressions for the density of various fluids are reported in Clauser (2006). Data on density of various minerals and rocks are listed, e.g., in Wohlenberg (1982a,b) or Olhoeft and Johnson (1989). Using Equations 17 and 18, Waples and Waples (2004) analyzed a substantial collection of density and specific heat capacity data from various authors and transformed specific heat capacity and thermal capacity to a uniform reference temperature of 20 °C (Table 3).

The mean thermal capacity of "impervious" rocks was found at 2.300(46) MJ m^{-3} K^{-1} by Roy et al. (1981). This is acceptably close to the mean of 2.460(65) MJ m^{-3} K^{-1} found by Waples and Waples (2004) for inorganic minerals.

Based on density, specific heat capacity, and thermal conductivity measured at room temperature, Mottaghy et al. (2005) determined thermal capacity as inverse slope of a regression of diffusivity on thermal conductivity according to Equation 21. All values fell well within ±20 % of the average of 2.3 MJ m^{-3} K^{-1} recommended by Beck (1988).

Mottaghy et al. (2008) determined average values for thermal capacity according to Equation 21 for metamorphic and magmatic crystalline rocks as the inverse slope of a linear regression of values of thermal diffusivity versus thermal conductivity measured at temperatures in the range 20 °C–300 °C:

$$\kappa(T) = \frac{\lambda(T)}{m + nT}, (T \text{ in } ^\circ C), \quad (23)$$

Regression of data measured on seven samples collected along a profile crossing the Eastern Alps from north to south and on nine samples from the northern rim of the Fennoscandian Shield near the Kola ultra-deep borehole SG-3 yielded m = 20 66(70) kJ m^{-3} K^{-1}, n = 2.2 (4) kJ m^{-3} K^{-2}, R^2 = 0.97 and m = 2404(91) kJ m^{-3} K^{-1}, n = 3.6(5) kJ m^{-3} K^{-2}, R^2 = 0.92, respectively. This yields a range of thermal capacity for the Alpine and Fennoscandian data of about 2.1 MJ m^{-1} K^{-1}–2.7 MJ m^{-1} K^{-1} and 2.4 MJ m^{-1} K^{-1}–3.5 MJ m^{-1} K^{-1}, respectively, in the temperature range 20 °C–300 °C.

The product of an average density of 2 700 kg m^{-3} for the crust and isobaric specific heat capacity calculated for an average molar mass of 0.22178 kg mol^{-1} according to Equation 20 (Whittington et al., 2009) yields the variation of thermal capacity ρc_P (Equation 21) with temperature shown in Figure 1. Its increase with temperature by about a factor of 1.7 in a temperature interval of 1 000 K demonstrates that the effect of temperature is stronger for the phonon component of thermal diffusivity κ than for phonon thermal conductivity λ_p due to the increase of specific heat capacity. Assuming a constant density throughout the crust implies that the increase and decrease in density due to the increase in pressure and temperature, respectively, partly cancel each other and that these changes are small compared to those of specific heat capacity and thermal diffusivity.

Latent heat

Solidification of magma and melting of rocks as well as freezing and thawing of water in soils or rocks liberates or consumes heat, respectively. The like applies to mineral

phase changes such as those associated with the seismic discontinuities at 410 km, 520 km, and 660 km in the transition zone from 400 km–600 km between the upper and lower mantle. These mineral phases are chemically identical but differ with respect to crystal structure and therefore elastic properties. This is why this transition is seen in the seismic wave field. Phase transitions require a certain pressure P and temperature T, but also a specific relation between these two state variables expressed by the so-called Clapeyron slope $dP/dT = \Delta S(P,T)/\Delta V(P,T)$ (the inverse of the Clausius–Clapeyron equation), where S and V are entropy and volume, respectively. This means that the depth where a certain phase transition occurs varies with the ambient temperature in the crust. Positive and negative values for the Clapeyron slope are associated with exothermic and endothermic reactions, respectively.

Phase changes generally consume or deliver much more latent heat than can be stored or delivered as sensible heat: It requires a temperature increase of more than 500 K to equal by sensible heat the amount of latent heat required to melt 1 kg of granite, and still an increase of more than 80 K to equal by sensible heat the amount of latent heat required to melt 1 kg of sea ice (Table 4).

The *discontinuity at 410 km* is generally associated with the transition in olivine from α-olivine to β-spinel, also named wadsleyite (e.g., Stacey and Davis, 2008). At expected pressure and temperature of 14 GPa and 1,600 K, respectively, corresponding values for the Clapeyron slope vary between 2.9 MPa K^{-1} and 4.8 MPa K^{-1} (Bina and Helffrich, 1994; Katsura et al., 2004; Stacey and Davis, 2008). The *discontinuity at 520 km* is associated with the transition from β-spinel (wadsleyite) into γ-spinel, also named ringwoodite (e.g., Stacey and Davis, 2008). At an expected pressure of 18 GPa, a temperature increase, such as by an ascending plume, would require a higher pressure for this transition according to inferred Clapeyron slopes of 4.0 MPa K^{-1}–5.3 MPa K^{-1} (Helffrich, 2000; Deuss and Woodhouse, 2001). A second transition occurs between garnet and calcium-perovskite ($CaSiO_3$), where the iron in garnet goes into Ilmenite, and its $CaSiO_3$-component into calcium-perovskite. This reaction has a negative Clapeyron slope. The two slopes of different sign may shift the depth for two transitions into opposite directions, which is observed as a splitting of the 520 km discontinuity. The *discontinuity at 660 km* defines the transition into the lower mantle. It is caused by the transition of γ-spinel (ringwoodite) into magnesium-perovskite ($MgSiO_3$) and ferrous periclase (magnesiowüstite, (Fe,Mg)O). At an expected pressure of 23.5 MPa, this endothermic transition is associated with a Clapeyron slope of − 2.8 MPa K^{-1} (Stacey and Davis, 2008).

The latent heat L that corresponds to these additional heat sources and sinks can be elegantly combined with the specific sensible heat capacities of the liquid and solid rock, c_l and c_s, respectively, into an effective bulk specific heat capacity c_{eff}. This effective specific heat capacity then accounts for the entire enthalpy change, including latent heat. In this approach, the latent heat effects are assumed to occur between the solidus and liquidus temperatures T_1 and T_2, respectively. The heat liberated by a solidifying

Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat, Table 3 Typical values or ranges for density ρ, isobaric specific heat capacity c_P, and thermal capacity ρ c_P of selected rocks at 20°C (Waples and Waples, 2004; Petrunin et al., 2004)

Rock	ρ (kg m^{-3})	c_P (J kg^{-1} K^{-1})	ρ c_P (kJ m^{-3} K^{-1})
Albite	2 540–2 560	755–780	1 922–1 991
Amphibole	3 010	700–1 134	2 110–3 410
Anhydrite	2 950–2 960	590–940	1 740–2 780
Anorthite	2 740	800	2 202
Basalt	2 870	880–900	2 526–2 583
Clay	2 680	860	2 300
Coal	1 350	1 300	1 760
Diabase	2 790	731–860	2 040–2 400
Dolomite	2 800	900	2 520
Gabbro	2 970–3 000	650–1 000	1 950–2 970
Gneiss	2 700	770–979	2 080–2 640
Granite	2 620–2 650	600–1 172	1 590–3 070
Gypsum	2 370	1 010	2 390
Limestone	2 760–2 770	680–880	1 880–2 430
Peridotite	2 740–3 190	705–1 005	1 930–3 210
Pyroxenite	3 190–3 240	660–1 000	2 140–3 190
Quartzite	2 640	731–1 013	1 930–2 670
Rock salt	2 160	880	1 900
Sandstone	2 640	775	2 050
Schist	2 770–2 900	790–1 096	2 190–3 180
Serpentinite	2 270–2 540	730–1 005	1 660–2 550
Siltstone	2 680	910	2 449
Slate	2 770–2 780	740–1 113	2 060–3 080
Syenite	2 820	460	1 300
Talc	2 780	1 000	2 780
Tuff	2 750	1 090	3 000

Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat, Table 4 Comparison of isobaric specific heat capacity c_P and latent heat of melting L of granite, basalt, seawater, and freshwater (Stacey and Davis, 2008, supplemented)

	Granite	Basalt	Seawater	Freshwater	Dry air (15°C, 1 atm)
c_P (kJ kg^{-1} K^{-1})	0.83	0.88	3.99	4.18	1.006
L (kJ kg^{-1})	420	420	335	333.55	196

("freezing") liquid phase is obtained by weighting by the volume fractions of liquid and solid phases, ϕ_l and ϕ_s, respectively. The enthalpy change of the rock volume then becomes $dH_{freezing} = (\phi_l\, c_l + \phi_s\, c_s)\, dT + L\, d\phi_l$, and the effective heat capacity c_{eff} is:

$$c_{eff}^{freezing} = \frac{dH}{dT} = \phi_l\, c_l + \phi_s\, c_s + L\, \frac{d\phi_l}{dT}. \quad (24)$$

Conversely, when considering melting the solid phase, the enthalpy change of the rock volume is $dH_{melting} = (\phi_l\, c_l + \phi_s\, c_s)\, dT + L\, d\phi_s$, and the effective heat capacity in this case c_{eff} is:

$$c_{eff}^{melting} = \frac{dH}{dT} = \phi_l\, c_l + \phi_s\, c_s + L\, \frac{d\phi_s}{dT}. \quad (25)$$

Acknowledgments

This contribution benefitted from insightful comments by an anonymous reviewer and editorial help by Dr. Sukanta Roy, Hyderabad.

Bibliography

Beck, A. E., 1988. Methods for determining thermal conductivity and thermal diffusivity. In Hänel, R., Rybach, L., and Stegena, L. (eds.), *Handbook of Terrestrial Heat Flow Density Determination*. Dordrecht: Kluwer, pp. 87–124.

Berman, R. G., 1988. Internally-consistent thermodynamic data for minerals in the system Na_2O-K_2O-CaO-MgO-FeO-Fe_2O_3-Al_2O_3-SiO_2-TiO_2-H_2O-CO_2. *Journal of Petrology*, **29**(2), 445–522.

Berman, R. G., and Brown, T. H., 1985. Heat capacity of minerals in the system Na_2O-K_2O-CaO-MgO-FeO-Fe_2O_3-Al_2O_3-SiO_2-TiO_2-H_2O-CO_2. *Contributions to Mineralogy and Petrology*, **89**, 168–183.

Bina, C. R., and Helffrich, G., 1994. Phase transition Clapeyron slopes and transition zone seismic discontinuity topography. *Journal of Geophysical Research*, **99**(B8), 15853–15860, doi:10.1029/94JB00462.

Birch, F., 1966. Section 7: compressibility; elastic constants. In Clark S. P., Jr. (ed.), *Handbook of Physical Constants.* Rev. ed., Memoir 97, Washington: Geologoical Society of America, pp. 97–173.

Brown, M. E., 2001. *Introduction to Thermal Analysis: Techniques and Applications (Hot Topics in Thermal Analysis and Calorimetry)*. Dordrecht: Kluwer.

Čermák, V., and Rybach, L., 1982. Thermal conductivity and specific heat of minerals and rocks. In Angenheister, G. (ed.), *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology*. Berlin: Springer. New Series, Vol. V(1a), pp. 305–343.

Clauser, C., 1992. Permeability of crystalline rocks. *EOS Transactions of the American Geophysical Union*, **73**(21), 233, 237. (for an update see also: Clauser, C., 2001. *Update of the permeability of crystalline rocks*, Report, Institute for Applied Geophysics and Geothermal Energy, E.ON Energy Research Center, RWTH Aachen University, http://www.eonerc.rwth-aachen.de/go/id/tsm/; retrieved 16 Oct 2010.)

Clauser, C., 2006. Geothermal Energy. In Heinloth, K. (ed.), *Landolt-Börnstein, Group VIII: Advanced Materials and Technologies*. Heidelberg: Springer. Energy Technologies, Subvol. C: Renewable Energies, Vol. 3, pp. 480–595.

Clauser, C., Griesshaber, E., and Neugebauer, H. J., 2002. Decoupled thermal and mantle helium anomalies – Implications for the transport regime in continental rift zones. *Journal of Geophysical Research*, **107**(B 11), 2269, doi:10.1029/2001JB000675.

CODATA (Committee on Data for Science and Technology), 2006. *CODATA Internationally Recommended Values of the Fundamental Physical Constants*, Paris: International Council of Scientific Unions (ICSU), http://physics.nist.gov/cuu/Constants/index.html, retrieved 9 Sept 2010.

Deuss, A., and Woodhouse, J., 2001. Seismic observation of splitting of the mid-transition zone discontinuity in Earth's mantle. *Science*, **294**, 354–357.

Dziewonski, A. M., and Anderson, D. L., 1981. Preliminary reference Earth model. *Physics of the Earth and Planetary Interiors*, **25**, 297–356.

Fei, Y., and Saxena, S. K., 1987. An equation for the heat capacity of solids. *Geochimica et Cosmochimica Acta*, **52**(2), 251–254.

Haines, P. (ed.), 2002. *Principles of Thermal Analysis and Calorimetry*, Cambridge, UK: The Royal Society of Chemistry.

Helffrich, G., 2000. Topography of the transition zone seismic discontinuities. *Reviews of Geophysics*, **38**, 141–158.

Hemminger, W. F., and Cammenga, H. K., 1989. *Methoden der Thermischen Analyse*. Berlin: Springer.

Holland, T. J. B., and Powell, R., 1996. An internally consistent thermodynamic data set for phases of petrological interest. *Journal of Metamorphic Geology*, **16**, 309–343.

Katsura, T., Yamada, H., Nishikawa, O., Song, M., Kubo, A., Shinmei, T., Yokoshi, S., Aizawa, Y., Yoshino, T., Walter, M. J., Ito, E., and Funakoshi, K-i, 2004. Olivine-wadsleyite transition in the system $(Mg, Fe)_2SiO_4$. *Journal of Geophysical Research*, **109**, B02209, doi:10.1029/2003JB002438.

Kelley, K., 1960. *Contributions to the Data on Theoretical Metallurgy: XIII High-Temperature Heat-Content, Heat-Capacity, and Entropy Data for the Elements and Inorganic Compounds*, U. S. Bureau of Mines Bulletin 584, Washington DC: U. S. Government Printing Office.

Kittel, C., 2004. *Introduction to Solid State Physics*, 8th edn. Hoboken: Wiley.

Maier, C. G., and Kelley, K. K., 1932. An equation for the representation of high temperature heat content data. *Journal of American Chemical Society*, **54**(8), 3243–3246.

Mottaghy, D. C., Schellschmidt, R., Popov, Y. A., Clauser, C., Kukkonen, I. T., Nover, G., Milanovsky, S., and Romushkevich, R. A., 2005. New heat flow data from the immediate vicinity of the Kola super-deep borehole: vertical variation in heat flow confirmed and attributed to advection. *Tectonophysics*, **401**(1–2), 119–142, doi:10.1016/j.tecto.2005.03.005.

Mottaghy, D., Vosteen, H.-D., and Schellschmidt, R., 2008. Temperature dependence of the relationship of thermal diffusivity versus thermal conductivity for crystalline rocks. *International Journal of Earth Sciences*, **97**, 435–442.

Olhoeft, G. R., and Johnson, G. R., 1989. Densities of rocks and minerals. In Carmichael, R. S. (ed.), *Practical Handbook of Physical Properties of Rocks and Minerals*. Boca Raton: CRC Press, pp. 139–176.

Petrunin, G. I., Popov, V. G., and Il'in, I. A., 2004. Conductive heat transfer in plagioclases. *Izvestiya, Physics of the Solid Earth (English Translation)*, **40**(9), 752–759.

Robertson, E. C., and Hemingway, B. S., 1995. *Heat Capacity and Heat Content of Rocks*, Open-file report 95-622. Reston, VA: U. S. Geological Survey.

Roy, R. F., Beck, A. E., and Touloukian, Y. S., 1981. *Thermophysical Properties of Rocks*, In Touloukian, Y. S., Judd, W. R., Roy, R. F., (eds.), *Physical Properties of Rocks and Minerals,* McGraw-Hill/CINDAS Data Series on Material Properties, New York: McGraw-Hill, Vol. II-2, pp. 409–502.

Somerton, W. H., 1992. *Thermal Properties and Temperature Related Behavior of Rock/Fluid Systems*. Amsterdam: Elsevier.
Stacey, F. D., and Davis, P. M., 2008. *Physics of the Earth*, 4th edn. Cambridge: University Press.
Tipler, P. A., and Mosca, G. P., 2007. *Physics for Scientists and Engineers: Extended Version*. Basingstoke: Palgrave Macmillan.
Waples, D. W., and Waples, J. S., 2004. A review and evaluation of specific heat capacities of rocks, minerals, and subsurface fluids. Part 1: Minerals and Nonporous rocks. *Natural Resources Research*, **13**(2), 97–122.
Whittington, A. G., Hofmeister, A. M., and Nabelek, P. I., 2009. Temperature-dependent thermal diffusivity of the Earth's crust and implications for magmatism. *Nature*, **458**, 319–321, doi:10.1038/nature07818.
Wohlenberg, J., 1982a. Density of minerals. In Hellwege, K.-H. (ed.), *Group V: Geophysics Vol 1: Physical Properties of Rocks, Subvol. A*. Berlin: Springer. Landolt-Börnstein, Vol. 1, pp. 66–113.
Wohlenberg, J., 1982b. Density of rocks. In Hellwege, K.-H. (ed.), *Group V: Geophysics Vol 1: Physical Properties of Rocks, Subvol. A*. Berlin: Springer. Landolt-Börnstein, Vol. 1, pp. 113–120.

Cross-references

Geothermal Record of Climate Change
Heat Flow, Continental
Heat Flow, Seafloor: Methods and Observations
Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity

THERMAL STORAGE AND TRANSPORT PROPERTIES OF ROCKS, II: THERMAL CONDUCTIVITY AND DIFFUSIVITY

Christoph Clauser
Institute for Applied Geophysics and Geothermal Energy, E.ON Energy Research Center, RWTH Aachen University, Aachen, Germany

Synonyms

Heat conductivity; Thermal conductivity; Thermal diffusivity

Definition

Thermal conductivity (also: heat conductivity) λ: Physical property governing heat diffusion in the steady state. It defines how much heat flows across a unit cross section of rock along a unit distance per unit temperature decrease per unit time; dimension: $W\,m^{-1}\,K^{-1}$.
Thermal diffusivity κ: Physical property governing transient heat diffusion. It is defined by the ratio of thermal conductivity and thermal capacity, i.e., by the ratio of heat flowing across the face of a unit volume and the heat stored in the unit volume per unit time; dimension: $m^2\,s^{-1}$.

Thermal conductivity

Fourier's law of heat conduction defines the vector of specific heat flow q_i, i.e., heat flow normalized by area, as the product of the thermal conductivity tensor λ_{ij} and the temperature gradient vector $\partial T/\partial x_j$:

$$q_i = -\lambda_{ij}\,\frac{\partial T}{\partial x_j}. \quad (1)$$

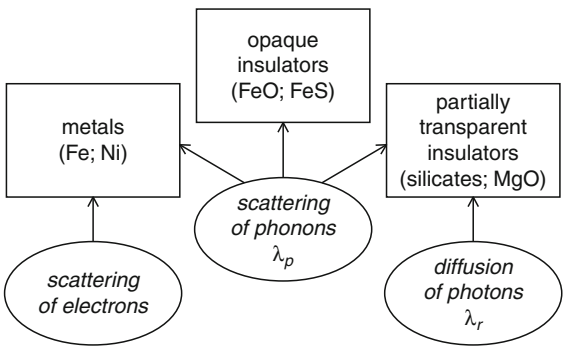
Temperature measurements are usually performed along vertical profiles in boreholes, yielding only the vertical component of the temperature gradient. Thermal conductivity in some rocks is, to a good approximation, isotropic, particularly in volcanic and plutonic rocks. Then heat will flow predominantly vertically, and it is sufficient to consider only the vertical component of Equation 1. In contrast, thermal conductivity of many sedimentary and metamorphic rocks is strongly anisotropic, and lateral heat flow may be significant. Hence, information on anisotropy is often required, demanding laboratory measurements in different directions. Anisotropy occurs over a wide range of scales, from microscopic, over laboratory and macroscopic, to tectonic.

Heat transport in most of the Earth's crust and mantle is diffusive and caused by (1) scattering of quantized lattice vibrations, the phonons, and (2) by diffusive (as opposed to ballistic) radiation of photons. These two processes are described by phonon thermal conductivity λ_p and radiative thermal conductivity, λ_r, respectively, the sum of which is often termed effective thermal conductivity, λ_{eff}. In most polycrystalline material heat radiation starts to dominate at temperatures above 2 500 K. In single crystals and glasses with little or no scattering (e.g., obsidian), however, radiation may become important at much lower temperatures of 500 k–1 000 K. In the metallic core, scattering of electrons provides a third heat transfer mechanism (Figure 1).

The following text deals mostly with phonon thermal conductivity. Radiative heat transfer is discussed in the paragraph on radiative thermal conductivity and in the chapter on thermal diffusivity.

Measuring techniques

Thermal conductivity can be measured in the laboratory on rock samples, i.e., cores or cuttings or in situ either in boreholes or with shallow penetration (3 m–20 m) marine



Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Figure 1 Different types of material in the Earth and associated mechanisms of heat transport (Redrawn after Hofmeister et al., 2009).

heat flow probes. There are numerous steady-state and transient techniques available for measuring thermal conductivity, the most prominent being the "divided bar," "needle probe," "optical scanning," and "(laser) flash," all of which are suitable to determine also the anisotropy of thermal conductivity. These methods are described in, for instance, Parker et al. (1961), Kappelmeyer and Haenel (1974), Beck (1988), Davis (1988), Somerton (1992), Schilling (1999), Popov et al. (1999b), Beardsmore and Cull (2001), and Blumm and Lemarchand (2002). Among these techniques, the transient ones are also suitable for determining thermal diffusivity.

As with most other petrophysical properties, in situ thermal conductivity may deviate significantly from laboratory values, even if the effects of temperature, pressure, and pore fluid are accounted for. This scale dependence involves different aspects: In situ measurements represent an average over a much larger rock volume than laboratory measurements performed on small samples, and cannot resolve small-scale variations. A subsequent upscaling may be necessary to identify the appropriate representative elementary volume (REV) for which reasonable transport parameter averages can be defined.

Indirect methods

When no data are available or no direct measurements can be performed, thermal conductivity can be inferred indirectly, either from mineralogical composition and saturating fluids or from correlations with other physical properties. While some of these methods are based on well-defined physical models, others are purely empirical.

Estimation from mineralogical composition and saturating fluids: Thermal conductivity of rocks may be estimated from their mineral and fluid content. Thermal conductivity of minerals varies much less than in rocks, due to their well-defined composition. As the bulk thermal conductivity of porous rocks varies with different saturants, it may be also of interest to know rock thermal conductivity for other saturants than those used in the laboratory measurement. Numerous models based on volume fractions of the individual mineral and fluid phases have been proposed, all with specific advantages and disadvantages: Some overestimate while others underestimate systematically the true bulk thermal conductivity. Most of them are valid only within a specific range of volume fractions (or porosities) and yield unreasonable results outside. To overcome this problem, additional parameters may be introduced in order to incorporate rock structure into a mixing law (see below).

The parallel and series models for thermal resistance of layered media are easy to understand, but have the disadvantage of being rather special cases, applicable mostly to bedded sediments. They correspond to the well-known weighted arithmetic and harmonic means, λ_{ari} and λ_{har}, Equations 2a and b, and are related to heat flowing parallel or perpendicular to bedding, respectively. They define upper and lower limits for all other models, sometimes also referred to as the Voigt upper bound and Reuss lower bound, respectively. Thus they constrain the maximum variance of possible predictions. The arithmetic mean between the two is known as the Voigt-Reuss-Hill average, λ_{VRH} (Equation 2c). It is useful for estimating an effective thermal conductivity, while the arithmetic and harmonic means are used to define a range for possible values.

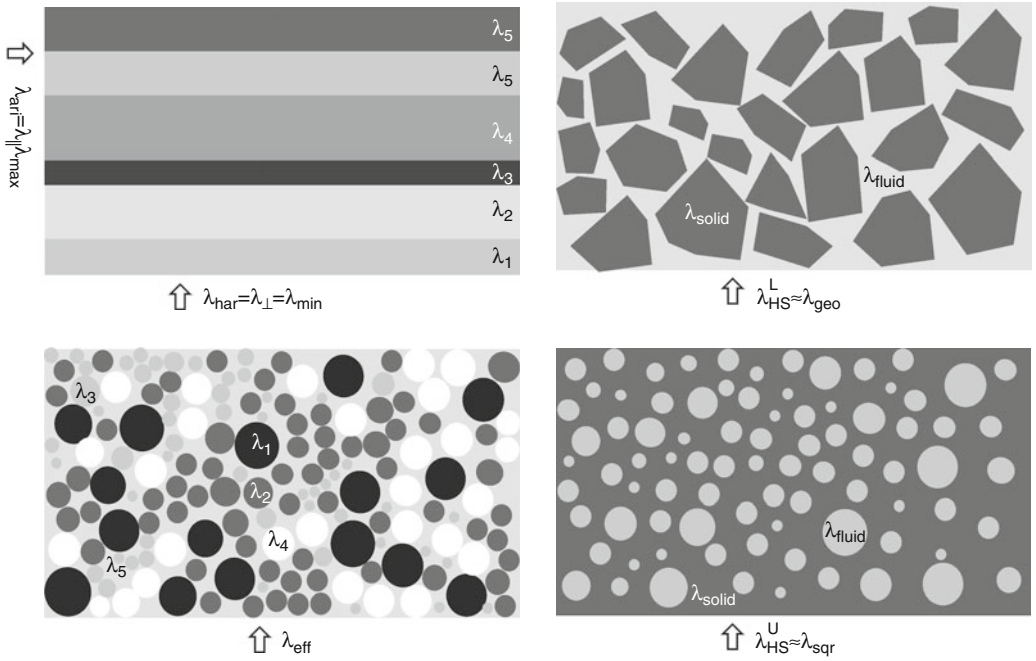
The weighted geometric and square root means λ_{geo} and λ_{sqr} (Equations 2d and e; e.g., Beardsmore and Cull, 2001), respectively, are associated with a mixture of different fluid and mineral phases of unspecified geometry. In many cases, both means successfully explain the data and are therefore widely used for lack of more specific information on the geometric arrangement of the individual volume fractions. In the special case where $N = 2$ and $n_1 = n_2 = 0.5$, the arithmetic, harmonic, and geometric means are related by $\lambda_{geo} = \sqrt{\lambda_{ari}\,\lambda_{har}}$.

Effective medium theory (Bruggeman, 1935) provides the effective medium mean λ_{eff} (Equation 2f) which is useful for macroscopically homogeneous and isotropic rocks consisting of randomly distributed grains and pores.

The upper and lower Hashin-Shtrikman bounds λ^U_{HS} and λ^L_{HS} (Equation 3, Hashin and Shtrikman, 1962), respectively, provide tighter constraints for the predictions of different models other than the arithmetic and harmonic means. The arithmetic average of both defines the Hashin-Shtrikman mean (Equation 2g). Geometrically, the lower Hashin-Shtrikman bound λ^L_{HS} corresponds to a rock model consisting of grains suspended in a fluid and is closely followed by the geometric mixing law. In contrast, the square root law is very close to the upper Hashin-Shtrikman bound λ^U_{HS} and could be related to a well-lithified rock with spherical, fluid-filled pores. Figure 2 illustrates the geometries corresponding to all of these models.

If λ_i denotes the thermal conductivity and n_i the volume fraction of the ith phase relative to the total volume ($1 = \sum n_i$), these seven weighted means are defined by:

$$
\begin{aligned}
&(a)\ \lambda_{\max} = \lambda_{ari} = \lambda_{||} = \sum_{i=1}^{N} n_i\ \lambda_i;\\
&(b)\ \lambda_{\min} = \lambda_{har} = \lambda_{\perp} = \left(\sum_{i=1}^{N} \frac{n_i}{\lambda_i}\right)^{-1};\\
&(c)\ \lambda_{VRH} = \frac{1}{2}\left(\lambda_{||} + \lambda_{\perp}\right);\ (d)\ \lambda_{geo} = \prod_{i=1}^{N} \lambda_i^{n_i};\\
&(e)\ \sqrt{\lambda_{sqr}} = \sum_{i=1}^{N} n_i\ \sqrt{\lambda_i};\ (f)\ \lambda_{eff}^{-1} = \sum_{i=1}^{N} \frac{3\ n_i}{2\ \lambda + \lambda_i};\\
&(g)\ \lambda_{HS} = \frac{1}{2}\left(\lambda^U_{HS} + \lambda^L_{HS}\right);
\end{aligned}
\tag{2}
$$



Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Figure 2 Geometrical arrangement of layers, mineral grains, and pores assumed in different models for calculating mean bulk thermal conductivity of a composite medium: arithmetic (λ_{ari}), harmonic (λ_{har}), geometric (λ_{geo}), square root (λ_{sqr}), Hashin-Shtrikman upper (λ^U_{HS}) and lower (λ^L_{HS}) bounds, and effective medium (λ_{eff}).

where:

$$\begin{aligned}
\lambda^U_{HS} &= \lambda_{\max} + \frac{A_{\max}}{1-\alpha_{\max} A_{\max}},\\
\text{with: } A_{\max} &= \sum_{i=1;\ \lambda_i \neq \lambda_{\max}}^{N} \frac{n_i}{\alpha_{\max} + 1/(\lambda_i - \lambda_{\max})};\\
\lambda_{\max} &= \max(\lambda_1, \ldots, \lambda_N);\ \alpha_{\max} = \frac{1}{3\,\lambda_{\max}}\\
\lambda^L_{HS} &= \lambda_{\min} + \frac{A_{\min}}{1-\alpha_{\min} A_{\min}},\\
\text{with: } A_{\min} &= \sum_{i=1;\ \lambda_i \neq \lambda_{\min}}^{N} \frac{n_i}{\alpha_{\min} + 1/(\lambda_i - \lambda_{\min})};\\
\lambda_{\min} &= \min(\lambda_1, \ldots, \lambda_N);\ \alpha_{\min} = \frac{1}{3\,\lambda_{\min}}.
\end{aligned} \tag{3}$$

For a two-component system consisting of pore fluid and solid rock with thermal conductivities λ_f and λ_s, respectively, Equation 3 simplifies to (Hashin and Shtrikman, 1962; Horai, 1971):

$$\lambda^U_{HS} = \lambda_s + \frac{\phi}{\frac{1}{\lambda_f - \lambda_s} + \frac{1-\phi}{3\,\lambda_s}};\ \lambda^L_{HS} = \lambda_f + \frac{1-\phi}{\frac{1}{\lambda_s - \lambda_f} + \frac{\phi}{3\,\lambda_f}}. \tag{4}$$

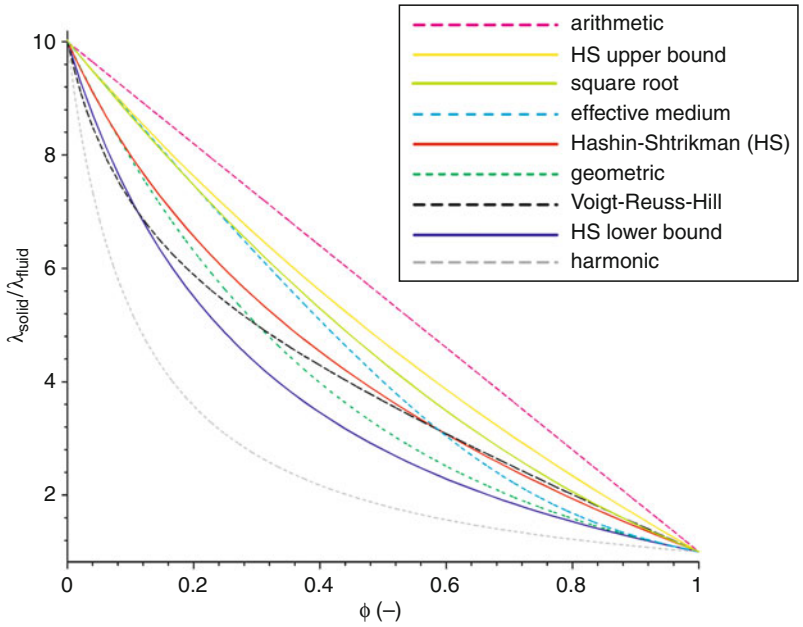
Generally, for a two-component system consisting of pore fluid and solid rock with thermal conductivities λ_f and λ_s, respectively, the implicit definition of λ_{eff} in Equation 2f can be resolved:

$$\lambda_{eff} = \frac{1}{4}\left\{ 3\,\phi(\lambda_f - \lambda_s) + 2\,\lambda_s - \lambda_f + \sqrt{\begin{array}{l} 9\,\phi^2\,\lambda_s^2 + 18\,\phi\,\lambda_s\,\lambda_f - 18\,\phi^2\,\lambda_s\,\lambda_f - 12\,\phi\,\lambda_s^2 \\ + \lambda_f^2 - 6\,\phi\,\lambda_f^2 + 4\,\lambda_s\,\lambda_f + 9\,\phi^2\,\lambda_f^2 + 4\,\lambda_s^2. \end{array}} \right\}. \tag{5}$$

Figure 3 compares the variation of thermal conductivity of a two-component system with volume fraction for these different mixing laws defined by Equations 2–5. By and large, and in particular for such a two-component system, thermal conductivity of a multiphase rock determined according to these models can be ordered as:

$$\begin{aligned}
\lambda_\perp = \lambda_{har} < \lambda^L_{HS} < \lambda_{VRH} < \lambda_{geo} < \lambda_{HS} < \lambda_{eff} \\
< \lambda_{sqr} < \lambda^U_{HS} < \lambda_{ari} = \lambda_{||}.
\end{aligned} \tag{6}$$

While only these nine models are presented and discussed here, various other mixing models are available which take into account additional factors, such as the shape of grains and voids. Several models assume spheroidal pores specifying the aspect ratio of the spheroids (Korvin, 1978, 1982; Schulz, 1981; Zimmerman, 1984, 1989; Buntebarth and Schopper, 1998; Popov et al.,



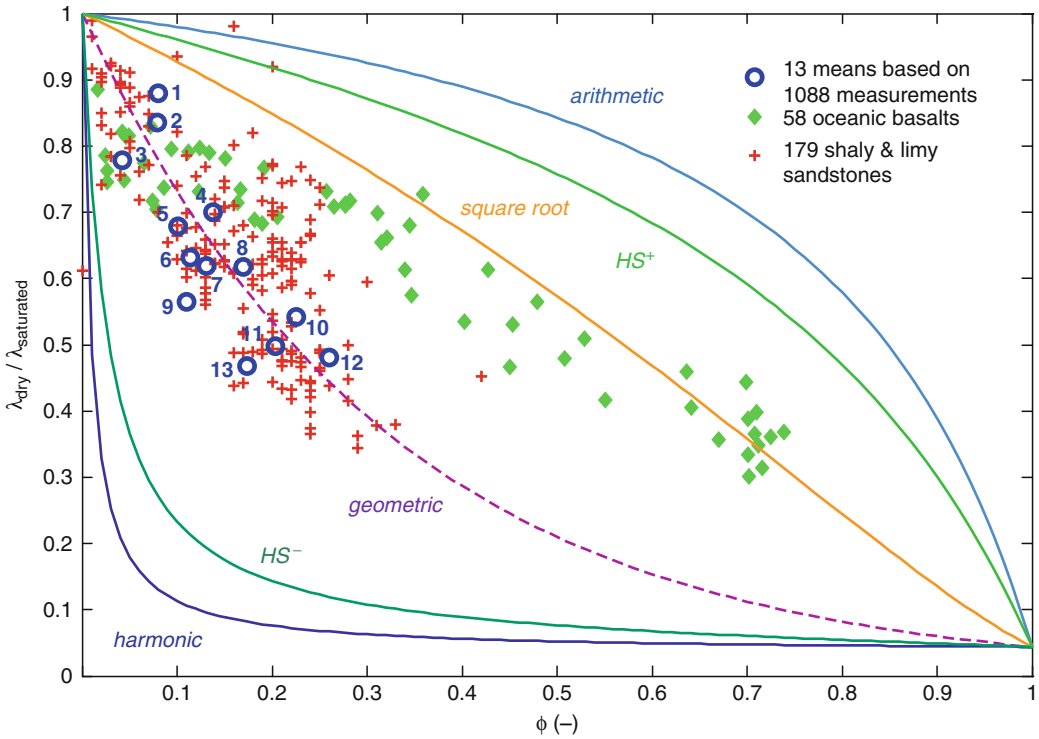
Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Figure 3 Variation of thermal conductivity λ of a two-phase rock with porosity ϕ according to the means in Equations 2–5 for solid and fluid thermal conductivities of $\lambda_{solid} = 6\,W\,m^{-1}\,K^{-1}$ and $\lambda_{fluid} = 0.6\,W\,m^{-1}\,K^{-1}$, respectively: arithmetic (λ_{ari}); Hashin-Shtrikman upper bound (λ^{U}_{HS}); square root (λ_{sqr}); effective medium (λ_{eff}); Hashin-Shtrikman (λ_{HS}); geometric (λ_{geo}); Voigt-Reuss-Hill average (λ_{VRH}); Hashin-Shtrikman lower bound (λ^{L}_{HS}); harmonic (λ_{har}).

2003). These models require information on the geometry of the internal rock structure and differ in the way of averaging over a representative elementary volume. Horai (1991) tested the results of predictions from several different mixing models on a remarkable data set in which porosity virtually varies from 0 % to 100 %. As can be expected, most of the models tested were valid only for certain porosity ranges. Only the two-phase models of Fricke-Zimmerman (Fricke, 1924; Zimmerman, 1989) and Schulz (1981) treating pores as spheroidal inclusions in a homogeneous and isotropic material are valid over the entire range of porosity. However, they require additional information on the spheroids' aspect ratio or orientation. Given the typical ratios of rock and fluid conductivities we observe in nature, i.e. less than 10, most of the conductivity models tested work to within an accuracy of 10 %–15 %.

Based on 1 325 individual measurements on sedimentary rocks and oceanic basalts in dry and saturated condition (λ_{dry}, λ_{sat}), Figure 4 compares the variation of the ratio $\lambda_{dry}/\lambda_{sat}$ with porosity measured with that predicted by the arithmetic, upper Hashin-Shtrikman, square root, geometric, lower Hashin-Shtrikman, and harmonic mixing laws, λ_{ari}, λ^{U}_{HS}, λ_{sqr}, λ_{geo}, λ^{U}_{HS}, and λ_{har}, respectively (Equations 2–4). It is evident that no single mixing law applies equally well to all rock types. But it appears that the geometric and square root model, on average, provide the best general fits to the data. It is also clear that the upper and lower Hashin-Shtrikman means, $\lambda^{L}_{HS}$and λ^{U}_{HS}, provide much tighter bounds than the arithmetic and harmonic means, λ_{ari} and λ_{har}, respectively.

Correlations with other physical properties: Physical properties measured in well-logs can be used to infer estimates for in situ thermal conductivity. The approach is based on an extension of the mixing-model approach to the borehole scale: The volume fractions n_i of the N different mineral (or fluid) phases are either taken directly from induced gamma ray spectroscopy logs (Williams and Anderson, 1990) or determined from a joint analysis of a suitable number J of geophysical logs, such as gamma ray (GR), sonic slowness (DT, the inverse of velocity), gamma density (DEN), and neutron porosity (NPHI) (e.g., Hartmann et al., 2005; Goutorbe et al., 2006). Let $\mathbf{x}$ and $\mathbf{b}$ be vectors composed of the N volume fractions n_i and the J theoretical log responses R^j with respect to the N different phases, respectively. Then, each element R^j of vector $\mathbf{b}$ is the cumulative response of the *j*th log to all phases weighted by their volume fractions:

$$R^j = \sum_i^N n_i R_i^j, \text{ where: } \sum_i^N n_i = 1, \text{and}$$
$$\mathbf{x} = [n_1, \ldots, n_N]^T, \mathbf{b} = [R^1, \ldots, R^J]^T. \quad (7)$$



Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Figure 4 Variation of thermal conductivity ratio $\lambda_{dry}/\lambda_{sat}$ (measured in dry and saturated condition) with porosity ϕ for different rock types. Numbered open circles represent means based on measurements on 1 058 sedimentary rock samples (Kobolev et al., 1990; Popov et al., 1995, 1999a; see also Clauser, 2006): (1) 21 limestones; (2) 54 limestones; (3) 13 quartz sandstones; (4) 44 quartz silt-stones; (5) 35 conglomerates; (6) 141 quartz sandstones; (7) 33 claystones; (8) 99 polymictic sandstones; (9) 30 quartz sandstones; (10) 22 claystones; (11) 65 quartz silt-stones; (12) 99 quartz silt-stones; (13) 241 quartz silt-stones. Shown for comparison are data measured on 58 oceanic basalts (diamonds) and 179 shaly and limy sandstones (crosses). Curves labeled arithmetic, HS^+, square root, geometric, HS^-, and harmonic, correspond to the arithmetic, upper Hashin-Shtrikman, square root, geometric, lower Hashin-Shtrikman, and harmonic mixing laws, λ_{ari}, λ^U_{HS}, λ_{sqr}, λ_{geo}, λ^L_{HS}, and λ_{har}, respectively (Equations 27–29) (Clauser, 2006).

The rows of matrix $\mathbf{A}$ contain the specific responses of each log to the N rock phases:

$$\mathbf{A} = \begin{bmatrix} R_1^1 & \cdots & R_N^1 \\ \vdots & \ddots & \vdots \\ R_1^J & \cdots & R_N^J \end{bmatrix}, \tag{8}$$

and the direct and inverse problems can then be written as

$$\mathbf{Ax} = \mathbf{b} \text{ and } \mathbf{x} = \mathbf{A}^{-1}\,\mathbf{b}, \tag{9}$$

respectively. Thus, in the direct problem, the log response vector $\mathbf{b}$ is computed from the volume fraction vector $\mathbf{x}$ and the specific log response matrix $\mathbf{A}$. In contrast, in the inverse problem, the volume fractions $\mathbf{x}$ are computed from the log responses $\mathbf{b}$ and the inverse of the specific log response matrix, $\mathbf{A}^{-1}$. Thus, for N solid rock constituents, solving the inverse requires $N - 1$ logs. Porosity does not count here, because it follows as the difference of one and the sum of the solid rock volume fractions. If more logs are available, making the problem over-determined, the inverse problem can also be solved in a least-squares sense. Once the volume fractions are known and assigned appropriate thermal conductivities, an appropriate mixing model can applied to compute rock thermal conductivity. Generally, the geometric and square root means, Equations 2d and e, often have turned out useful, but other mixing models may be appropriate in specific cases.

Assigning representative thermal conductivities to the solid rock constituents is not trivial. Tabulated values of rock thermal conductivity should be used only if they characterize specimens from the logged formations. In all other cases, these formations or their outcrops need to be sampled and these specimens tested in the laboratory. If measurements are performed at ambient conditions, the values need to be corrected for the effect of temperature, and in some cases for pressure as well.

In general, the effect of temperature is more pronounced than that of pressure. However, for greater depth

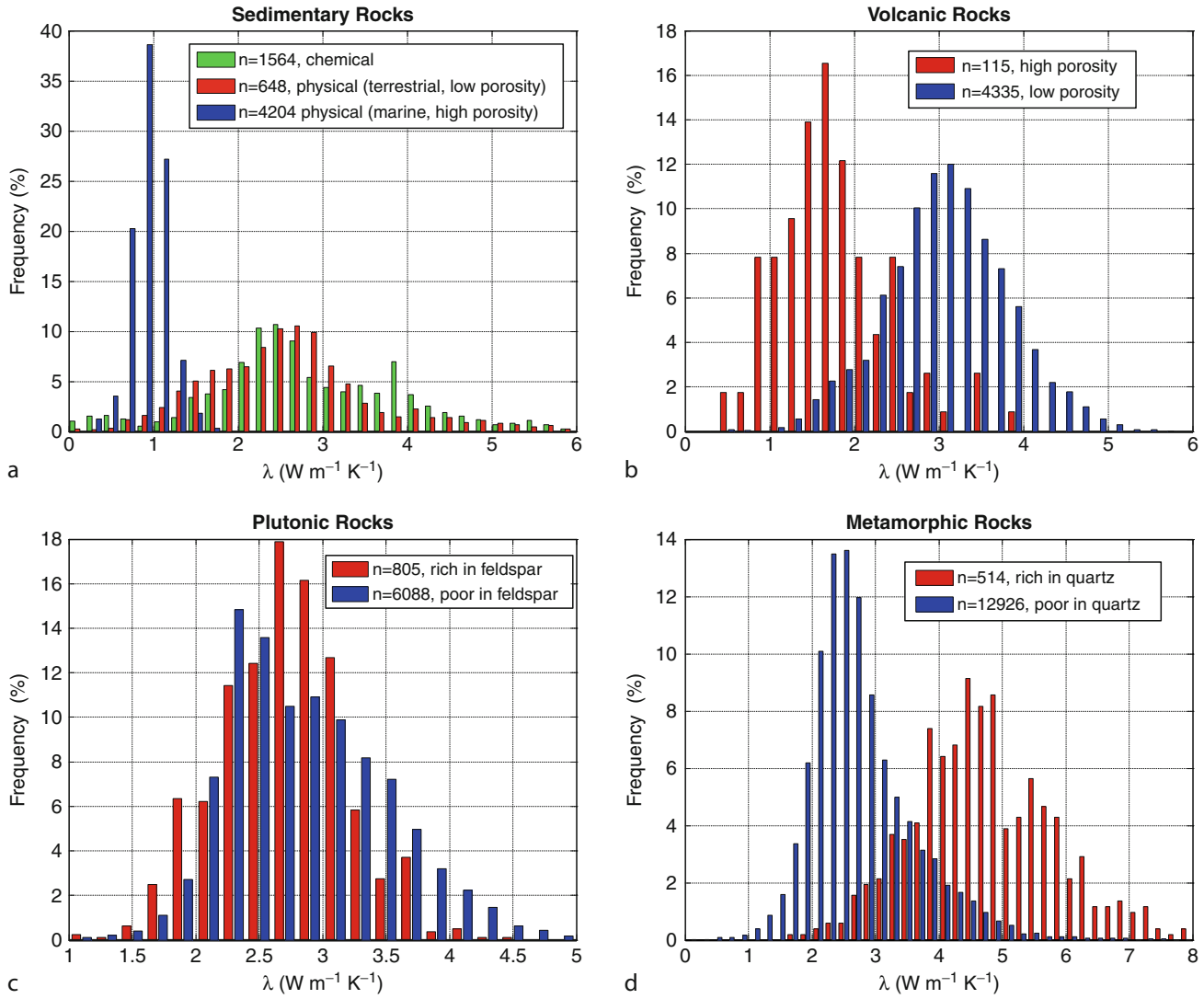
and little or less consolidated rocks it also needs to be accounted for. If commercial log interpretation software is used to perform the inversion, the theoretical log responses R_i^j with respect to the different rock constituents are usually supplied by the software. Alternatively, values for the theoretical log responses R^j can be obtained from the literature (e.g., Crain, 1986).

Thermal conductivity of minerals and rocks

Thermal conductivity of minerals is much better constrained than that of rocks, due to the well-defined crystal structure and chemical formula for each mineral. Substantial collections of mineral thermal conductivities were compiled by Birch (1942), Clark (1966), Horai and Simmons (1969), Dreyer (1974); Roy et al. (1981), Čermák and Rybach (1982), Carmichael (1984), Popov et al. (1987), Diment and Pratt (1988); Somerton (1992), Clauser and Huenges (1995), and Clauser (2006).

Rocks are less well defined. In fact, rock type is a rather poor descriptor for physical properties as it rarely characterizes the dominating factors for any given property. With respect to thermal conductivity, these comprise mineral content, porosity, pore fluid, saturation, and anisotropy for each rock type. As these factors are variable for each rock, the variation of rock thermal conductivity was characterized in a statistical manner in Clauser (2006, 2009) according to the four main diagenetic classes of rocks: sedimentary, volcanic, plutonic, and metamorphic. This discussion is summarized here.

For *sedimentary rocks* (Figure 5a), thermal conductivity is mainly controlled by porosity and sediment type.



Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Figure 5 Histograms of thermal conductivity for (**a**) sedimentary, (**b**) volcanic, (**c**) plutonic, and (**d**) metamorphic rocks (Clauser, 2009).

For *volcanic rocks* (Figure 5b), porosity is the controlling factor on thermal conductivity: Both mean and median of the high- and low-porosity histograms differ by nearly a factor of two, and the high-porosity distribution is skewed toward low conductivities. Plutonic and metamorphic rocks are generally much less porous. Here, the controlling factor is the dominant mineral phase. For *plutonic rocks* (Figure 5c), the feldspar content determines the shape of the histogram: Rocks with a high feldspar content (i.e., >60 %) display a nearly symmetrical conductivity distribution about a lower mean conductivity than rocks with low feldspar content. In spite of these differences, the means and medians for both distributions are nearly identical. For *metamorphic rocks* (Figure 5d), it is the quartz content which controls thermal conductivity: Both mean and median of the distributions for high and low quartz content differ by nearly a factor of two, similar as for *volcanic rocks* (Figure 5b) with regard to porosity. While the histogram for high quartz content rocks (mostly quartzites) is nearly symmetrical, the low quartz content histogram is strongly skewed toward low conductivities.

Radiative thermal conductivity

In addition to heat conduction, heat radiation emitted from hot bodies propagates through a sufficiently transparent and little absorbing medium as electromagnetic waves with velocity

$$c_m = c_0/n = \Lambda_m \nu_m, \tag{10}$$

where Λ_m and ν_m are radiation frequency and wavelength in the medium, and $n = c_0/c_m$ is the ratio of the speed of light in vacuum and in the medium, i.e., the real part of the index of refraction (Table 1).

Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Table 1 Refractive index n of selected substances

Substance	$n = c_0/c_m$ (at 589 nm)
Air	1.0003
Water(H_2O)	1.33
Magnesium fluoride (MgF_2)	1.38
Calcium fluoride (CaF_2)	1.43
Rock salt (NaCl)	1.54
Quartz (SiO_2)	1.54
Glass	1.5–1.6
Carbon disulfide (CS_2)	1.63
Corundum (Al_2O_3)	1.76
Zircon ($ZrSiO_4$)	1.92
Sulfur (S)	2.00
Sinc sulfide (ZnS)	2.37
Diamond	2.42
Carborundum (SiC)	2.55
Rutile (TiO_2)	3.10
Galena (PbS) (@ 590 nm)	3.90

Radiation theory is based on Planck's distribution law defining the spectral radiance L_Λ or L_ν of a black body in the wavelength or frequency interval $d\Lambda_m$ or $d\nu_m$ in the medium in terms of the corresponding wavelength or frequency in vacuum, respectively:

$$L_\Lambda(\Lambda_m, T)\, d\Lambda_m = \frac{2hc_0^2 n^2(\Lambda, T)}{\Lambda^5 \left(e^{hc_0/(\Lambda kT)} - 1\right)} d\Lambda \quad (\mathrm{W\,m^{-2}\,sr^{-1}\,m^{-1}} = \mathrm{W\,m^{-3}\,sr^{-1}}),$$

$$L_\nu(\nu_m, T)\, d\nu_m = \frac{2h\nu^3 n^2(\nu, T)}{c_0^2 \left(e^{h\nu/(kT)} - 1\right)} d\nu \quad (\mathrm{W\,m^{-2}\,sr^{-1}\,Hz^{-1}} = \mathrm{W\,s\,m^{-2}\,sr^{-1}}), \tag{11}$$

where, $h = 6.626\,068\,96(33) \times 10^{-34}$ J s is the Planck constant, $k = 1.380\,6504(24) \times 10^{-23}$ J K^{-1} the Boltzmann constant, and $d\Lambda = n\, d\Lambda_m$. Radiance is the radiant energy flux emitted per unit solid angle in a given direction per unit projected area of a source. According to Wien's displacement law, the maximum of spectral radiance increases with temperature and decreases with wavelength (or increases with frequency) of radiation. Hofmeister (2005) estimated the variation of n(ν,T) with temperature and frequency and found both to be very small suggesting that n is approximately constant in the visible frequency range.

The intensity of radiation, I, is related to the incident intensity, I_0, radiation path, x, and opacity, ε, by $I = I_0 \exp(-\varepsilon\, x)$, where opacity $\varepsilon = 1/\ell$ is the reciprocal mean free path ℓ of radiation. It defines the average travel distance of photons before being absorbed or scattered. In general, opacity is a function of the radiation wavelength. Opacity in an absorbing and scattering medium comprises contributions from both of these processes:

$$\varepsilon = \alpha + \varsigma, \tag{12}$$

where α and ς are absorption and scattering coefficients, respectively. The scattering coefficient ς is usually identified with the inverse grain size, suggesting orders of magnitude ranging from 10^2 m to 10^6 m^{-1}. The absorption coefficient α may be as large as $\alpha = 7{,}000\,\mathrm{m^{-1}}$, but values below $\alpha = 1\,500\,\mathrm{m^{-1}}$ are reported for absorption coefficients in single crystal olivines with different proportions of forsterite (Mg_2SiO_4) and fayalite ($Fe_2^{2+}SiO_4$) ($Fo_{94}Fa_6$ – $Fo_{86}Fa_{14}$) in the two pass bands at 0.5 µm and 0.5 µm–6.0 µm (e.g., Shankland et al., 1979; Clauser, 1988; Hofmeister, 2005). The width and level of these pass bands depend critically on the iron content in the minerals. The complex index of refraction of all materials, m, is defined by $m = \sqrt{\varepsilon_r \mu_r} = n - i \cdot k$, where ε_r and μ_r are (complex) relative electrical permittivity and relative magnetic permeability, respectively, $n = c_0/c_m$ is the real and k the imaginary part of the index (sometimes also called extinction coefficient), and

$i = \sqrt{-1}$ is the imaginary unit. The absorption coefficient α is related to the complex index of refraction k by

$$\alpha = 4\,\pi\,k/\Lambda, \tag{13}$$

where Λ is the radiation wavelength (Aronsson et al., 1970).

Radiation therefore becomes important for rocks with a larger free mean path of radiation corresponding to smaller values of opacity or absorption coefficient and scattering coefficient. This holds in particular at larger wavelength, in the infrared part of the absorption spectrum's transmission window between about 0.5 μm–6.0 μm.

Radiated heat is diffused if photons emitted by mineral grains are scattered or absorbed by neighboring grains. If the mean free path of radiation, ℓ, is small compared to the distance to material discontinuities (such as grain boundaries) and for moderate temperature gradients, i.e., no large anisotropy in the intensity of radiation, an effective thermal conductivity

$$\lambda_{\mathrm{eff,ij}} = \lambda_{\mathrm{p,ij}} + \lambda_{\mathrm{r,ij}} \tag{14}$$

replaces λ_{ij} in Equation 1, where $\lambda_{p,ij}$ and $\lambda_{r,ij}$ are phonon and radiative thermal conductivities, respectively.

Rosseland (1930) and Clark (1957) defined the radiative thermal conductivity based on the temperature derivative of the spectral radiance of a blackbody. Following practice in engineering, Hofmeister (2005) accounts in this expression additionally for the emissivity $\eta \leq 1$ of a grainy material by applying the temperature derivative and the subsequent integration to the product of emissivity η and spectral radiance L_ν, yielding

$$\lambda_{\mathrm{r}} = \frac{4\,\pi}{3}\int_0^\infty \frac{1}{\varepsilon(\nu,T)}\,\frac{\partial(\eta(\nu,T)\;L_\nu(\nu,T))}{\partial T}\,\mathrm{d}\nu. \tag{15}$$

Different practical expressions were derived depending on which parameters can be assumed as independent of temperature or frequency (see, e.g., Clark (1957); Shankland et al. (1979); Schärmeli, 1982; Clauser, 1988; Hofmeister, 1999, 2005; Hofmeister et al., 2009). Based on grain size δ and attenuation

$$\mathrm{A} = (\ln \mathrm{I}_0 - \ln \mathrm{I}_\mathrm{t})/\delta \approx \alpha, \tag{16}$$

where α is the absorption coefficient and I_0 and I_t are incident and transmitted intensities, respectively, Hofmeister (2005) suggests the following expressions for emissivity and opacity:

$$\eta = 1 - e^{-\delta\,\alpha(\nu)} = 1 - e^{-(\ln \mathrm{I}_0 - \ln \mathrm{I}_\mathrm{t})}, \quad \varepsilon = \frac{1+\delta\,\alpha}{\delta}, \tag{17}$$

and estimates values for the product $\delta\,\alpha$ based on the interface reflectivity R (%) between two neighboring grains. Values for $\delta\,\alpha$ considered realistic for the mantle range between 5 and 10 (corresponding to a range for R between 0.07 % and 0.05 %) with a preferred value of 7 (R = 0.01 %).

The "gray body" approximation assumes opacity and emissivity as finite, constant, and independent of radiation wavelength. If the real part of the refractive index, n, and the spectral radiance, L_Λ or L_ν, are also independent of temperature and wavelength, Equation 15 simplifies into:

$$\lambda_{\mathrm{r}} = \frac{16\,\eta\,\sigma\,n^2\,T^3}{3\,\varepsilon}, \tag{18}$$

where σ = 5.670 400(40) × $10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}}$ is the Stefan-Boltzmann constant. An example of magnitude is obtained when emissivity η = 0.99 and opacity ε is identified with the olivine ($Fo_{92}Fa_8$) absorption coefficient α at 1 700 K (Equation 12, neglecting contributions from scattering), with $1\,000\,\mathrm{m^{-1}} < \alpha < 1\,500\,\mathrm{m^{-1}}$. Additionally, a typical silicate value is assumed for the index of refraction, n = 1.6 (Table 1). This yields a range for radiative thermal conductivity at 1 700 K of $3.8\,\mathrm{W\,m^{-1}\,K^{-1}} > \lambda_\mathrm{r} > 2.5\,\mathrm{W\,m^{-1}\,K^{-1}}$.

Inserting Equations 16 and 17, and the temperature derivative of the spectral radiance (11) as provided by Shankland et al. (1979) into (15) yields radiative thermal conductivity as:

$$\begin{aligned}\lambda_\mathrm{r} &= \frac{4\,\pi\,\delta}{3}\int_0^\infty \frac{1-e^{-\delta\,\alpha(\nu)}}{1+\delta\,\alpha(\nu)}\,\frac{\partial(L_\nu(\nu,T))}{\partial T}\,\mathrm{d}\nu \\ &= \frac{8\,\pi\,\delta\,\mathrm{n\,T^3\,k^4}}{3\,c_0^2\,h^3}\sum\int_{lower}^{upper}\frac{1-e^{-\delta\,\alpha(\nu)}}{1+\delta\,\alpha(\nu)} \\ &\times\left[\frac{n\,e^x\,x^4+}{e^x-1}+\frac{2\,T\,x^3}{e^x-1}\,\frac{\partial n}{\partial T}\right]dx\;(\mathrm{W\,m^{-1}\,K^{-1}}),\end{aligned} \tag{19}$$

where x = hν/(k T) and dx = h dν/(k T), and the summation allows for the transparent regions above and below the strong absorption bands in the visible part of the spectrum. The second term of the sum in the integrand vanishes if $\partial n/\partial T \approx 0$ as suggested by Hofmeister (2005). Because α varies nonlinearly with frequency or wavelength, and the cutoff frequencies in the integral in (19) depend on δ and α, Hofmeister (2005) evaluated the integral numerically for polynomials in T whose exponents ranged from 0 to 6, thus yielding other than a purely cubic relationship between radiative thermal conductivity and temperature.

Variation with temperature

Since the pioneering experiments of Eucken (1911), thermal conductivity of minerals and rocks is known to decrease with temperature, generally with its inverse. Eucken's empirical result was corroborated theoretically by Peierls (1929) based on Debye's (1914) theory of phonon scattering. Eucken (1911) observed in his experiments

in contrast to crystals an increase in thermal conductivity of amorphous siliceous glass, a clear indication of radiative heat transfer.

The decrease is primarily due to the decrease of phonon (or lattice) thermal conductivity λ_p with temperature and to a smaller degree to thermal cracking. Since the thermal expansion coefficient increases with temperature (but differently for all minerals) differential expansion may create contact resistances between mineral grains. The effect of contact resistance is less pronounced in water-saturated than in dry rocks, the condition in which most rocks are tested at elevated temperatures. For single-mineral aggregates, a linear relationship between temperature and thermal resistivity, λ^{-1}, discriminates between contributions which depend on temperature T and others which do not, such as micro-cracks, grain boundaries, shape, and orientation of crystals and their fragments:

$$\lambda^{-1}(T) = c_1 + c_2\ T, \tag{20}$$

where λ is in $W\,m^{-1}\,K^{-1}$ and T is in K. By measuring thermal conductivity λ and plotting its inverse, thermal resistivity, λ^{-1}, versus temperature, constants c_1 and c_2 are obtained from intercept and slope of a linear regression. Table 2 provides values for the constants c_1 and c_2 in Equation 20 which may be used to infer the temperature dependence of thermal resistivity for some single-mineral aggregates (Clark, 1969).

Based on measurements on 113 samples of metamorphic rocks from the KTB research borehole in Germany in the temperature range 50 °C–200 °C, Buntebarth (1991) determined mean values for the constants c_1 and c_2 in Equation 20 for gneissic and metabasitic rocks. The arithmetic means determined from measurements on 66 gneiss samples are $\bar{c}_1 = 0.16(3)\,W^{-1}\,m\,K$ and $\bar{c}_2 = 0.37(14) \times 10^{-3}\,W^{-1}\,m$. The corresponding means determined from measurements on 36 metabasite samples are $\bar{c}_1 = 0.33(3)\,W^{-1}\,m\,K$ and $\bar{c}_2 = 0.22(14) \times 10^{-3}\,W^{-1}\,m$.

In contrast to phonon conductivity λ_p, the radiative contribution to thermal conductivity, λ_r, generally increases with the cube of temperature (see above). Thus, measurements of thermal conductivity as function of temperature generally first exhibit a decrease with temperature until, from about 1 000 °C–1 200 °C onward, the radiative component balances and sometimes even reverses the decreasing trend.

The temperature dependence of rock thermal conductivity was characterized in a statistical manner in Clauser (2006, 2009) according to the four main diagenetic classes of rocks: sedimentary, volcanic, plutonic, and metamorphic. This discussion is summarized here (Figure 6).

For *sedimentary rocks* (Figure 6a) up to 300 °C there is a reduction by nearly a factor of two, both for physical and chemical sediments. Above 300 °C, the decrease in thermal conductivity is less, but it is stronger for chemical sediments than for physical sediments. However, there are very few data for this temperature range, which makes this last observation statistically weak. Above 300 °C, the mean thermal conductivity of sediments varies between $1.0\,W\,m^{-1}\,K^{-1}$ – $1.5\,W\,m^{-1}\,K^{-1}$.

Volcanic rocks (Figure 6b) vary quite differently with temperature depending on their opacity, i.e., on how well they transmit thermal energy by radiation. Due to this additional "radiative thermal conductivity," volcanic glasses and rocks with small iron content experience an increase in thermal conductivity for temperatures above 800 °C–1 000 °C (e.g., Clauser, 1988; Hofmeister et al., 2009). In contrast, thermal conductivity of conduction dominated rocks, such as rocks with high iron content, decreases with temperature. An inversion of this trend is indicated by few available high-temperature measurements (above 1 300 °C) but with too few measurements for a statistical appraisal. At about 1 000 °C thermal conductivity for these rocks is at about 50 % of the room-temperature value. Again, there are few data points above 700 °C.

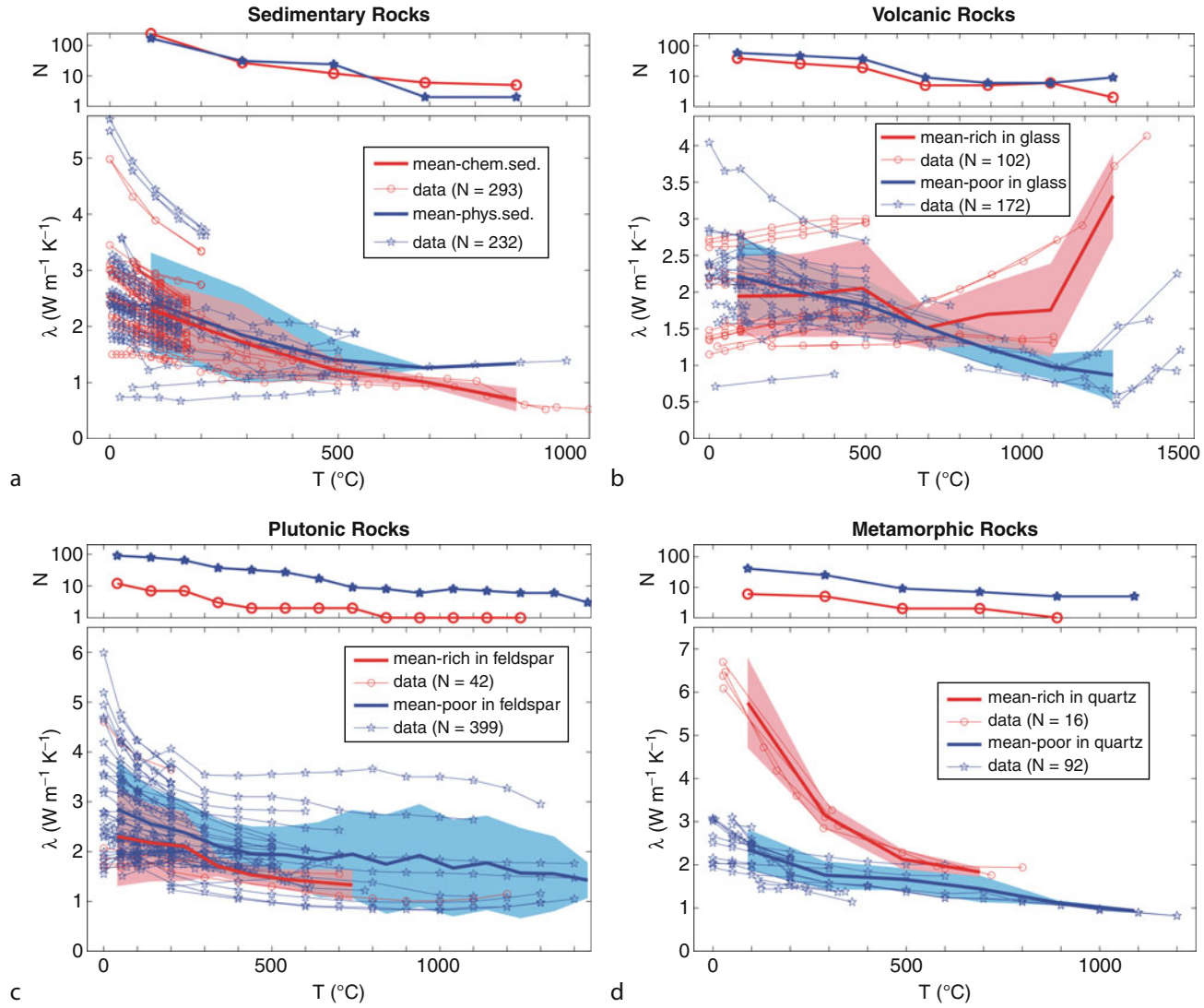
P*lutonic rocks* (Figure 6c) show no strong radiative contribution. At temperatures above 600 °C, thermal conductivity decreases only very little. However, in these rocks the variation of thermal conductivity with temperature depends on their feldspar content. For rocks enriched in feldspar, thermal conductivity decreases little up to 300 °C, while for those poor in feldspar the decrease is stronger, becoming more gentle above 300 °C, and spreading an additional 20 % over the next 1 000 K. The different behavior of rocks with high feldspar content is due to the increase in thermal conductivity with temperature of some plagioclase feldspars (e.g., Höfer and Schilling, 2002; Petrunin et al., 2004) which compensates the decrease in thermal conductivity with temperature observed for most other minerals and rocks. Other notable exceptions are fused silica as well as volcanic and silica glasses.

For *metamorphic rocks* (Figure 6d), the decrease of thermal conductivity with temperature depends on the content in a dominant mineral phase, similar as for plutonic rocks. For quartzites, the decrease is strong, by nearly a factor of three up to a temperature of about 500 °C with only a very mild further decrease beyond this

Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Table 2 Values of c_1 and c_2 in Equation 20 for single-mineral aggregates; data: Clark (1969)

Mineral	T (°C)	$c_1 \times 10^3$ ($W^{-1}\,m\,K$)	$c_2 \times 10^3$ ($W^{-1}\,m$)
Halite, NaCl	0–400	−52.55	0.788
Periclase, MgO	100–800	−21.50	0.127
Corundum, Al_2O_3	100–800	−28.66	0.155
Quartz, SiO_2[a]	100–400	62.10	0.387
Spinel, $MgAl_2O_4$	100–1 000	19.11	0.122
Zircon, $ZrSiO_4$	100–800	131.37	0.093
Forsterite, Mg_2SiO_4,	100–600	85.98	0.282
Enstatite, ferrosilite, $(Mg_2,Fe_2)SiO_3$	100–300	200.63	0.222

[a]Single SiO_2 crystal, heat flowing ⊥ to optical axis



Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Figure 6 Variation of thermal conductivity with temperature for (**a**) sedimentary, (**b**) volcanic, (**c**) plutonic, and (**d**) metamorphic rocks. Color shading indicates a range defined by plus and minus one standard deviation and N is the number of data at each temperature (Clauser, 2009).

temperature. For rocks poor in quartz the decrease in conductivity is not quite as strong, amounting to about one third of the room-temperature value up to 300 °C. Then it remains roughly constant up to 500 °C and decreases again to about one third of the room-temperature value up to 750 °C.

In summary, for moderate temperatures thermal conductivity of rocks is well described by a linear relationship with inverse temperature, similar as in Equation 20. For this temperature range several approaches are available for inferring thermal conductivity at elevated temperatures. Based on the analysis of available tabulated data of thermal conductivity as function of temperature Zoth and Hänel (1988) suggested the following form:

$$\lambda(T) = A + \frac{B}{350 + T} (0\,^{\circ}C \le T \le 800\,^{\circ}C), \qquad (21)$$

where average values of the coefficients A and B for different rock types are given in Table 3.

Linear relationships between temperature and thermal resistivity, such as Equations 20 and 21, discriminate between temperature-dependent contributions and other factors, which are independent of temperature, such as micro-cracks, grain boundaries, pore volume, mineralogical composition, shape, and orientation of crystals and their fragments.

Sass et al. (1992) and Vosteen and Schellschmidt (2003) distinguish between the effects of composition

and temperature. They propose a general empirical relation for $\lambda(T)$, the thermal conductivity in W m^{-1} K^{-1} at temperature T in °C, as a function of λ_0, the thermal conductivity at 0 °C:

$$\lambda(T) = \frac{\lambda_0}{a + T\left(b - \frac{c}{\lambda_0}\right)} \text{ or } \frac{\lambda_0}{\lambda(T)} = \underbrace{a}_{\text{intercept}} + \underbrace{(b - \frac{c}{\lambda_0})}_{\text{slope}}\ T \quad (22)$$

For different rock types, the slopes and intercepts of this equation can be determined from linear regressions of Equation 22 yielding a mean intercept ā and its uncertainty Δa. Coefficients b and c and associated uncertainties σ_b and σ_c are determined from a second linear regression of the different slopes ($b - c/\lambda_0$) as a function of $1/\lambda_0$ (Table 4).

Since thermal conductivity is usually measured at room temperature, λ_0 is expressed as a function of λ_{25}, the room temperature thermal conductivity, by Sass et al. (1992) for crystalline rocks (felsic gneiss to amphibolite) as:

$$\lambda_0 = \lambda_{25}\left(1.007 + 25\left(0.0037 - \frac{0.0074}{\lambda_{25}}\right)\right). \quad (23)$$

Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Table 3 Values for constants A and B in Equation 21 for different rock types (Zoth and Hänel, 1988)

Rock type	T (°C)	A (W m^{-1} K^{-1})	B (W m^{-1})
1. Rock salt	−20 to 0	−2.11	2 960
2. Limestones	0–500	0.13	1 073
3. Metamorphic rocks	0–1 200	0.75	705
4. Acidic rocks	0–1 400	0.64	807
5. Basic rocks	50–1 100	1.18	474
6. Ultra-basic rocks	20–1 400	0.73	1 293
7. Rock types (2)–(5)	0–800	0.70	770

Vosteen and Schellschmidt (2003) find for magmatic and metamorphic rocks:

$$\lambda_0 = 0.53\ \lambda_{25} + 0.5\ \sqrt{1.13\ \lambda_{25}^2 - 0.42\ \lambda_{25}}, \quad (24)$$

and for sedimentary rocks:

$$\lambda_0 = 0.54\ \lambda_{25} + 0.5\ \sqrt{1.16\ \lambda_{25}^2 - 0.39\ \lambda_{25}}. \quad (25)$$

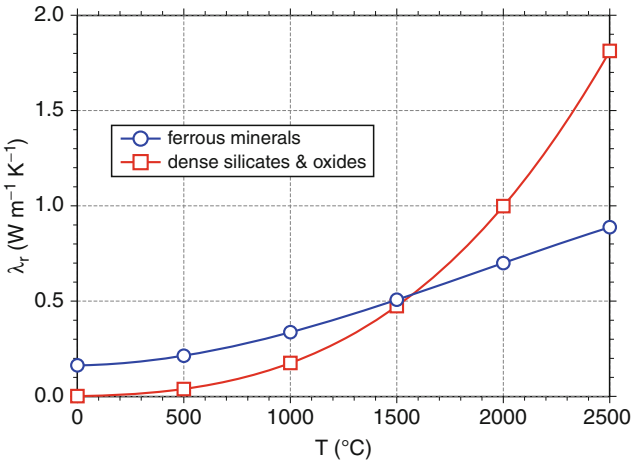
Hofmeister (1999) provided a detailed analysis of heat transport based on an analysis of phonon lifetimes obtained from infrared reflectivity. It accounts for the variation of the phonon contribution λ_p to thermal conductivity with both temperature and pressure as well as for the pressure dependent radiative contribution λ_r. It allows approximation of thermal conductivity for mantle conditions if: (1) K'_0, the pressure derivative of the isothermal bulk modulus K_T, is constant: $K'_0 = dK_T/dP$ = const; (2) the variations of the bulk modulus with temperature and pressure are mutually independent; (3) the pressure derivative of the thermodynamic Grüneisen parameter γ (Equation 12 or 13 in *Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat*, this volume) is constant: $d\gamma/dP = f$. For mantle material, γ varies from 1 to1.4, K'_0 from 4 to 5, and $f \approx 0$ vanishes approximately. According to Hofmeister (1999) and within the uncertainty of these parameters, thermal conductivity in the mantle is:

$$\lambda(T,P) = \lambda_{298\,K,\,101.33\,kPa} \times (298/T)^a (1 + K'_0 P/K_T) e^{-(4\gamma+1/3)\int_{298}^{T}\alpha(T')dT'} + \lambda_r, \quad (26)$$

where T is absolute temperature, $\lambda_{298\ K,\ 101.33\ kPa}$ is thermal conductivity at room temperature and atmospheric pressure, α(T) is volume coefficient of thermal expansion as a function of temperature, and the exponent a is the fitting parameter. The radiative contribution λ_r in (26) may be approximated by Equations 15, 18, or 19. Hofmeister (1999) provides alternative expressions for λ_r for ferrous minerals or dense silicates and oxides (Figure 7):

Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Table 4 Coefficients a, b, and c in Equation 22 and associated uncertainties Δa and σ_b, σ_c; Δa is the error of the mean intercept ā for all rock types of the linear regressions of the normalized thermal resistance $\lambda_0/\lambda(T)$ as a function of temperature T; σ_b and σ_c are the errors defined by the linear regression of the slopes ($b - c/\lambda_0$) as a function of the thermal resistance $1/\lambda_0$ (see Equation 22)

Rock type	ā (–)	Δa (%)	$b \times 10^3$ (K^{-1})	$\sigma_b \times 10^3$ (K^{-1})	$c \times 10^3$ (W m^{-1} K^{-2})	$\sigma_c \times 10^3$ (W m^{-1} K^{-2})	T (°C)	Reference
Basement rocks I (from felsic gneiss to amphibolite)	1.007	–	3.6	–	7.2	–	0–250	Sass et al., 1992
Basement rocks II (magmatic and metamorphic)	0.99	1	3.0	1.5	4.2	0.6	0–500	Vosteen and Schellschmidt, 2003
Sediments	0.99	1	3.4	0.6	3.9	1.4	0–300	Vosteen and Schellschmidt, 2003



Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Figure 7 Variation of radiative thermal conductivity λ_r of ferrous minerals, dense silicates, and oxides with temperature according to Equation 27 (Hofmeister, 1999).

ferrous minerals :

$$\lambda_r = 0.01753 - 1.0365 \times 10^{-4} T + 2.2451 \times 10^{-7} T^2 - 3.407 \times 10^{-11} T^3;$$

dense silicates and oxides

$$\lambda_r = 8.5 \times 10^{-11} T^3 \ (\lambda_r \text{ in W m}^{-1}\text{ K}^{-1} \text{ and T in K}). \quad (27)$$

Figure 1 in the companion article in this volume (see *Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat*) shows the variation of phonon thermal conductivity λ_p with temperature in an average crust of density $\rho = 2\,700\,\text{kg m}^{-3}$ molar mass of $0.22178\,\text{kg mol}^{-1}$. It is derived from $\lambda = \kappa\,\rho\,c$ as the product of thermal capacity $\rho\,c_P$ and thermal diffusivity κ calculated according to Equation 20 (in Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat, this volume"; Whittington et al., 2009). Due to the balancing increase of specific heat capacity with temperature, the decrease in thermal conductivity with temperature is less than in thermal diffusivity by a factor of about 1.3. Assuming a constant density throughout the crust, this implies that the increase and decrease in density due to the increase in pressure and temperature, respectively, partly cancel each other and that these changes are small compared to those of specific heat capacity and thermal diffusivity.

For mantle minerals, such as olivine and its high-pressure polymorphs, the β- and γ-spinels wadsleyite and ringwoodite, respectively, Xu et al. (2004) fitted phonon thermal conductivity measured to 1 373 K and 20 GPa to an exponential equation in temperature yielding values for the exponent between −0.406(35) and −0.537(11) suggesting that fitted phonon thermal conductivity varies with $T^{-1/2}$ (Table 5).

Variation with pressure

The effect of pressure on phonon thermal conductivity λ_p is different in two distinct pressure ranges. First, fractures and micro-cracks (developed during stress release after sampling) begin to close with increasing pressure. This reduces thermal contact resistance as well as porosity which is usually filled with a low conductivity fluid. This process ends when a pressure of about 15 MPa is reached. A compilation of measurements on various sedimentary, volcanic, plutonic, and metamorphic rocks (Clauser and Huenges, 1995) indicates that this effect accounts for an increase of about 20 % relative to thermal conductivity at atmospheric pressure. A further pressure increase to 40 MPa does not affect thermal conductivity significantly. If pressure is increased further, however, a second process becomes effective, the reduction of intrinsic porosity, i.e., voids which are not created by stress release. For granite and for metamorphic rocks, data indicate an increase of thermal conductivity by about 10 % within the pressure range 50 MPa–500 MPa.

For mantle minerals, such as olivine and its high-pressure polymorphs, the β- and γ-spinels wadsleyite and ringwoodite, respectively, Xu et al. (2004) determined values for the pressure coefficient between $0.022\,\text{GPa}^{-1}$ and $0.032\,\text{GPa}^{-1}$ for phonon thermal conductivity measured to 1 373 K and 20 GPa (Table 5). A table of numerical values for pressure derivatives of phonon thermal conductivity measured by a variety of authors was compiled by Hofmeister et al. (2009). Most values for $\lambda^{-1}(\partial\lambda/\partial P)$ fall into the range $0.04\,\text{GPa}^{-1}$–$0.36\,\text{GPa}^{-1}$, exceeded only by values of $0.69\,\text{GPa}^{-1}$ and $0.5\,\text{GPa}^{-1}$ for sulfur and quartz$_{\perp c}$ (measured perpendicular to the optical *c*-axis), respectively. Osako et al. (2004) fitted thermal conductivity measured to 1 100 K and 8.3 GPa on isotropic, single-crystal garnet and anisotropic olivine ($Fo_{93}Fa_7$; in three crystallographic directions) as upper and lower mantle constituents, respectively, to a linear equation in pressure (Table 6). A pressure dependence of garnet and olivine was found on the order of $4\,\%\,\text{GPa}^{-1}$–$5\,\%\,\text{GPa}^{-1}$ and $3\,\%\,\text{GPa}^{-1}$–$4\,\%\,\text{GPa}^{-1}$, respectively.

In contrast, radiative thermal conductivity λ_r was found much less variable with pressure than with temperature (Clark, 1957). In particular, this holds once the spectral radiance overlaps the infrared pass band in the absorption spectrum, at temperatures above 1 900 K (Hofmeister, 2005).

Variation with other factors

Apart from temperature and pressure, thermal conductivity also varies with porosity, pore fluid, saturation, dominant

mineral phase, and anisotropy. These effects are summarized here from the detailed discussion in Clauser (2006):

For large *porosity* (i.e., $\phi >> 1\,\%$) thermal conductivity of the saturating fluid affects significantly the bulk rock thermal conductivity. The influence varies with the thermal conductivity of the saturants, e.g., water, oil, natural gas, air. The resulting bulk thermal conductivity can be estimated from a suitable mixing model, e.g., Equations 2–4.

The effect of partial *saturation* is different for porous or fractured rocks. In porous rocks, porosity comprises both bulk pore space and bottlenecks formed by the contact between individual grains. Dry bottlenecks act as thermal contact resistances between grains, while the bulk pore volume contributes proportionally to the effective rock thermal conductivity. In fractured rocks, in contrast, there are no bottlenecks between grains as in porous rocks, and the small void volume in the fractures corresponds to the bulk pores space of porous rocks.

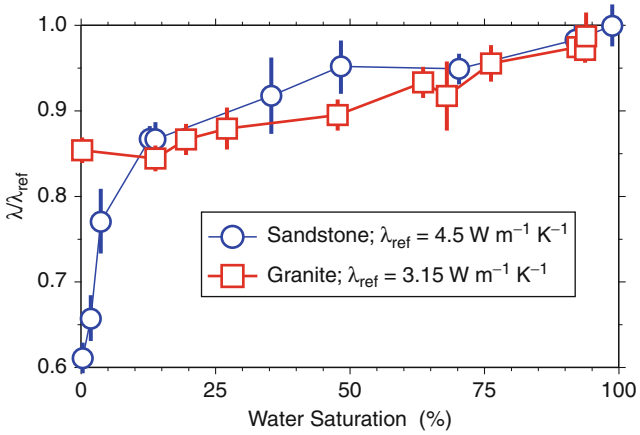
Figure 8 illustrates these two effects for a water-saturated medium-porosity sandstone and a low-porosity granite: Starting completely dry with an unsaturated conductivity of about 60 % of the saturated value, a level of 85 % is reached for the sandstone at about 10 % saturation. The 15 % conductivity residual is then spread almost linearly over the remaining 90 % of saturation. Physically this observation indicates that the filling of inter-granular bottlenecks accounting for only about 10 %−15 % of the total porosity significantly reduces the contact resistances between the individual grains. In contrast, the replacement of low conductivity air by the more conductive fluid in the major part of the pore volume accounts for the second effect. If only fractures contribute to the total porosity, as in the granite, there are no bottlenecks and only the second effect is observed: Starting completely dry with an unsaturated conductivity of about 85 % of the saturated conductivity, a quasi linear increase is observed

Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Table 5 Reference values λ_{298} and κ_{298} at 298 K, pressure coefficients a and a′, and fitting functions for thermal conductivity λ and thermal diffusivity κ of lower mantle minerals with absolute temperature T and pressure P (Xu et al., 2004)

		$\lambda = \lambda_{298}\,(298/T)^{1/2}\,(1 + a\,P)$		$\kappa = \kappa_{298}\,(298/T)^{n}\,(1 + a'P)$		
Mineral	P (GPa)	λ_{298} (W m^{-1} K^{-1})	a (GPa^{-1})	$\kappa_{298} \times 10^6$ (m^2 s^{-1})	n	a′ (GPa^{-1})
Olivine	4–10	4.13(11)	0.032(3)	1.31(5)	0.681(22)	0.036(4)
	4	4.49(4)	–	1.29(5)	0.563(35)	–
	7	5.19(4)	–	1.74(4)	0.720(26)	–
	10	5.56(4)	–	1.84(5)	0.723(31)	–
Wadsleyite	14	8.10(4)	0.023	2.55(3)	0.721(13)	–
Ringwoodite	20	9.54(5)	0.022	3.09(4)	0.793(17)	–

Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Table 6 Coefficients and fitting functions for the variation thermal conductivity λ and thermal diffusivity κ of garnet and olivine with absolute temperature T and pressure P (Osako et al., 2004)

Garnet				Olivine				
$\lambda = C_0 + C_1/T$		$\lambda = A_0 + A_1\,P$			$\lambda = C_0 + C_1/T$		$\lambda = B_0\,\exp(B_1\,P)$	
C_0 (W m^{-1} K^{-1})	C_1 (W m^{-1})	A_0 (W m^{-1} K^{-1})	A_1 (W m^{-1} K^{-1} GPa^{-1})		C_0 (W m^{-1} K^{-1})	C_1 (W m^{-1})	B_0 (W m^{-1} K^{-1})	B_1 (GPa^{-1})
2.01(8)	704(43)	3.48(33)	0.160(26)	[100]	1.91(28)	2,088(163)	6.61(13)	0.038(5)
				[010]	0.84(36)	1,377(157)	3.98(15)	0.042(5)
				[001]	2.08(38)	1,731(86)	5.91(25)	0.034(5)
$\kappa = c_0 + c_1/T$		$\kappa = a_0 + a_1\,P$			$\kappa = c_0 + c_1/T$		$\kappa = b_0\,\exp(b_1\,P)$	
$c_0 \times 10^6$ (m^2 s^{-1})	$c_1 \times 10^6$ (m^2 s^{-1} K)	$a_0 \times 10^6$ (m^2 s^{-1})	$a_1 \times 10^6$ (m^2 s^{-1} GPa^{-1})		$c_0 \times 10^6$ (m^2 s^{-1})	$c_1 \times 10^6$ (m^2 s^{-1} K)	$b_0 \times 10^6$ (m^2 s^{-1})	b_1 (GPa^{-1})
0.29(6)	374(31)	1.19(6)	0.046(1)	[100]	−0.06(11)	938(46)	2.50(4)	0.033(5)
				[010]	−0.13(8)	626(45)	1.52(6)	0.040(7)
				[001]	−0.03(17)	832(98)	2.16(14)	0.035(3)



Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Figure 8 Variation of thermal conductivity with partial saturation for a sandstone (circles; $\phi = 18\,\%$) and granite (squares; $\phi = 1\,\%$) saturated with water and standard deviations (bars); values normalized by reference thermal conductivities shown in legend (Clauser, 2006 based on data of Reibelt, 1991).

due to the replacement of low conductivity air until the 100 % level is reached for complete saturation.

Anisotropy of sedimentary and metamorphic rocks is due to the conditions of their formation. Anisotropy exists on several scales: (1) On the microscopic scale, many minerals are anisotropic; (2) on the laboratory scale, thermal conductivity of many rocks is also anisotropic. However, even if rocks are composed of anisotropic minerals, random orientation of the crystals within the rock may render the rock's bulk thermal conductivity isotropic on a macroscopic scale; (3) on a still larger scale, if rocks are exposed to folding, orogenic or other tectonic processes, thermal conductivity of the resulting rock formation may be anisotropic.

As a result, thermal conductivity parallel to the direction of layering or foliation, $\lambda_{||}$, is greater than thermal conductivity in the perpendicular direction, $\lambda_{\perp}$. The factor of anisotropy, the ratio $\lambda_{||}/\lambda_{\perp}$, generally falls into the range 0.9–3, with most values between 1 and 2 (e.g., Clauser and Huenges, 1995; Popov and Mandel, 1998; Popov et al., 1999a, b; Clauser, 2006; Davis et al., 2007). For sedimentary rocks a general trend has been reported of decreasing $\lambda_{\perp}$ with factor of anisotropy $\lambda_{||}/\lambda_{\perp}$, but no such trend was identified for metamorphic rocks (Clauser, 2006).

Thermal diffusivity

Thermal diffusivity is required in the analysis of transient heat transfer. A sizable compilation of room-temperature data of phonon thermal diffusivity κ_p measured on various mantle minerals is given by Hofmeister et al. (2009) from measurements by various groups.

If both conductivity and thermal capacity are known, thermal diffusivity κ can be calculated from $\kappa = \lambda/(\rho\, c)$. As for steady-state thermal conduction, transient heat diffusion in most of the Earth's crust and mantle is caused by scattering of quantized lattice vibrations, the phonons, and by diffusive (as opposed to ballistic) radiation of photons. These two processes are described by phonon thermal conductivity λ_p and radiative thermal conductivity, λ_r, respectively, the sum of which is often termed effective thermal conductivity, λ_{eff}. As thermal diffusivity is the ratio of thermal conductivity and thermal capacity, it is also influenced by the variation of density and specific heat capacity. This is of particular interest with respect to the variation with temperature.

Measuring techniques

All of the transient laboratory methods used to determine thermal conductivity are useful to determine thermal diffusivity as well. Recently, heat pulse (Schilling, 1999; Höfer and Schilling, 2002; Gibert et al., 2003) and laser flash methods (Parker et al., 1961; Blumm and Lemarchand, 2002; Hofmeister, 2006) have been used for measurements at high temperature. Compared to other methods their advantage lies in a reduction or even complete absence of physical contacts between samples on the one hand and temperature sensors and heat sources on the other hand. While most methods measure the effective diffusivity comprising contributions from phonon conduction and diffused heat radiation, the laser flash method yields the diffusive component without contributions from heat radiation. This difference becomes important particularly at high temperatures (for a critical assessment see e.g., Hofmeister et al., 2009).

Variation with temperature

Thermal diffusivity κ of rocks varies even more strongly with temperature than thermal conductivity λ. This is caused by the opposite behavior of thermal conductivity and thermal capacity (ρ c) with respect to temperature. Because of several self-compensating factors, thermal capacity (ρ c) with few exceptions generally varies within $\pm 20\,\%$ of $2.3\,\mathrm{MJ\,m^{-3}\,K^{-1}}$ for the great majority of minerals and rocks (Beck, 1988). This is confirmed by a linear regression of thermal diffusivity on thermal conductivity which was measured on a suite of metasedimentary, volcanic, magmatic, and metamorphic rocks together with density and specific heat capacity (Mottaghy et al., 2005):

$$\kappa = \frac{\lambda}{\rho\, c} = \frac{\lambda}{2.3} = 0.44\,\lambda\,(\kappa \text{ in } 10^{-6}\,\mathrm{m^2\,s^{-1}}). \qquad (28)$$

A linear regression of thermal capacity as a function of temperature yields also a linear relationship. This allows to determine thermal diffusivity $\kappa(T)$ at any temperature,

based only on the known variation of thermal conductivity λ(T) with temperature (Mottaghy et al., 2005):

$$\kappa(T) = \frac{\lambda(T)}{2.134 + 0.0044\,T} \quad (\kappa \text{ in } 10^{-6}\,\mathrm{m^2\,s^{-1}};\ \mathrm{T\ in\ ^\circ C}). \tag{29}$$

Thus, thermal diffusivity can be derived from thermal conductivity and vice versa. For the suite of rocks studied by Mottaghy et al. (2005), thermal conductivity decreased by 4 %–7 % in the range 1–100 °C while thermal diffusivity decreased by 18 %–22 %.

Thermal diffusivity was measured at temperatures of up to 550 °C by Ray et al. (2006) on a suite of 16 samples comprising Archean granulitic rocks which are considered a major component of the middle and lower continental crust. Phonon scattering was found dominating heat transport with radiative diffusion of photons setting in at 450 °C for most, but not all rocks. Based on their measurements, Ray et al. (2006) propose an equation by which thermal diffusivity κ(T) at elevated temperature below 450 °C can be derived from room temperature values, κ_{rt}:

$$\kappa(T) = 0.7 + 144\,\frac{\kappa_{rt} - 0.7}{T - 150} \quad (\kappa \text{ in } 10^{-6}\,\mathrm{m^2\,s^{-1}};\ \mathrm{T\ in\ K}), \tag{30}$$

They also proposed an additional term in Equation 30, proportional to T^3, by which the radiative contribution above 450 °C is fitted. However, this assumes a similar variation with temperature of thermal conductivity and thermal diffusivity and neglects the additional variation of specific heat capacity. This, in fact, makes thermal diffusivity vary stronger with temperature than thermal conductivity, which is why the T^3-term is omitted here. Whittington et al. (2009) measured the phonon component κ_p of thermal diffusivity at temperatures of up to 1 260 K on garnet schist, leucogranite, and welded rhyolitic ash-flow tuff using laser flash analysis. This characterizes purely the phonon heat transfer component without any radiative contribution, in contrast to the values discussed above. Below and above the transition between α- and β-quartz at 846 K (~573 °C) the data are reasonably fitted by

$$\kappa_p(T) = \begin{cases} -0.062 + \frac{567.3}{T}; & \mathrm{T} > 846\ \mathrm{K} \\ 0.732 - 0.000135\,T; & \mathrm{T} < 846\ \mathrm{K} \end{cases} \quad (\kappa \text{ in } 10^{-6}\,\mathrm{m^2\,s^{-1}};\ \mathrm{T\ in\ K}), \tag{31}$$

assuming an average molar mass of 0.22178 kg mol^{-1} and an average density of 2 700 kg m^{-3} for the crust (see Figure 1 in *Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat*, this volume). The leucogranite and rhyolite samples were homogeneous and isotropic. The schist was anisotropic owing to alternating mica- and quartz-rich layers, requiring testing in the direction parallel and perpendicular to foliation.

For mantle minerals, such as olivine and its high-pressure polymorphs, the β- and γ-spinels wadsleyite and ringwoodite, respectively, Xu et al. (2004) fitted phonon thermal diffusivity measured to 1 373 K and 20 GPa to an exponential equation in temperature yielding values for the exponent between −0.563(35) and −0.793(17) suggesting that fitted phonon thermal conductivity varies with $T^{-½}$–T^{-1} (Table 5).

Pertermann and Hofmeister (2006) measured thermal diffusivity on oriented single crystals and polycrystalline samples of olivine-group minerals with the laser-flash method at temperatures of up to about 1 500 °C. They fitted the data to a second order polynomial in T:

$$\kappa_p(T) = a + b/T + c/T^2 (\mathrm{T\ in\ K}). \tag{32}$$

Values for the coefficients a, b, and c fitted to data measured on single crystal and polycrystalline samples are shown in Table 7.

Osako et al. (2004) fitted thermal diffusivity measured to 1 100 K and 8.3 GPa on isotropic, single-crystal garnet and anisotropic olivine ($Fo_{93}Fa_7$; in three crystallographic directions) as upper and lower mantle constituents, respectively, to a linear equation in inverse temperature (Table 6). They found a strong anisotropy in olivine which they assume to prevail throughout the olivine stability field in the mantle down to 410 km.

Variation with pressure

Tommasi et al. (2001) measured thermal diffusivity in the crystallographic [100] and [010] directions parallel and perpendicular to a strain-induced foliation, respectively, as a function of temperature at atmospheric pressure on spinel lherzolites and spinel harzburgite. These rocks are considered representative for the subcontinental and suboceanic mantle, respectively. They found an anisotropy in thermal diffusivity on the order of 25 % with the maximum aligned in the direction of strain. Support of their experimental findings was provided by corroborating petrophysical modeling.

For mantle minerals, such as olivine and its high-pressure polymorphs, the β- and γ-spinels wadsleyite and ringwoodite, respectively, Xu et al. (2004) fitted phonon thermal diffusivity measured to 1 373 K and 20 GPa to a linear equation in pressure yielding a pressure coefficient of 0.036(4) GPa^{-1} (Table 5).

Osako et al. (2004) fitted thermal diffusivity measured to 1 100 K and 8.3 GPa on isotropic, single-crystal garnet and anisotropic olivine ($Fo_{93}Fa_7$; in three crystallographic directions) as upper and lower mantle constituents, respectively, to an exponential equation in pressure (Table 6). The pressure dependence of garnet and olivine was found to be on the order of 4 % GPa^{-1}–5 % GPa^{-1} and 3 % GPa^{-1}–4 % GPa^{-1}, respectively.

Thermal Storage and Transport Properties of Rocks, II: Thermal Conductivity and Diffusivity, Table 7 Coefficients for fitting the variation of thermal diffusivity (10^{-6} m^2 s^{-1}) with temperature (K) according to Equation 32 (Pertermann and Hofmeister, 2006)

Sample		Chemical composition	$a \times 10^6$ (m^2 s^{-1})	$b \times 10^6$ (m^2 s^{-1} K)	$c \times 10^6$ (m^2 s^{-1} K^2)	T_{max} (°C)
		Single crystals				
Olivines	Fo[001]	Mg_2SiO_4	0.3081	679.6	213 492	985
	FoCo[001]	$Mg_{1.99}Co_{0.01}SiO_4$	0.2347	587.8	172 482	1 477
	[010]	”	0.2415	115.4	165 515	1 181
	Needles[100]	$Mg_{1.84}Fe_{0.16}SiO_4$	0.7088	57.7	202 533	985
	[010]	”	0.3100	100.6	86 470	739
	[001]	”	0.3805	381.3	79 703	886
	Sumput[010]	$Mg_{1.87}Fe_{0.13}SiO_4$	0.3135	127.2	73 824	983
Sinhalite	[010]	$MgAlBO_4$	0.5546	−128.1	432 571	741
Chrysoberyl	[100]	$BeAl_2O_4$	0.5366	551.7	566 872	990
	[010]	”	0.3516	415.2	388 978	989
	[001]	”	0.6371	428.2	543 382	990
		Polycrystalline samples				
Dunites	#1	~$Mg_{1.8}Fe_{0.2}SiO_4$	0.2291	290.0	92 938	888
	#2	”	0.3563	178.4	93 356	1 083
Monticellite-bearing rock		$Ca_{1.15}$ $Mg_{0.79}Mn_{0.06}SiO_4$	0.3816	153.0	23 706	985
Hortonolite-bearing rock		$Mg_{1.2}Fe_{0.8}SiO_4$	0.2826	301.2	12 638	705
Fayalite-bearing slag		~$Fe_{1.98}Mn_{0.02}SiO_4$	0.2637	83.9	21 216	886
Fayalite-bearing rock		~$Fe_{1.84}Mn_{0.02}$ $Mg_{0.14}SiO_4$	0.1798	265.3	15 688	887

Fo forsterite, *FoCo* Co-doped forsterite

Variation with other factors

Micro-cracks and grain boundaries give rise to increased thermal resistance and to a reduction of the mean free path of radiation due to scattering of radiation. It is somewhat debated below which grain size the effect is negligible: Based on the agreement between diffusivities measured on minerals and rocks, Gibert et al. (2003) concluded that the effect of grain boundaries, thermal cracking, and secondary phases is negligible. Along the same lines, Seipold (1998) argued that grain size is much larger than the phonon mean free path and therefore grain boundaries should not interfere with heat diffusion. Branlund and Hofmeister (2008) confirm this for quartzites, but find diffusivities measured on chert, agate, and chalcedony to be lowered by grain boundaries. They propose that grain sizes above 1 μm should not affect heat transfer.

Additional advective heat transfer was identified by Seipold and Schilling (2003) due to the release of water adsorbed at the inner surfaces of voids (i.e., pores and cracks) in rocks at about 450 K and by dehydration of serpentinite at 850 K. Both processes create high local overpressures which are relieved by cracking if the overpressure exceeds the tensile strength of the rock. The resulting flow is then accompanied by a corresponding advective heat transfer. This phenomenon was observed and studied in laboratory experiments (Seipold and Schilling, 2003) but has implications for the lower crust and upper mantle with respect to recrystallization processes involving the discharge of fluids. These “crustal burps” provide the only conceivable way how some fluids from the mantle or lower crust my find their way to the Earth’s surface. The example discussed by Seipold and Schilling (2003) involves liberating water of crystallization during the conversion of serpentinite into forsterite and talc, followed by the formation of enstatite. As this involves heat advection as a separate heat transfer mechanism, this process is better addressed separately and not parameterized into some sort of “effective” heat transport property not directly linked to a physical process.

Summary

Understanding the thermal regime of the Earth requires appreciation of properties and mechanisms for storage, transport, and generation of heat with the Earth. Both experimental and indirect methods are available for inferring the corresponding rock properties. Steady-state heat conduction or transient heat diffusion is the dominant transport process in the Earth’s crust, except when appreciable fluid flow provides a mechanism for heat advection. For most crustal and mantle rocks, heat radiation sets in at temperatures above about 450 °C and becomes significant only at temperatures above 1 200 °C. At temperatures above 2 500 °C heat radiation becomes a dominant mechanism.

Acknowledgments

Extensive and constructive comments provided by an anonymous reviewer are greatly appreciated. Dr. Sukanta Roy kindly helped to meet the size limit of this contribution.

Bibliography

Aronsson, J. R., Bellottti, L. H., Eckroad, S. W., Emslie, A. G., McConnell, R. K., and von Thüna, P. C., 1970. Infrared spectra and radiative thermal conductivity of minerals at high temperatures. *Journal of Geophysical Research*, **75**(17), 3443–3456.

Beardsmore, G. R., and Cull, J. P., 2001. *Crustal Heat Flow*. Cambridge: Cambridge University Press.

Beck, A. E., 1988. Methods for determining thermal conductivity and thermal diffusivity. In Hanel, R., Rybach, L., and Stegena, L. (eds.), *Handbook of Terrestrial Heat Flow Density Determination*. Dordrecht: Kluwer, pp. 87–124.

Birch, F., 1942. Thermal conductivity and diffusivity. In Birch, F., Schairer, J. F., Spicer, H. C. (eds.), *Handbook of Physical Constants*, Special Paper 36. New York: Geological Society of America, pp. 243–266.

Blumm, J., and Lemarchand, S., 2002. Influence of test conditions on the accuracy of laser flash measurements. *High Temp.-High Pres.*, **34**, 523–528.

Branlund, J. M., and Hofmeister, A. M., 2008. Factors affecting heat transfer in natural SiO_2 solids. *Ameican Mineralogist*, **93**, 1620–1629.

Bruggeman, D. A. G., 1935. Berechnung verschiedener Konstanten von heterogenen Substanzen – I. Dielektrizitätskonstanten und Leitfähigkeiten der Mischkörper aus isotropen Substanzen. *Annalen der Physik*, **24**, 636–679.

Buntebarth, G., 1991. Thermal properties of the KTB-Oberpfalz VB core samples at elevated temperature and pressure. *Scientific Drilling*, **2**, 73–80.

Buntebarth, G., and Schopper, J. R., 1998. Experimental and theoretical investigations on the influence of fluids, solids and interactions between them on thermal properties of porous rocks. *Physics and Chemistry of the Earth*, **23**, 1141–1146.

Carmichael, R. S. (ed.), 1984. *CRC Handbook of Physical Properties of Rocks III*. Boca Raton: CRC Press.

Čermák, V., and Rybach, L., 1982. Thermal conductivity and specific heat of minerals and rocks. In Angenheister, G. (ed.), *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology,* New Series, V(1a). Berlin: Springer, pp. 305–343.

Clark, S. P., Jr., 1957. Radiative transfer in the Earth's mantle. *EOS. Transactions of the American Geophysical Union*, **38**, 931–938.

Clark, S. P., Jr., 1966. Thermal conductivity. In Clark, S. P., Jr. (ed.), *Handbook of Physical Constants,* Memoir 97. New York: Geological Society of America, pp. 459–482.

Clark, S. P., Jr., 1969. Heat conductivity in the mantle. In Hart, P. J. (ed.), *The Earth's Crust and Upper Mantle*, Geophysical Monograph 13. Washington, DC: American Geophysical Union, pp. 622–626.

Clauser, C., 1988. Opacity – the concept of radiative thermal conductivity. In Hänel, R., Rybach, L., and Stegena, L. (eds.), *Handbook of terrestrial heat flow density determination*. Dordrecht: Kluwer, pp. 143–165.

Clauser, C., 2006. Geothermal energy. In Heinloth, K. (ed.), *Landolt-Börnstein, Group VIII: Advanced Materials and Technologies*. Heidelberg-Berlin: Springer. Energy Technologies, Subvol. C: Renewable Energies, Vol. 3, pp. 480–595.

Clauser, C., 2009. Heat transport processes in the Earth's crust. *Surveys in Geophysics*, **30**, 163–191, doi:10.1007/s10712-009-9058-2.

Clauser, C., and Huenges, E., 1995. Thermal conductivity of rocks and minerals. In Ahrens, T. J. (ed.), *Rock Physics and Phase Relations – a Handbook of Physical Constants*. Washington, DC: American Geophysical Union. AGU Reference Shelf, Vol. 3, pp. 105–126.

Crain, E. R., 1986. *The Log Analysis Handbook*. Tulsa: Pennwell Publishing. Quantitative Log Analysis Methods Series, Vol. 1.

Davis, E. E., 1988. Oceanic heat-flow density. In Hänel, R., Rybach, L., and Stegena, L. (eds.), *Handbook of Terrestrial Heat Flow Density Determination*. Dordrecht: Kluwer, pp. 223–260.

Davis, G. M., Chapman, D. S., Van Wagoner, T. M., and Armstrong, P. A., 2007. Thermal conductivity anisotropy of metasedimentary and igneous rocks. *Journal of Geophysical Research*, **112**, B05216, doi:10.1029/2006JB004755.

Debye, P., 1914. Zustandsgleichung und Quantenhypothese mit einem Anhang über Wärmeleitung. In *Vorträge über die kinetische Theorie der Materie und der Elektrizität, Mathematische Vorlesungen an der Universität Göttingen,* VI. Leipzig: Teubner.

Diment, W. H., and Pratt, H. R., 1988. *Thermal Conductivity of Some Rock-Forming Minerals: A Tabulation*, Open File Report 88-690. Denver, CO: US Geological Survey.

Dreyer, W., 1974. *Materialverhalten Anisotroper Festkörper: Thermische und Elektrische Eigenschaften*. Wien: Springer.

Eucken, A., 1911. Über die temperaturabhängigkeit der wärmeleitfähigkeit fester nichtmetalle. *Annalen der Physik, 4. Folge*, **34**(2), 185–220.

Fricke, H., 1924. A mathematical treatment of the electric conductivity and capacity of disperse systems. *Physical Review*, **24**, 575–587.

Gibert, B., Seipold, U., Tommasi, A., and Mainprice, D., 2003. Thermal diffusivity of upper mantle rocks: influence of temperature, pressure, and the deformation fabric. *Journal of Geophysical Research*, **108**(B8), 2359, doi:10.1029/2002JB002108.

Goutorbe, B., Lucazeau, F., and Bonneville, A., 2006. Using neural networks to predict thermal conductivity from geophysical well logs. *Geophysical Journal International*, **166**(1), 115, doi:10.1111/j.1365-246X.2006.02924.x.

Hartmann, A., Rath, V., and Clauser, C., 2005. Thermal conductivity from core and well log data. *International Journal of Rock Mechanics and Mining Sciences*, **42**, 1042–1055, doi:10.1016/j.ijrmms.2005.05.015.

Hashin, Z., and Shtrikman, S., 1962. A variational approach to the theory of the effective magnetic permeability of multiphase materials. *Journal of Applied Physics*, **33**(10), 3125–3131.

Höfer, M., and Schilling, F. R., 2002. Heat transfer in quartz, orthoclase, and sanidine at elevated temperature. *Physics and Chemistry of Minerals*, **29**, 571–584.

Hofmeister, A. M., 1999. Mantle values of thermal conductivity and the geotherm from phonon lifetimes. *Science*, **283**, 1699–1706.

Hofmeister, A. M., 2005. Dependence of diffusive radiative transfer on grain-size, temperature, and Fe-content: Implications for mantle processes. *Journal of Geodynamics*, **40**, 51–72.

Hofmeister, A. M., 2006. Thermal diffusivity of garnets at high temperature. *Physics and Chemistry of Minerals*, **33**, 45–62.

Hofmeister, A. M., Branlund, J. M., and Pertermann, M., 2009. Properties of rocks and minerals – thermal conductivity of the Earth. In Price, G. D. (ed.), *Mineral Physics, Treatise on Geophysics*. Amsterdam: Elsevier, Vol. 2, pp. 543–577.

Horai, K., 1971. Thermal conductivity of rock-forming minerals. *Journal of Geophysical Research*, **76**(5), 1278–1308.

Horai, K., 1991. Thermal conductivity of Hawaiian Basalt: a new interpretation of Robertson and Peck's data. *Journal of Geophysical Research*, **96**(B3), 4125–4132.

Horai, K., and Simmons, G., 1969. Thermal conductivity of rock-forming minerals. *Earth and Planetary Science Letters*, **6**(5), 359–368.

Kappelmeyer, O., and Haenel, R., 1974. *Geothermics – with special reference to application*. Berlin-Stuttgart: Bornträger.

Kobolev, V. P., Kutas, R. I., and Popov, Y. A., 1990. Method and results of research of thermal properties of Ural region rocks with

laser scanning. *Geophysical Journal*, **12**(4), 29–37, Naukova Dumka, Kiev (in Ukrainian).
Korvin, G., 1978. The hierarchy of velocity formulae: generalized mean value theorems. *Acta Geodaetica, Geophysica et Montanista*, **13**, 211–222.
Korvin, G., 1982. Axiomatic characterization of the general mixture rule. *Geoexploration*, **19**, 267–276.
Mottaghy, D. C., Schellschmidt, R., Popov, Y. A., Clauser, C., Kukkonen, I. T., Nover, G., Milanovsky, S., and Romushkevich, R. A., 2005. New heat flow data from the immediate vicinity of the Kola super-deep borehole: vertical variation in heat flow confirmed and attributed to advection. *Tectonophysics*, **401**(1–2), 119–142, doi:10.1016/j.tecto.2005.03.005.
Osako, M., Ito, E., and Yoneda, A., 2004. Simultaneous measurement of thermal conductivity and thermal diffusivity for garnet and olivine under high pressure. *Physics of the Earth and Planetary Interiors*, **143–144**, 311–320.
Parker, W. J., Jenkins, R. J., Butler, C. P., and Abbott, G. I., 1961. Flash method of determining thermal diffusivity, heat capacity, and thermal conductivity. *Journal of Applied Physics*, **32**(9), 1679–1684.
Peierls, R., 1929. Zur kinetischen Theorie der Wärmeleitung in Kristallen. *Annalen der Physik, 5. Folge*, **3**, 1055–1101.
Pertermann, M., and Hofmeister, A. M., 2006. Thermal diffusivity of olivine-group minerals at high temperature, *American Mineralogist*, **91**, 1747–1760.
Petrunin, G. I., Popov, V. G., and Il'in, I. A., 2004. Conductive heat transfer in Plagioclases. *Izvestiya, Physics of the Solid Earth (English Translatrion)*, **40**(9), 752–759.
Popov, Y. A., and Mandel, A. M., 1998. Geothermal study of anisotropic rock masses. *Izvestiya, Physics of the Solid Earth (English Translation)*, **34**(11), 903–915.
Popov, Y. A., Berezin, V. V., Solov'yev, G. A., Romushkevich, R. A., Korostelev, V. M., Kostyurin, A. A., and Kulikov, I. V., 1987. Thermal conductivity of minerals. *Izvestiya, Physics of the Solid Earth (English Translation)*, **23**(3), 245–253.
Popov, Y. A., Pevzner, L. A., Romushkevich, R. A., Korostelev, V. M., and Vorob'ev, M. G., 1995. Thermophysical and geothermal sections obtained from Kolvinskaya well logging data. *Izvestiya, Physics of the Solid Earth (English Translation)*, **30**(9), 778–789.
Popov, Y. A., Pevzner, L. A., Pimenov, V. P., and Romushkevich, R. A., 1999a. New geothermal data from the Kola superdeep well SG-3. *Tectonophysics*, **306**(3–4), 345–366.
Popov, Y. A., Pribnow, D. F. C., Sass, J. H., Williams, C. F., and Burkhardt, H., 1999b. Characterization of rock thermal conductivity by high-resolution optical scanning. *Geothermics*, **28**, 253–276.
Popov, Y., Tertychnyi, V., Romushkevich, R., Korobkov, D., and Pohl, J., 2003. Interrelations between thermal conductivity and other physical properties of rocks: experimental data. *Pure and Applied Geophysics*, **160**, 1137–1161.
Ray, L., Förster, H.-J., Schilling, F. R., and Förster, A., 2006. Thermal diffusivity of felsic to mafic granulites at elevated temperatures. *Earth and Planetary Science Letters*, **251**, 241–253.
Reibelt, M., 1991. *Study on the Influence of Surface Structure and Fluid Saturation of Rocks on the Determination of Thermal Conductivity by a Half-space Line Source*. Unpublished diploma thesis (in German). Germany: Institut für Angewandte Geophysik, Petrologie und Lagerstättenforschung, Technische Universität Berlin.
Rosseland, S., 1930. The principles of quantum theory. In Eberhard, G., Kohlschütter, A., and Ludendorff, H. (eds.), *Handbuch der Astrophysik*. Berlin: Springer, Vol. 3(1), pp. 452–474.
Roy, R. F., Beck, A. E., and Touloukian, Y. S., 1981. *Thermophysical Properties of Rocks*. In Touloukian, Y. S., Judd, W. R., and Roy, R. F. (eds.), *Physical Properties of Rocks and Minerals*, McGraw-Hill/CINDAS Data Series on Material Properties, Vol. II-2. New York: McGraw-Hill, pp. 409–502.
Sass, J. H., Lachenbruch, A. H., Moses, T. H., Jr., and Morgan, T., 1992. Heat flow from a scientific research well at Cajon Pass, California. *Journal of Geophysical Research*, **97**(B4), 5017–5030.
Schärmeli, G., 1982. Anisotropy of olivine thermal conductivity at 2.5 GPa up to 1500 K measured on optically non-thick sample. In Schreyer, W. (ed.), *High-Pressure Researches in Geoscience*. Schweizerbart, Stuttgart: Germany, pp. 349–373.
Schilling, F. R., 1999. A transient technique to measure thermal diffusivity at elevated temperatures. *European Journal of Mineralogy*, **11**, 1115–1124.
Schulz, B., 1981. Thermal conductivity of porous and highly porous materials. *High Temperatures-High Pressures*, **13**, 649–660.
Seipold, U., 1998. Temperature dependence of thermal transport properties of crystalline rocks – a general law. *Tectonophysics*, **291**(1–4), 161–171.
Seipold, U., and Schilling, R. R., 2003. Heat transport in serpentinites. *Tectonophysics*, **370**, 147–162.
Shankland, T. J., Nitsan, U., and Duba, A. G., 1979. Optical absorption and radiative heat transport in olivine at high temperature. *Journal of Geophysical Research*, **84**(B4), 1603–1610.
Somerton, W. H., 1992. *Thermal Properties and Temperature Related Behavior of Rock/Fluid Systems*. Amsterdam: Elsevier.
Tommasi, A., Gibert, B., Seipold, U., and Mainprice, D., 2001. Anisotropy of thermal diffusivity in the upper mantle. *Nature*, **411**, 783–786.
Vosteen, H.-D., and Schellschmidt, R., 2003. Influence of temperature on thermal conductivity, thermal capacity, and thermal diffusivity for different types of rock. *Physics and Chemistry of the Earth*, **28**(9–11), 499–509.
Whittington, A. G., Hofmeister, A. M., and Nabelek, P. I., 2009. Temperature-dependent thermal diffusivity of the Earth's crust and implications for magmatism. *Nature*, **458**, 319–321, doi:10.1038/nature07818.
Williams, C. F., and Anderson, R. A., 1990. Thermophysical properties of the Earth's Crust: In situ measurements from continental and ocean drilling. *Journal of Geophysical Research*, **95**(B6), 9209–9236.
Xu, Y., Shankland, T. J., Linhardt, S., Rubie, D. C., Langenhorst, F., and Klasinski, K., 2004. Thermal diffusivity and conductivity of olivine, wadsleyite and ringwoodite to 20 GPa and 1373 K. *Physics of the Earth and Planetary Interiors*, 143–144, 321–336.
Zimmerman, R. W., 1984. *The Effect of Pore Structure on the Pore and Bulk Compressibilities of Consolidated Sandstones*. Ph.D. thesis. Berkeley, CA: University of California.
Zimmerman, R. W., 1989. Thermal conductivity of fluid-saturated rocks. *Journal of Petroleum Science and Engineering*, **3**(3), 219–227.
Zoth, G., and Hänel, R., 1988. Appendix. In Hänel, R., Rybach, L., and Stegena, L. (eds.), *Handbook of Terrestrial Heat Flow Density Determination*. Dordrecht: Kluwer, pp. 449–466.

Cross-references

Heat Flow, Continental
Heat Flow, Seafloor: Methods and Observations
Thermal Storage and Transport Properties of Rocks, I: Heat Capacity and Latent Heat

TIME REVERSAL IN SEISMOLOGY

Carène Larmat[1], Clarence S. Clay[2]
[1]Geophysics Group, EES-17, MS D443, Los Alamos National Laboratory, Los Alamos, NM, USA
[2]Marine Geophysics and Wave Propagation, Department of Geoscience, University of Wisconsin-Madison, Madison, WI, USA

Definition

Time-Reversal – This refers to a collection of techniques that aim to focus wave energy onto a specific point in space and time, implicitly using the *source-receiver reciprocity* and the *invariance* of the wave propagation equation to the transformation $t \rightarrow -t$. The usual way to perform a time-reversal experiment is to simultaneously rebroadcast from a set of receivers the recorded signals read backward. By doing this, an approximate and reverse version of an original wave propagation movie is created, which will display wavefronts converging back onto the source point. A variant of this procedure is *reciprocal Time-Reversal*, where the read backward signals are rebroadcast from the source point instead of from the recording point. Based on the *source-receiver reciprocity*, the wavefield will in this case focus on the receiver.

Invariance of the wave equation – In the absence of loss terms, the wave equation only contains second-order space and time derivatives. This implies that if u(x,t) is a wavefield solution, then u(x,−t), u(−x,t), and u(−x,−t) are also solutions of the equation. This means that the propagation direction of a wavefield may be reversed if it has been measured at any point of space at a given moment. Consideration about the correct boundary conditions to be applied can be found in the subsequent references.

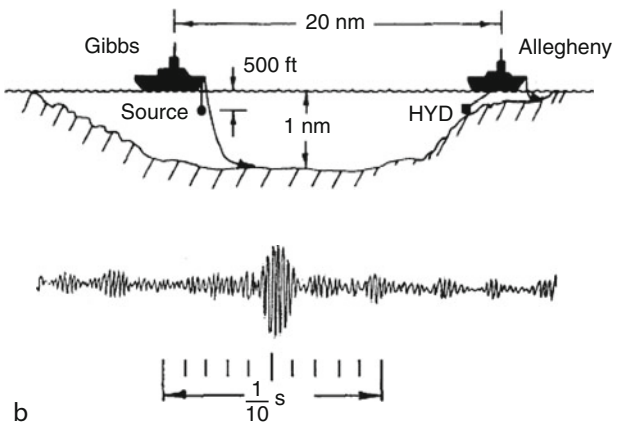
Source-receiver reciprocity – The principle of reciprocity states that a source and receiver may be interchanged and the same waveform will be observed. The conditions required are that the considered signal is transient, nonlinear effects are negligible, and that the medium remains unchanged.

History

Imagine filming the ripples formed when a pebble is dropped into a pond, and playing the movie backward. You will observe the ripples converge back onto the original impact point, indicating its location by a local concentration of wave energy. The history of developing Time-Reversal is strongly related to the ability to physically or numerically realize this thought experiment based on progress in understanding wave propagation and in developing the necessary tools.

The story of Time-Reversal starts in the 1950–1960s in a cluttered room of the Hudson Laboratory of Columbia University where Antares Parvulescu (1923–1998) tested his idea of time-reversed transmissions. A microphone was placed about 25 ft from a loud speaker, from which a series of short impulses were sent 1 s apart. A sequence of the signal due to the room reverberations was received at the microphone and tape-recorded. The tape was removed and played backward. These "time-reversed signals" (called then "matched signals") were played by the loud speaker (notice it is a reciprocal Time-Reversal experiment that was carried out). When an indiscernible rushing noise could be heard over the rest of the room, the scope at the microphone displayed a signal that looked like the autocorrelation of the reverberation signal. When we indeed brought our ears to the microphone, the rushing noise became a simple "click," indicating that localization in space and time of the wave energy did occur! This result was first published in a terse abstract (Parvulescu, 1961) and is further documented in Parvulescu (1995) and in chapter 7 of the book of Tolstoy and Clay (1966 and its revision in 1987).

In 1961–1962, Parvulescu and colleagues decided to take this matched signal experiment to sea. This was a serious logistic and scientific challenge as most of the geophysicists at that time did not believe that transmission in ocean was reproducible. The marine experiment used two oceanographic research ships, the Gibbs and Allegheny, with deep anchoring capability in the Tongue of Ocean, which is an oceanic trench. Figure 1a shows the setting of the experiment. The received signals at the Allegheny, lengthened by multiple reverberations, were radioed back to the Gibbs to be then transmitted from the source (this was again a reciprocal Time-Reversal experiment). Figure 1b shows the peak of pressure created on the hydrophone. The report of this first "field" Time-Reversal experiment is documented in Parvulescu and Clay (1965). Details of further experimentations can be found in chapters 4–8 of Tolstoy and Clay (1966–1987).



Time Reversal in Seismology, Figure 1 **(a)** Ship positions and instruments used for the marine experiment. Gibbs has the source, and Allegheny has the receiver (HYD). **(b)** Signal received at the hydrophone when performing a reciprocal time-reversal between the two ships in an oceanic trench. Figures from Parvulescu and Clay (1965) and Tolstoy and Clay (1966).

In 1980s, Time-Reversal was first used in seismology by George McMechan and colleagues at the University of Texas. They modified techniques developed for exploration geophysics to image the primary source of excitation instead of geologic reflectors and rediscovered the principle of Time-Reversal. They successfully imaged the source of a small 1983 aftershock earthquake in California using a purely acoustical 2D model to numerically rebroadcast signals recorded at a very dense network in a 15 by 12 km vertical slice of the formation (McMechan, 1982).

In the late 1980s, advances in microelectronics, storage capacity, and automated signal processing led to the development of the *Time-Reversal Mirror* (TRM) by Mathias Fink and colleagues at the Laboratoire Ondes et Acoustique (LOA) of the University of Paris VII (Fink, 1992). TRMs are finite-size arrays of transmit-receive transducers that are at the base of active research on theoretical limit of Time-Reversal (Fink et al., 2000; Fink, 2006). TRMs have made possible novel applications of Time-Reversal ranging from nondestructive evaluation (Ulrich et al., 2008a, b) to medical imaging (Fink and Tanter, 2010).

For references and more technical details about Time-Reversal, see the reviews by Fink et al., 2000; Fink, 2006; Anderson et al., 2008; and Larmat et al., 2010.

Time-reversal and complexity

The Earth and oceans are complex environments in which transmitted signals quickly become difficult to interpret, for example, instead of a single pulse for the first arrival, the P-wave is accompanied by a coda train due to the scattering within the crust. Contrary to many location methods based on simple paradigms, Time-Reversal thrives with complexity. In a homogeneous and purely acoustic medium, only one phase will be recorded on a single receiver from a source point. Introduce a reflecting interface and two phases will be recorded: the primary and a copy of the source impulse due to the reflection on the interface. The two combined phases contain more information about the source location than the primary phase alone. All these different phases will converge and add up onto the source point when traveling backward. Several studies (e.g., Dowling and Jackson, 1992; Fink et al., 2000; Blomgren et al., 2002) have actually demonstrated the width of the focus is smaller in randomly heterogeneous media than in the homogeneous case (implying better resolution on the location). This is referred as *super-resolution.* Theoretical discussion about Time-Reversal resolution can be found in Fink et al. (2000). A second important point for complex media is that any arbitrary time-segment will return to the source. Time-Reversal does not need any interpretation of the signals.

These unique facets of Time-Reversal explain the wide use of the method in very different disciplines handling with complex media, for example, target detection method and communication in underwater acoustics, nondestructive evaluation (i.e., location of defects in intact specimens), improved imaging and destruction of stones/tumors in medicine, and source location and imaging techniques in geophysics.

Location of seismic sources

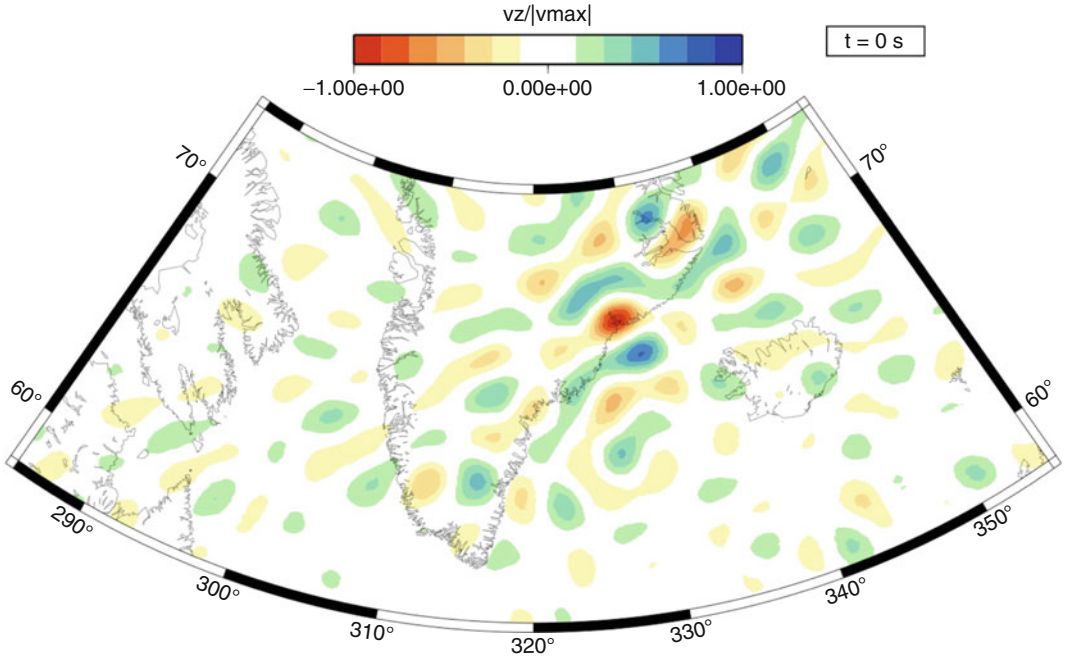
Since the late 1990s, several groups have performed Time-Reversal location of seismic events (for references, see Anderson et al., 2008; Larmat et al., 2010), harnessing the developing computing power. These applications have demonstrated that contrary to what was thought by McMechan, Time-Reversal does work with "sparse" rebroadcasting networks (i.e., with receivers way more than one wavelength apart). Figure 2 shows the time-reversal result from a 2001 magnitude-5 glacial earthquake in Greenland by Larmat et al. (2008). Glacial earthquakes were discovered in 2003 by Ekström et al. (2003) as long-period signal barely above the noise level. The signal from 146 stations in the Northern hemisphere was filtered between 0.01 and 0.02 Hz, time-reversed and rebroadcast into a 3D Earth model. The figure shows the snapshot of the vertical component of the time-reversed displacement at the moment of the focus (actually the vertical component normalized by the maximum value reached on the snapshot). The focus appears as a two-lobe pattern of opposite polarity indicating the origin of the seismic signal. The two-lobe pattern reveals that the nature of the source excitation was a single force vector, which is consistent with recent source analysis by Tsai et al. (2008). Had the source been a simple monopole, the focus would have manifested itself as a single point (see details in Larmat et al., 2008).

In the last 5 years, Time-Reversal application to seismology has evolved into using 3D full-waveform modeling to locate any type of seismic sources (e.g., point source, long lasting, finite) with "no specific interpretation . . . made of the various arrivals" as pointed out by Brian Kennett (1983). One application can be location of the origin of various emergent signals such as tremors (Steiner et al., 2008; Larmat et al., 2009; Lokmer et al., 2009) and weaker microseismicity. It must be noted that several Time-Reversal source location have been carried out of late.

Other location methods are based on the core idea of sending back some recorded signals to a given point in space and time. All back-projection related methods can be seen as simplified Time-Reversal procedures for which the backward propagation is reduced to shifting the time series by predicted arrival times (e.g., Source Scanning Algorithm of Kao et al. (2005).

Location of features in the velocity model

Time-reversed propagation is also used to image the subsurface structure. The exploration methods used by McMechan have evolved into reverse-time migration (e.g., Baysal et al.,1983). Next generation tomography



Time Reversal in Seismology, Figure 2 One hundred and forty-six vertical displacement signals of a glacial earthquake were time-reversed and sent back in a 3D Earth model. The vertical component (vz) of the created time-reversal signal at the moment of the focus displays a two-lobe pattern of opposite polarity consistent with a single force vector as the source mechanism (adapted from Larmat et al., 2008).

models are currently created based on the adjoint method that uses a forward and a backward propagation to locate the source of discrepancies between data and predicted waveforms (Tromp et al., 2005).

Time-reversal and correlation methods

The signal produced at point A from source S can be expressed with the Green function formalism: $S_A(t) = G(x_S, x_A, t) \otimes s(t)$ where $s(t)$ is the source time function and $\otimes$ denotes the convolution operation. If you time-reverse the signal $S_A(t)$, and rebroadcast it from A, the signal produced on point S will be:

$$\begin{aligned} S_S^{TR}(t) &= G(x_A, x_S, t) \otimes S_A(T-t) \\ &= [G(x_S, x_A, T-t) \otimes G(x_S, x_A, -(T-t)] \otimes s(t) \end{aligned} \quad (1)$$

where the reciprocity between A and S of the Green function has been used. The right term of this expression is similar to the expression of the autocorrelation of S_A in terms of Green functions. More generally, the cross-correlation of the signal recorded at a point A and B is:

$$\begin{aligned} C_{AB}(t) = [G(x_S, x_A, t) \otimes G(x_S, x_B, -t)] \otimes [s(T-t) \otimes s(t)] \\ \rightarrow S[G(x_A, x_B, t) + G(x_A, x_B, -t)] \end{aligned} \quad (2)$$

which states that under certain conditions the cross-correlation of two signals is composed of the superposition of the Green function between the two receivers and its time-reversed version. Equation (2) is the basis of what is known as *interferometry* or imaging with ambient noise, as the Green function contains information about the properties of the medium (see *Seismic, Ambient Noise Correlation*, the review by Snieder and Wapenaar, 2010 and the pioneer paper of Claerbout, 1968).

To come back to Time-Reversal, if the source time function is considered to be a dirac, the time-reversed signal at the original point source S is:

$$S_S^{TR}(t) = [G(x_S, x_A, T-t) \otimes G(x_S, x_A, -(T-t)] \quad (3)$$

This expression is typical of a *matched filter*, whose output is optimal in some sense. Whatever the Green function, the Time-Reversal wavefield is maximal at the focus time and at the source point (Fink et al., 2000, p. 1946). Applications derived from this interpretation of Time-Reversal consist in locating seismic sources, such as trapped miners, by comparing their transmitted distress signals with a database of prerecorded Green functions (created by recording the seismic response for each possible source location) (Hanafi et al., 2009).

Summary

Interest in seismic source study has recently shifted from event-based temporary observation to continuous

monitoring using dense seismic networks. Seismologists now study a wealth of signals due to less energetic sources than earthquakes such as tremors and microseismicity. Glacial earthquakes, landslides, Earth Hum, and meteorological phenomena such as hurricanes belong to a group of events that defines "environmental seismology," a term coined by the French seismologist Jean-Paul Montagner.

Microseismic noise is now used for imaging the Earth. Due to their emergent nature, these signals represent a challenge to classical location methods based on triangulation. Time-Reversal as a method and Time-Reversal ideas are emerging as a promising framework to foster future location and imaging methods. The success of Time-Reversal as a source location method shall rely on the coverage of seismic networks, the accuracy of high-resolution velocity models, and the performance of numerical schemes to be used for the backward propagation. Still unclear are the convergence and resolution power of Time-Reversal under various conditions. Our ability to reliably observe and model scattering phenomena is expected to be an important part of the answer.

Acknowledgment

We gratefully acknowledge the support of the U.S. Department of Energy through the LANL/LDRD Program for this work.Carène Larmat & Clarence S. Clay.

Bibliography

Anderson, B., Griffa, M., Larmat, C., Ulrich, T. J., and Johnson, P. A., 2008. Time reversal. *Acoustics Today*, **4**(1), 5.

Baysal, E., Kosloff, D. D., and Sherwood, J. W. C., 1983. Reverse time migration. *Geophysics*, **48**(11), 1514–1524.

Blomgren, P., Papanicolaou, G., and Zhao, H., 2002. Super-resolution in time-reversal acoustics. *The Journal of the Acoustical Society of America*, **111**(1), 230–248.

Claerbout, J. F., 1968. Synthesis of a layered medium from its acoustic transmission response. *Geophysics*, **33**, 264–269.

Dowling, D. R., and Jackson, D. R., 1992. Narrow-band performance of phase-conjugate arrays in dynamic random media. *The Journal of the Acoustical Society of America*, **91**, 3257–3277.

Ekström, G., Nettles, M., and Abers, G., 2003. Glacial earthquakes. *Science*, **302**, 622–624.

Fink, M., 1992. Time-reversal of ultrasonic fields–Part I: basic principles. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, **39**(5), 555–566.

Fink, M., 2006. Time-reversal acoustics in complex environments. *Geophysics*, **71**(4), SI151–SI164.

Fink, M., and Tanter, M., 2010. Multiwave imaging and super resolution. *Physics Today*, **63**(2), 28–33.

Fink, M., Cassereau, D., Derode, A., Prada, C., Roux, P., Tanter, M., Thomas, J.-L., and Wu, F., 2000. Time-reversed acoustics. *Reports on Progress in Physics*, **63**(12), 1933–1995.

Hanafi, S. M., Cao, W., McCarter, K., and Schuster, G. T., 2009. Using super-stacking and super-resolution properties of time-reversal mirrors to locate trapped miners. *The Leading Edge*, **28**(3), 302–307.

Kao, H., Shan, S.-J., Dragert, H., Rogers, G., Cassidy, J. F., and Ramachandran, K., 2005. A wide depth distribution of seismic tremors along the northern Cascadia margin. *Nature*, **436**, 841–844.

Kennett, B. L. N., 1983. A new way to estimate source parameters. *Nature*, **302**, 659–660.

Larmat, C., Tromp, J., Liu, Q., and Montagner, J.-P., 2008. Time reversal location of glacial earthquakes. *Journal of Geophysical Research*, **113**, B09314.

Larmat, C. S., Guyer, R. A., and Johnson, P. A., 2009. Tremor source location using time reversal: selecting the appropriate imaging field. *Geophysical Research Letters*, **36**, L22034.

Larmat, C. S., Guyer, R. A., and Johnson, P. A., 2010. Time-reversal methods in geophysics. *Physics Today*, **63**, 31–35.

Lokmer, B., O'Brien, G. S., Stich, D., and Bean, C. J., 2009. Time reversal imaging of synthetic volcanic tremor sources. *Geophysical Research Letters*, **36**, L12308.

McMechan, G. A., 1982. Determination of source parameters by wavefield extrapolation. *Geophysical Journal of the Royal Astronomical Society*, **71**, 613–628.

Parvulescu, A., 1961. MESS processing. *The Journal of the Acoustical Society of America*, **33**, 1674.

Parvulescu, A., 1995. Matched-signal (MESS) processing by the ocean. *The Journal of the Acoustical Society of America*, **98**, 943–960.

Parvulescu, A., and Clay, C. S., 1965. Reproducibility of signal transmissions in the ocean. *Radio Engineering and Electronic Physics*, **29**(4), 223–228.

Snieder, R., and Wapenaar, K., 2010. Imaging with ambient noise. *Physics Today*, **63**(9), 44–49.

Steiner, B., Saenger, E. H., and Schmalholz, S. M., 2008. Time reverse modeling of low-frequency microtremors: application to hydrocarbon reservoir localization. *Geophysical Research Letters*, **35**, L03307.

Tolstoy, I., and Clay, C. S., 1966. *Ocean Acoustics, Theory and Experiment in Underwater Sound*. McGraw-Hill (Translated into Russian and published, 1969. Revised-edition 1987, the Acoustical Society of America – the American Institute of Physics).

Tromp, J., Tape, C., and Liu, Q., 2005. Seismic tomography, adjoint methods, time reversal, and banana-doughnut kernels. *Geophysical Journal International*, **160**, 195–216.

Tsai, V. C., Rice, J. R., and Fahnestock, M., 2008. Possible mechanisms for glacial earthquakes. *Journal of Geophysical Research*, **113**, F03014.

Ulrich, T. J., Sutin, A. M., Guyer, R. A., and Johnson, P. A., 2008a. Time reversal and non-linear elastic wave spectroscopy (TR NEWS) techniques. *International Journal of Non-Linear Mechanics*, **43**, 209–216.

Ulrich, T. J., Sutin, A. M., Claytor, T., Pallin, P., Le Bas, P.-Y., and TenCate, J. A., 2008b. The time reversal elastic nonlinearity diagnostic applied to the evaluation of diffusion bonds. *Applied Physics Letters*, **93**, 151914.

Cross-references

Earthquake, Focal Mechanism
Earthquake, Location Techniques
Propagation of Elastic Waves: Fundamentals
Seismic Data Acquisition and Processing
Seismic Tomography
Seismic, Ambient Noise Correlation
Seismic, Migration
Seismic, Waveform Modeling and Tomography
Seismological Networks

TRAVELTIME TOMOGRAPHY USING CONTROLLED-SOURCE SEISMIC DATA

Colin A. Zelt
Department of Earth Science, Rice University, Houston, TX, USA

Synonyms

Active-source data; Traveltime inversion

Definition

Traveltime. The total time for a particular wave to travel from a source to a receiver.
Tomography. Determination of model structure by back projection of the data along a path connecting a source and a receiver.
Controlled-source seismology. Determination of earth structure using seismic waves generated by artificial sources such as chemical explosives, mechanical vibrators, weight drops, gun shots, etc.

Background

Traveltime tomography is the main method by which the Earth's seismic velocity structure is determined on all scales, from the upper few meters to the whole mantle. It was adapted from algorithms used in medical imaging in the 1970s (Dziewonski and Anderson, 1984). In seismology, the term tomography refers to the back projection of data along a path connecting a source and receiver using a mathematical inverse method. Ideally, the traveltimes of seismic arrivals corresponding to many criss-crossing rays are used to construct a two-dimensional (2D) or 3D image of the Earth's seismic velocity variations. Tomography is a type of inverse problem, although the terms traveltime inversion and traveltime tomography are used in the seismological community without a well-defined distinction. The latter usually implies a uniform, fine grid model parameterization, in which a smooth model is sought. Any departure from this approach is usually referred to as traveltime inversion, although the distinction is somewhat arbitrary (Levander et al., 2007); this article will discuss both approaches. This article concerns traveltime tomography using controlled-source data and does not discuss other types of tomography that use seismic amplitudes, phases, or waveforms as input data, nor earthquake or surface-wave data.

Data from earthquakes and artificial (controlled) sources are used in traveltime tomography, the former usually for crustal or mantle structure, the latter usually for crustal or near-surface structure. Much of the theory and practice of traveltime tomography is the same or similar for controlled-source and earthquake data. However, earthquake tomography usually uses delay times and includes a determination of the source locations (hypocenters), whereas controlled-source tomography usually uses total times and the source locations are known very accurately. Also, the degree of lateral velocity heterogeneity in the crust, and in particular the near-surface, can be much greater than in the mantle, so there is often no sense of a reference model in controlled-source seismology as there is in earthquake seismology, and this in turn means that a nonlinear tomographic method must be applied to controlled-source data in which the rays must be determined along with the unknown velocity structure as part of the inverse problem. Finally, controlled-source data are often acquired using sources and receivers distributed along a straight or nearly straight line, yielding a so-called 2D dataset. In this case, 2D modeling is used in which lateral homogeneity perpendicular to the line is assumed. Earthquake tomography is almost always 3D given the typical distribution of earthquakes and stations. Today, 3D controlled-source data is not uncommon, but it is expensive to acquire, especially for crustal-scale studies. In some studies, both earthquake and controlled-source data have been inverted simultaneously (e.g., Ramachandran et al., 2005).

Controlled-source seismology can be divided into two main approaches: (1) near-vertical reflection methods, and (2) refraction/wide-angle reflection methods. Traveltime tomography is applied to data from both experiments, although it is more common for refraction data. The most common applications of traveltime tomography using controlled-source data are (1) tectonic studies of the crust and uppermost mantle, (2) earthquake risk studies for upper crustal fault and basin geometries, (3) petroleum exploration in crosswell studies, model building to facilitate migration, and for refraction statics, and (4) environmental/engineering studies of the near surface (as shallow as a few meters). The traveltimes utilized are often limited to those of the first arrivals corresponding to direct, refracted (turning) or diffracted waves. Sometimes the traveltimes of primary reflections are included, for example, from the basement overlying sediments or the crust-mantle (Moho) boundary, necessitating the need for layers or interfaces in the model parameterization. Only rarely are later refracted arrivals, multiples or conversions utilized. Traveltimes are picked from the seismograms usually either interactively by eye, or using a semi-automated scheme whereby a few picks are made interactively, and the intervening picks are determined automatically using a cross-correlation scheme (Zelt, 1999). Usually only very high quality, spatially-dense data can be picked using fully automated schemes.

Most controlled-source tomography uses only P-wave traveltimes to constrain the compressional wave-velocity structure because P wave are easiest to generate and therefore the P wave arrivals are usually the strongest and easiest to pick. However, S wave experiments are quite common in environmental/engineering studies (e.g., Chambers et al., 2009). Theoretically, traveltime tomography is the same for P- and S-wave data, and if both data are available in the same study, Poisson's ratio can be estimated. The forward modeling component of traveltime tomography is almost always based on ray theory, an infinite-frequency approximation of wave propagation, although recently a finite-frequency approach for

controlled-source data has been proposed (Zelt, 2009). Almost all applications of traveltime tomography assume isotropic media because it is rarely necessary to invoke anisotropy to model the observed data. Exceptions to this are mainly for crosswell data and in studies of the oceanic crust (e.g., Pratt and Chapman, 1992; Dunn et al., 2000). Traveltime tomography is applied to land and marine data, as well as onshore-offshore data, using the same tomographic methodology. Arrays of airguns are the most common marine source, while on land chemical explosives, mechanical vibrators, weight drops, hammers, shotguns and rifles are common sources.

The ideal dataset for traveltime tomography is one with as dense a spacing of sources and receivers as possible, along a line for a 2D experiment, or areally distributed for a 3D experiment. As a result, a subsurface point will be sampled by rays at as many different angles as possible, which in turn will yield the best possible spatial model resolution as discussed later. There is usually very little processing of seismic data before traveltime picking, only that required to "clean up" the data to make it easier for picking. This may include bandpass filtering, trace editing, velocity filtering, trace mixing or binning, or deconvolution. Picks should not be interpolated to provide uniform spatial coverage if there are significant trace gaps, since this will provide an incorrect sense of model constraint (Zelt, 1999). Uncertainties should be assigned to the pick times to avoid over- or under-fitting the data, and to allow the appropriate up- and down-weighting of less noisy and more noisy data, respectively (Zelt, 1999). Elevation or bathymetric corrections should be avoided as they depend on the unknown velocity structure. Instead, the known surface topography or bathymetry should be incorporated into the 2D or 3D model. For a 2D experiment in which the shot and receiver locations deviate significantly from a straight line, it may be necessary to perform 2.5D modeling, i.e., using a 3D model and 3D ray tracing, and the known topographic or bathymetric surfaces, but keeping the velocity model homogeneous in a direction perpendicular to the main trend of the profile (e.g., Van Avendonk et al., 2004). Given the spatial resolution of most velocity models derived from traveltime tomography, the degree of profile "crookedness" must be substantial to warrant 2.5D modeling (Schmelzbach et al., 2008; Figure 1).

There are several review papers relevant to this article. Nowack and Braile (1993) review traveltime tomography methods in 1D and 2D media. Zelt (1999) focuses on 2D traveltime tomography and inversion, as well as model assessment methods. Rawlinson and Sambridge (2003a) present an overview of traveltime tomography for 2D and 3D structure. Levander et al. (2007) review all aspects of crustal-scale controlled-source seismology, including traveltime tomography.

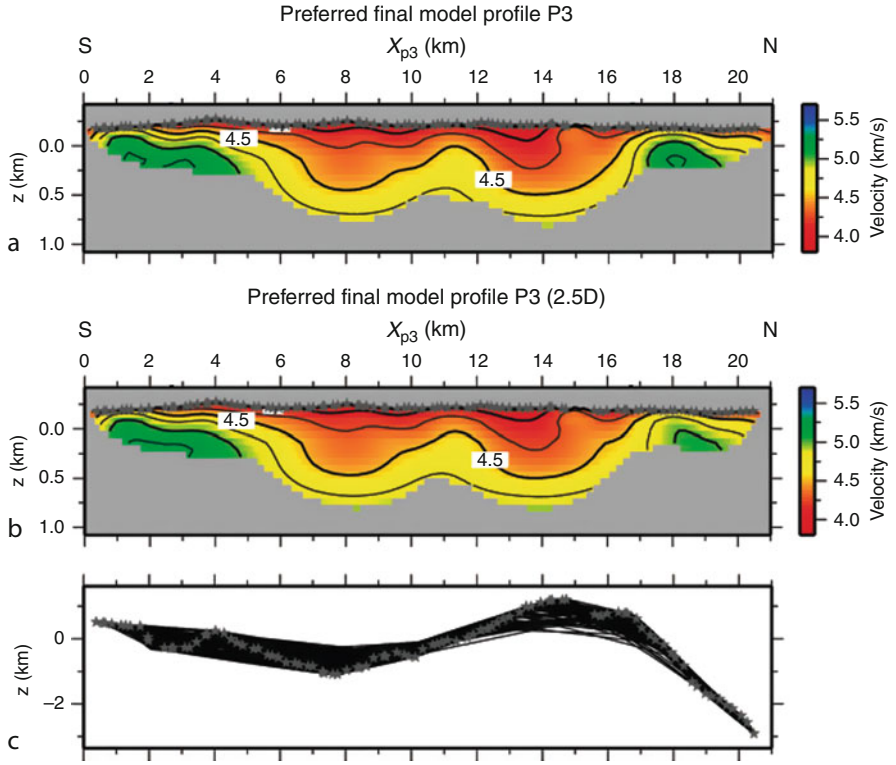
Forward modeling

In the 1970s, ray theory, an infinite-frequency approximation of wave propagation, for laterally heterogeneous media was developed (Cerveny et al., 1977, McMechan and Mooney, 1980). This allows earth models of arbitrary 2D and 3D complexity to be considered, provided the tenants of ray theory are honored, that is, the velocity field varies slowly with respect to the seismic wavelengths. Sharp velocity changes are modeled as velocity discontinuities (layer boundaries) using Snell's law. As a result, ray theory models are typically composed of one or a few layers, bounded by smoothly varying interfaces, within which the velocity varies smoothly. Ray theory forms the basis for the forward modeling used today in almost all traveltime tomography algorithms. There are two main types of infinite-frequency modeling: ray tracing and wavefront tracking.

Two end-member approaches to ray tracing are used: numerical and analytical. In the former case, the velocity field is specified by a regular (e.g., McMechan and Mooney, 1980) or irregular (e.g., Zelt and Ellis, 1988) grid of nodes, together with a scheme for interpolation between nodes, and a ray is traced by specifying a take-off angle from a point source and solving a set of ordinary differential equations, the ray tracing equations (Cerveny et al., 1977). In the analytic case, the velocity field is specified by a regular (e.g., Chapman and Drummond, 1982; White, 1989; Rawlinson et al., 2001) or irregular (e.g., Spence et al., 1984) grid of nodes, but with a form of interpolation, for example linear, such that the rays can be calculated analytically within each cell of the grid. In practice, the numerical and analytical techniques may be about equally efficient if the analytical method uses a fine grid of nodes to represent a sufficiently smooth velocity field as required by ray theory, and if the numerical scheme adjusts the step length according to the velocity gradient (Zelt and Ellis, 1988).

For an arbitrary 2D velocity model, a non-trivial consideration is the determination of ray take-off angle from the source such that an arbitrary receiver location is reached, or such that a particular layer or interface in the model is reached. Today, most 2D ray tracing algorithms offer an automatic determination of take-off angles by shooting trial rays followed by a bisection scheme to sufficiently refine the take-off angles so that the receivers are hit (e.g., Zelt and Smith, 1992). The robustness of the shooting method is crucial because the data from receivers for which it is not possible to find rays will not be used in the inverse step. Also, it is important to be sure that an inability to find rays is because of the velocity model, e.g., a shadow zone, as opposed to a shortcoming of the ray tracing algorithm.

To avoid the limitations of shooting/bisection algorithms, particularly for 3D models, ray bending methods were developed (Um and Thurber, 1987). In this approach, a ray connecting a source and receiver is estimated, typically using a 1D reference model, and the ray's path is iteratively updated using Fermat's principle until the minimum-time path is determined. In practice, shooting methods are more efficient and sufficiently robust for 2D models, whereas bending methods are favored for 3D models.



Traveltime Tomography Using Controlled-Source Seismic Data, Figure 1 Shallow crustal example from the South Portuguese Zone fold-and-thrust belt of first-arrival tomography by Schmelzbach et al. (2008) using the smoothing-regularized Zelt and Barton (1998) algorithm to seek a minimum-structure model. There were 99 shots (grey stars) along the crooked line. (**a**) Velocity model obtained by projecting all the data onto a straight line and performing a 2D inversion. (**b**) Velocity model obtained by performing a 2.5D inversion that preserves the true positions of all shots and receivers. (**c**) Rays projected into the horizontal plane for the 2.5D modeling to illustrate the extreme crookedness of the experiment. Model regions not sampled by rays are grey. The models in (**a**) and (**b**) are very similar, confirming the appropriateness of 2D modeling of traveltime data, even in the case of extreme crooked lines.

In the late 1980s, a new infinite-frequency forward modeling approach was introduced in which first-arrival traveltimes are calculated on a fine grid using a finite-difference solution of the eikonal equation (Vidale, 1988, 1990; Podvin and Lecomte, 1991). Rays are obtained by following the gradient of the time field. These methods are known as wavefront tracking algorithms. Reflected rays can be calculated in a two-step procedure by calculating downgoing and upgoing wavefronts (Hole and Zelt, 1995). Hole and Zelt (1995) also presented a modification to Vidale's approach which otherwise breaks down for velocity contrast of more than ~40%. The advantage of wavefront tracking methods is that they find the ray between any two points with the shortest traveltime, including the diffracted path in the case of a geometrical shadow zone, and they can be very efficient, especially for 3D models. The disadvantages are that they are computationally cumbersome for calculating later arrivals, and they can be inaccurate and/or computationally intensive for high-contrast media. Another type of wavefront tracking scheme that solves the eikonal equation is the fast-marching method, which is computationally efficient and unconditionally stable (Sethian and Popovici, 1999).

Nakanishi and Yamaguchi (1986) and Moser (1991) introduced the shortest path ray-tracing method based on Fermat's principle which uses a fine network of nodes and graph theory. It is capable of handling arbitrarily heterogeneous media and calculating first arrivals, reflections, and multiples. Its advantages over other methods are robustness and the fact that the grid does not have to be rectangular or regular. However, its memory and computation requirements are greater. Van Avendonk et al. (2001) developed a hybrid of the shortest path and ray bending methods that uses the shortest path algorithm to find an initial ray, and then refines it using a bending method. The result is an efficient and robust algorithm to calculate accurate traveltimes and rays for refractions and reflections in 2D and 3D media.

Traveltime tomography and inverison

The traveltime t between a source and receiver along a ray L is given in integral form for a velocity field $v(\mathbf{r})$ as

$$t = \int_L \frac{1}{v(\mathbf{r})} \mathrm{d}l$$

where $\mathbf{r}$ is the position vector in 2D or 3D media. This is a nonlinear inverse problem given the relationship between the measured data (traveltimes) and the unknown model parameters (the velocity field). However, by transforming variables to use slowness $s(\mathbf{r})$, the reciprocal of velocity, instead of velocity as the unknown, a seemingly linear inversion problem is created:

$$t = \int_L s(\mathbf{r}) \mathrm{d}l.$$

However, the ray L is also dependent on the velocity (or slowness) model, thus making the inverse problem nonlinear regardless of what form of model variable or parameterization is used. In global seismology, a 1D (radially-symmetric) earth model is often used to predict the rays given the relatively small velocity heterogeneities in the mantle, and thereby converting the inverse problem into a linear one. In controlled-source seismology, there is no concept of a reference model that is sufficiently accurate to predict the rays in advance given the relatively large velocity heterogeneities that are typically encountered in the crust. This means the controlled-source tomography or inverse problem is always treated as a nonlinear one. This also means the model can be parameterized any number of ways using velocity or slowness, and cells, nodes, or splines, since the problem's nonlinearity must be dealt with regardless of the parameterization. Most often a linearized gradient approach is applied in which a starting model is used and both the model and rays are updated over a series of iterations with the hope that there will be convergence to an acceptable model (the final model).

The model is almost always discretized using cells, nodes, splines, or other interpolating functions; in the latter two cases, the discrete model parameters are the coefficients of the interpolating functions. In the simplest formulation of the tomography problem, the model is parameterized using constant-slowness cells, in which case the equation for the ith data becomes

$$t_i = \sum_j l_{ij} s_j$$

where l_{ij} is the length of the ith ray in the jth model cell and s_j is the slowness in the jth cell. In vector–matrix form this is

$$\mathbf{t} = \mathbf{L}\mathbf{s}.$$

Applying a Taylor series expansion to this equation keeping only the linear term and assuming the rays are independent of slowness yields the perturbation equation

$$\delta\mathbf{t} = \mathbf{G}\delta\mathbf{s}$$

where $\delta\mathbf{t}$ is the data misfit vector equal to the difference between the observed traveltimes and those predicted by a prior model, and $\delta\mathbf{s}$ is the difference between the unknown slowness model and the prior slowness model; the unknown model is also called the estimated model. The partial derivative matrix, $\mathbf{G}$, contains the elements $g_{ij} = \partial t_i / \partial s_j$, and for the constant-slowness cell parameterization, $g_{ij} = l_{ij}$, or $\mathbf{G} = \mathbf{L}$. In the general case, the elements of the partial derivative matrix are $g_{ij} = \partial t_i / \partial m_j$ where m_j is the jth model parameter, which could be the velocity at a node or the coefficient of an interpolating function, or the position of an interface within the velocity model (e.g., Zelt and Smith, 1992). In the general case, the perturbation equation becomes

$$\delta\mathbf{t} = \mathbf{G}\delta\mathbf{m}.$$

The elements of the partial derivative matrix $\mathbf{G}$ are usually calculated analytically to avoid the potential inaccuracy of numerical differencing and the extra computation needed to trace additional rays. The approximations involved in the analytic partial derivatives, e.g., the stationary ray assumption, are typically not a problem since the resultant gradient direction in model space will be improved over a series of iterations.

The perturbation equation is not solved directly because the unknown model parameters are typically under- or mix-determined (Menke, 1989), depending on the model parameterization; with relatively few model parameters, it is possible for the problem to be overdetermined, but even in this case the perturbation equation is not solved directly because it is wise to constrain the magnitude of the model perturbation to avoid violating the linearization assumption. Thus, model constraints, in addition to the observed data, are usually included to stabilize the solution. Also, all observed data contain noise and to avoid over-fitting noisy data, additional model constraints are required to select one model from an infinite number that will statistically predict the observed data at the equivalent desired misfit according to the estimated noise level. This issue points out the non-uniqueness of all inverse problems involving real data, regardless of whether the system of equations is overdetermined or underdetermined.

The additional model constraint is called regularization (e.g., Scales et al., 1990), and in addition to stabilizing a solution, it is an effective way of steering the solution towards models with desirable pre-defined characteristics. The regularization is most often in the form of the zero-, first- or second-order spatial derivatives of the estimated model parameters, or their perturbation from a background model. This is because it is often desirable to seek an estimated model that is as close as possible to another model, and/or a model that has a minimum structure as measured by its spatial derivatives in keeping with Occams's principle (Constable et al., 1987). Regularization is always arbitrary, and the specific form of

regularization that is best for a particular problem depends on the model parameterization, the data, including its coverage and noise level, and the experimental objectives.

The most common forms of regularization in traveltime problems are the minimization of the model's perturbation from a background model, and/or the minimization of the second-order spatial derivative of either the model, or its perturbation from a background model. A model with minimum perturbation is called the smallest model; a model with minimum second-order derivative is called the smoothest model. Any combination of model constraints may be included as a part of the regularization to form the objective function, $\Phi(\mathbf{m})$, where $\mathbf{m}$ is the vector containing the estimated model parameters. The objective function usually measures the square of the data misfit vector and the square of the perturbation or spatial derivative operators comprising the regularization so that minimizing the objective function leads to a linear system of equations that is amenable to efficient algorithms for their exact or approximate solution. One example of an objective function for the traveltime problem is

$$\Phi(\mathbf{m}) = \delta\mathbf{t}^{\mathrm{T}}\mathbf{C}_d^{-1}\delta\mathbf{t} + \lambda\left\{\beta\left[\alpha\left(\mathbf{v}^{\mathrm{T}}\mathbf{W}_h^{\mathrm{T}}\mathbf{W}_h\mathbf{v}+s_z\mathbf{v}^{\mathrm{T}}\mathbf{W}_v^{\mathrm{T}}\mathbf{W}_v\mathbf{v}\right)\right.\right.$$
$$\left.\left.+(1-\alpha)\Delta\mathbf{v}^{\mathrm{T}}\mathbf{W}_p^{\mathrm{T}}\mathbf{W}_p\Delta\mathbf{v}\right]+(1-\beta)\left[\mathbf{z}^{\mathrm{T}}\mathbf{W}_z^{\mathrm{T}}\mathbf{W}_z\mathbf{z}\right]\right\}$$

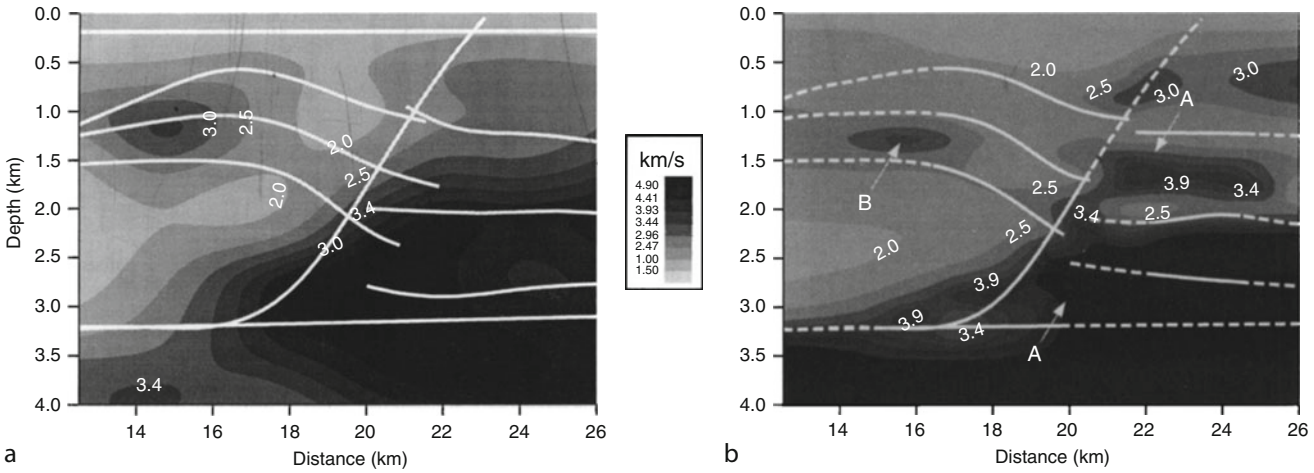
where $\mathbf{m} = \mathbf{v} + \mathbf{z}$; $\mathbf{v}$ and $\mathbf{z}$ are vectors containing the estimated model parameters defining the velocity (or slowness) field and the interfaces (if any); $\Delta\mathbf{v}$ is the velocity perturbation vector equal to $\mathbf{v}$-$\mathbf{v}_{\mathrm{o}}$, and $\mathbf{v}_{\mathrm{o}}$ is the background velocity model vector. $\mathbf{C}_{\mathrm{d}}$ is the data covariance matrix containing the estimated pick uncertainties providing the appropriate up-weighting and down-weighting of the low-noise and high-noise data, respectively. $\mathbf{W}_{\mathrm{h}}$ and $\mathbf{W}_{\mathrm{v}}$ are the horizontal and vertical roughness matrices containing the second-order spatial finite-difference operators that measure the roughness of the velocity field in the horizontal and vertical directions, respectively; $\mathbf{W}_{\mathrm{p}}$ is the velocity perturbation weighting matrix which is a diagonal matrix containing the weights applied to the perturbation between each estimated and background velocity model parameter; $\mathbf{W}_z$ is the interface roughness matrix containing the second-order spatial finite-difference operators that measure the roughness of the interfaces. When applying spatial derivative operators to the velocity field, it is common to use different operators to measure the horizontal and vertical structure, and weight these differently, since one would normally expect the earth to contain more heterogeneity vertically compared to horizontally. Any or all of these operators may be normalized by the reference model values so that relative quantities are penalized as opposed to absolute values (e.g., Toomey et al., 1994). In addition, the operators can be weighted according to their spatial position in the model, e.g., penalize model structure more in the deep portion of the model versus the shallow portion since one would generally expect decreasing resolution with depth.

There are four free parameters in this example objective function to control the relative weight of each term. λ determines the overall amount of regularization, that is, the trade-off between fitting the data and constraining the model; β determines the trade-off between velocity and interface regularization; α determines the trade-off between second derivative and perturbation regularization of the velocity parameters; and s_z specifies the relative weight of vertical versus horizontal smoothing regularization of the velocity parameters. In some algorithms, λ is not a free parameter because it is reduced automatically by the algorithm at each iteration from a free-parameter starting value, λ_{o} (e.g., Zelt and Barton, 1998). The reduction of λ stabilizes the inversion by constraining the long-wavelength structure in the initial iterations and allowing progressively finer model structure to enter in later iterations.

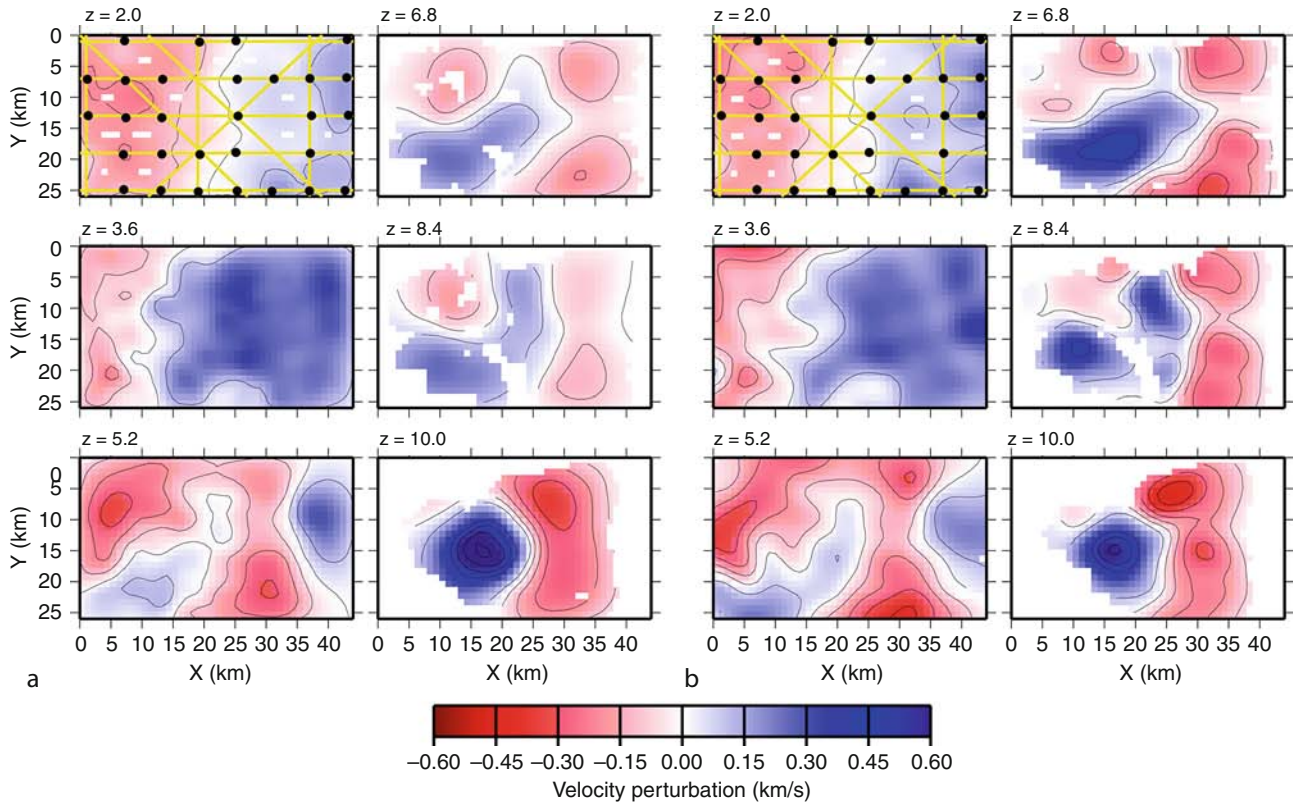
Minimizing the objective function with respect to the unknown model parameters ($\mathbf{v}$ and $\mathbf{z}$) leads to a linear system of equations that may be relatively small or very large, depending on the number of data and model parameters. If the linear system is large, it will also typically be very sparse (at least 99.9%), in which case there are efficient algorithms for their solution (e.g., Paige and Saunders, 1982; Rawlinson et al., 2001). If the linear system is relatively small, it can be solved directly using standard matrix inversion routines, such as LU decomposition (Press et al., 1992).

The objective function presented is designed to illustrate in a single example the different types of regularization that are most commonly applied. In practice, the objective function will typically be simpler depending on the data, model parameterization, and experimental goal. This objective function can be simplified or modified to correspond to those used in most of the popular traveltime inversion and tomography algorithms used today. For example, the Zelt and Smith (1992) inverse method includes only perturbation constraint on the velocity and interface parameters with respect to the model from the previous iteration; the Zelt and Barton (1998) tomographic method includes only smoothness constraints on the velocity field. Thus, together with their very different model parameterizations to be discussed in the next section, these two algorithms can been viewed as end-members of the same regularized inverse approach, and most algorithms in use today fall somewhere in between. In fact, the main differences between the algorithms discussed in the next section lies not in the details of the objective function, but more so in the form of model parameterization, forward calculation, and the types of arrivals considered.

For layered models, including both velocity and interface parameters in a single inversion scheme, as opposed to solving for each parameter separately, either one after the other, or by alternating between the two, makes the



Traveltime Tomography Using Controlled-Source Seismic Data, Figure 2 Synthetic example of reflection tomography for a smooth velocity model by Lailly and Sinoquet (1996) based on real data from a passive margin with a salt structure and a listric fault. (**a**) True model. The reflectors (white curves) are embedded in the smooth velocity field and are allowed to cross each other. (**b**) Estimated model using reflections from the continuous portions of the reflectors (white curves); the parts of the reflectors represented by dashed lines are not illuminated by rays.



Traveltime Tomography Using Controlled-Source Seismic Data, Figure 3 Example of 3D first-arrival tomography applied to ocean-bottom-seismometer (OBS) data from the Faeroe Basin to find the minimum-structure model (Zelt and Barton, 1998). (**a**) Model obtained using smoothing-regularized tomography. (**b**) Model obtained using back-projection tomography method of Hole (1992). Horizontal slices of velocity perturbations with respect to a 1D starting model are shown at the depths labeled from 2 to 10 km. The OBS locations and shot lines are overlaid on the z = 2 km slice. Regions not sampled by rays are blank. Contour interval is 0.1 km/s. Both models provide the same level of fit to the data, but the model from back-projection contains more structure overall. The high-velocity anomaly at z = 10 km likely indicates a basement high.

problem significantly more challenging. There will be more model non-uniqueness because of trade-offs between the two parameter types, and more prior information will be required to deal with the increased ambiguity. Also, more testing will be required to ensure that the trade-offs are fully understood. But solving for both parameter types simultaneously allows one to explore the trade-offs more thoroughly and is the only way to ensure the overall minimum-structure model is obtained, if that is the objective.

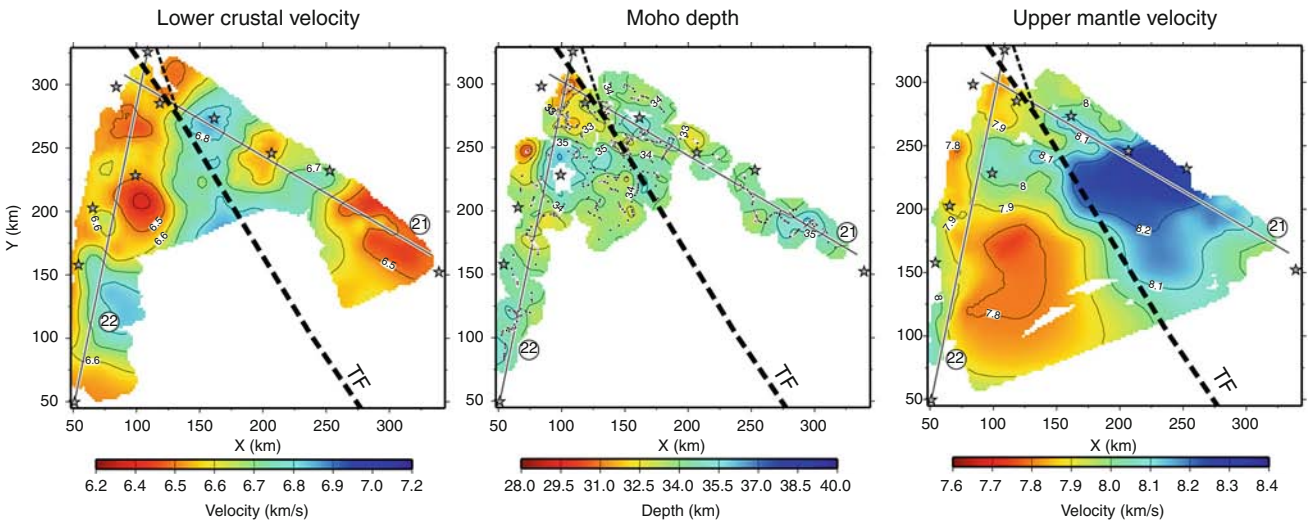
Assuming uncorrelated, Gaussian picking errors, and a model parameterization that does not limit the degree to which the predicted data can match the variability of the observed data, a final model should be chosen that provides a normalized misfit, χ^2, of one (Bevington, 1969).

$$\chi^2 = \frac{1}{N}\left[\delta\mathbf{t}^{\mathrm{T}}\mathbf{C}_d^{-1}\delta\mathbf{t}\right]$$

where N is the number of data. A model that provides a χ^2 value less than one means the observed data are being over-fit, and therefore to some extent, the noise in the data is also being fit. Although the χ^2 criterion is a good rule of thumb, there are several instances in which it should not be followed. First, traveltime picking errors are likely not uncorrelated. Second, for spatially sparse data from a region with strong lateral heterogeneities, it may be necessary to use so many independent model parameters in order to achieve $\chi^2 = 1$ that the constraint on some parameters is unacceptably small, in which case fewer parameters may be preferable and a χ^2 value greater than one is allowed (Zelt, 1999). An inability to achieve $\chi^2 = 1$ may be because of inconsistent picking, for example for reciprocal pairs, in which case the inconsistencies should either be corrected or used as the basis for assigning the pick uncertainties in the first place (e.g., Zelt et al., 2006a).

Algorithms

This section describes many of the algorithms available today, each with its own capabilities and limitations. Some of the earliest 2D traveltime tomography and inversion algorithms were presented by Firbas (1981) and White (1989) which used regular parameterizations and only first arrivals, and Spence et al. (1985) which allowed an irregular model grid and later arrivals. Lutter and Nowack (1990) and Lutter et al. (1990) developed a 2D inversion algorithm using a regular grid of nodes and numerical ray tracing that allows for the independent inversion of first arrivals and reflections for velocity and interface geometry, respectively. The Zelt and Smith (1992) 2D algorithm is the opposite in many respects. An irregular grid of velocity and interface nodes can be used, and any type of refracted or reflected arrival can be inverted simultaneously for velocity and interface geometry. In addition, unique to the Zelt and Smith (1992) algorithm, it allows any subset of the total set of model parameters to be inverted for, holding all others fixed, facilitating an automated forward modeling approach and making it straightforward to incorporate prior information. Clowes et al. (1995) used the Zelt and Smith (1992) algorithm to simultaneously invert refraction and reflection times from

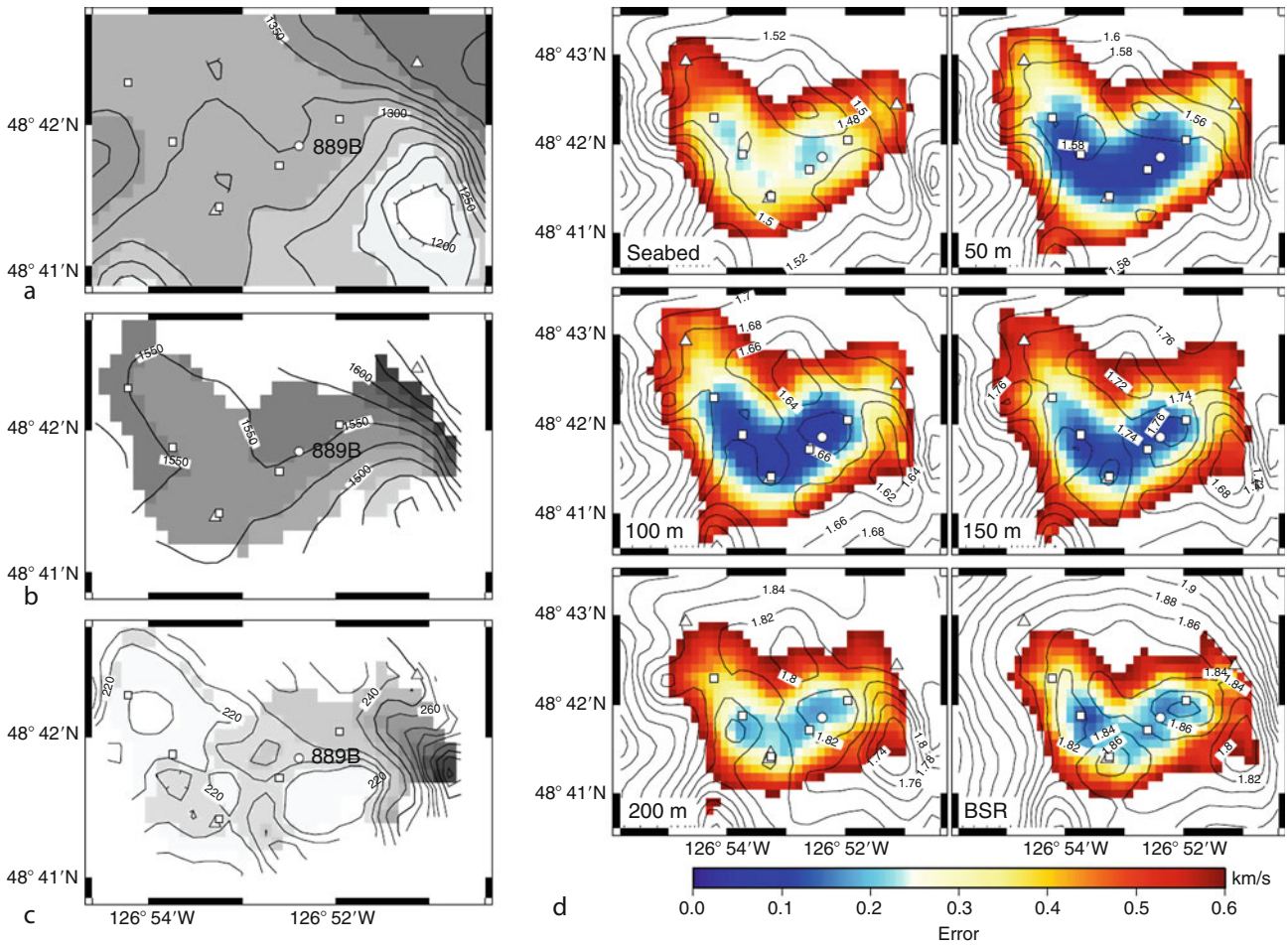


Traveltime Tomography Using Controlled-Source Seismic Data, Figure 4 Example of simultaneous 3D first-arrival and reflection tomography of sparse crustal data across the Tintina strike-slip fault in the northern Canadian Cordillera (Zelt et al., 2006b). Left panel shows the lower crustal velocity averaged vertically between 1 and 15 km above the Moho. Center panel shows depth to Moho; dots show reflection points. Right panel shows the upper mantle velocity averaged vertically between 1 and 15 km below the Moho. Stars are shot locations. The straight lines labeled 21 and 22 indicated the best-fit lines of the crooked profiles containing the receivers locations. Regions not sampled by rays are blank. Dashed line is the surface location of the fault. There is a 0.3–0.4 km/s change in upper mantle velocity across the fault.

a network of 2D profiles such that the model parameters at the intersection points were linked to ensure consistency. This approach can be used to infer 3D structure from 2D data, or develop a starting model for full 3D inversion if offline data is available in addition to inline data (e.g., Zelt et al., 1999).

Hole (1992) developed a 3D first-arrival tomography algorithm using the Vidale (1990) forward modeling scheme and back projection to solve for a smooth velocity field. Hammer et al. (1994) developed a 3D first-arrival tomography algorithm specially suited to sparse data using a spectral, continuous function model parameterization. Toomey et al. (1994) developed a 3D first-arrival tomography algorithm to obtain a smooth velocity field in which the velocity grid is "draped" from an irregular bathymetric surface by vertically shearing the columns of nodes to accurately model seafloor relief. Vertically sheared grids are particularly important for marine data, especially at mid-ocean ridges, where a large velocity discontinuity may exist. Other algorithms with sheared grids have been developed by Van Avendonk et al. (1998) and Korenaga et al. (2000), although they are limited to 2D models, but include reflections. The main difference between these two algorithms is the former uses first arrivals that can turn above or below a reflecting interface at which there is no velocity discontinuity, whereas the latter only uses first arrivals that turn above a reflecting interface.

Lailly and Sinoquet (1996) developed a 2D algorithm for inverting reflection times to estimate a smooth velocity

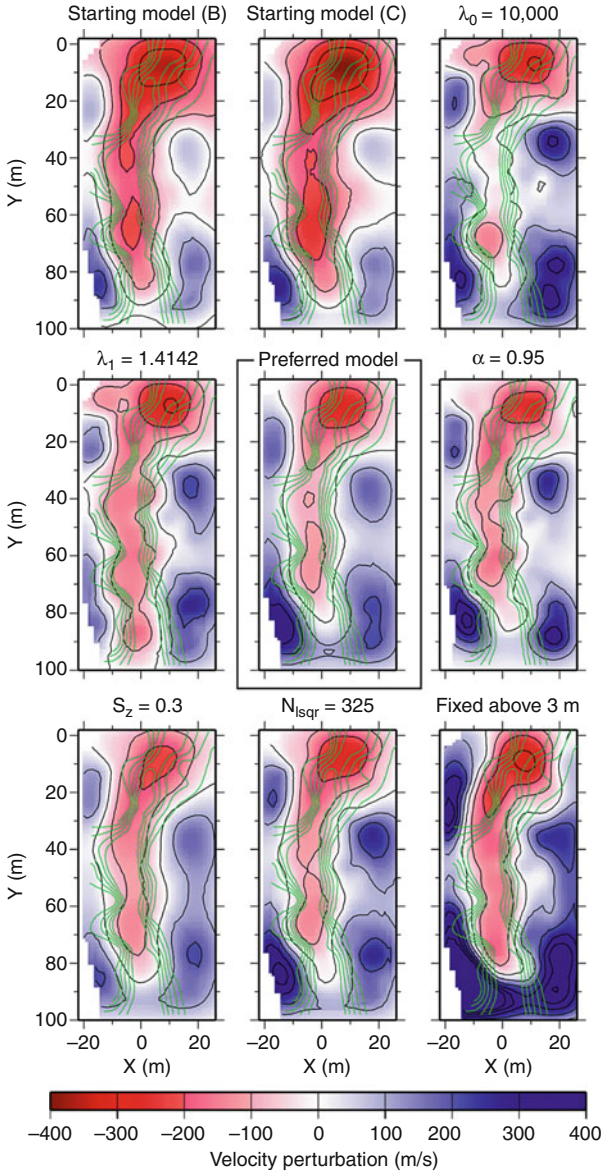


Traveltime Tomography Using Controlled-Source Seismic Data, Figure 5 Example of simultaneous 3D refraction and reflection tomography of OBS data and coincident single channel reflection profiles over a gas hydrate stability zone offshore Vancouver Island (Hobro et al., 2005). (**a**) Seafloor depth. (**b**) Bottom simulating reflector (BSR) depth. (**c**) BSR depth below the seafloor. (**d**) Slices through the velocity model at the seafloor, at 50-m intervals through the hydrate stability zone and immediately above the BSR. Labeled contours indicate velocities (km/s) and color scale marks formal uncertainty estimates. Ten OBS positions and ODP site 889 are marked. Refracted arrivals from above and below the BSR were inverted along with wide-angle and normal-incidence reflections from the BSR.

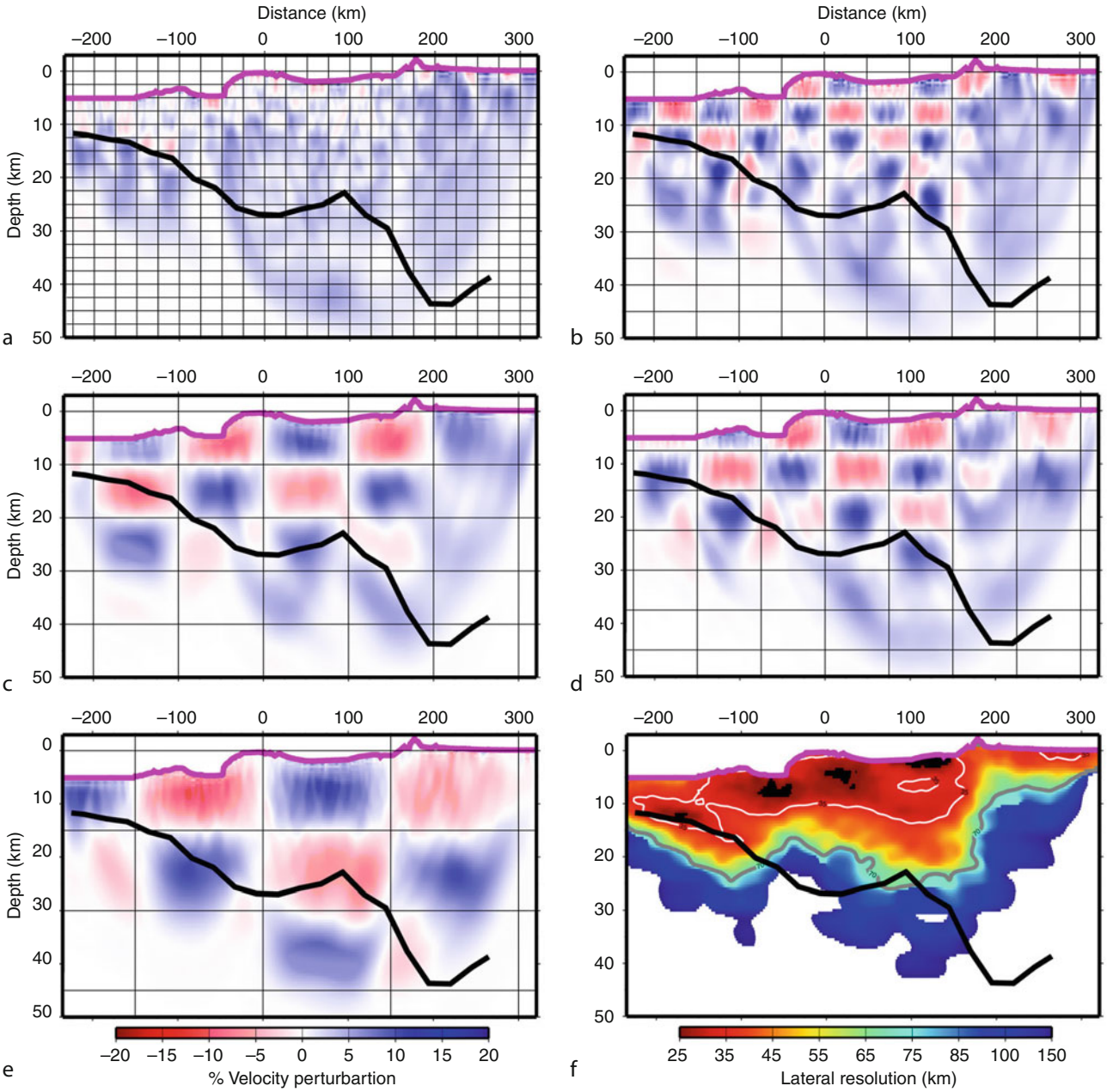
model for pre-stack depth migration of reflection data using the wavefront tracking method of Podvin and Lecomte (1991) for the forward calculation (Figure 2). Zelt and Barton (1998) developed a 3D first-arrival tomography algorithm using smoothing regularization and compared the results with those from the back-projection method of Hole (1992) (Figure 3). The results showed that for the same level of fit to the data, regularized inversion can provide a simpler model. Zelt et al. (2006b) applied 3D simultaneous refraction and reflection tomography to solve for smooth velocities and multiple interfaces using a sparse crustal-scale 3D dataset (Figure 4). McCaughey and Singh (1997) and Hobro et al. (2003) developed related 2D and 3D simultaneous refraction and reflection tomography algorithms that solve for smooth velocities and interfaces with the allowance for discontinuities across the layer boundaries (Figure 5).

Most of the algorithms described above are general purpose in nature and follow more or less from the objective function presented in the previous section. However, more specialized algorithms have been developed and a few are described here to give a sense of what is possible. Hole et al. (1992) developed an algorithm for determining a 3D interface using first arrivals with known velocities above and below the interface. Zhang et al. (1998) inverted traveltime curves instead of points, specifically average slowness and apparent slowness. They claim this balances the contribution from short and long rays, and enhances resolution and convergence speed. Rawlinson and Sambridge (2003b) developed an algorithm for the inversion of refraction and reflection times using a 3D multi-layered model parameterization. Interfaces are defined by a non-uniform node distribution and velocities vary linearly with depth so rays are calculated analytically as piece-wise circular arcs using a shooting method. Trinks et al. (2005) presented a method for simultaneous refraction and reflection inversion using a 2D layered model parameterization which adapts to non-uniform ray coverage such that the cell size is inversely proportional to the local ray density.

Refraction/wide-angle reflection traveltimes can be jointly inverted with coincident near-vertical reflection times in several ways (Zelt, 1999). Arrivals in pre-stack reflection data can be picked and inverted jointly with the wide-angle data, but it is often difficult to pick weak events in pre-stack data. The most common approach is to pick reflections from a stacked reflection section and invert these data simultaneously with the wide-angle data by modeling zero-offset reflections (e.g., McCaughey and Singh, 1997; Zelt et al., 2003). One potential pitfall when using coincident reflection data is incorrectly correlating the near-vertical reflection event with the corresponding wide-angle event or layer boundary; Jaiswal et al. (2006) present a way to avoid this problem. As an alternative to joint inversion, the reflector geometries and interval velocities from stacked data can be used as prior information in the regularization to constrain a layered velocity model (e.g., Bosch et al., 2005).

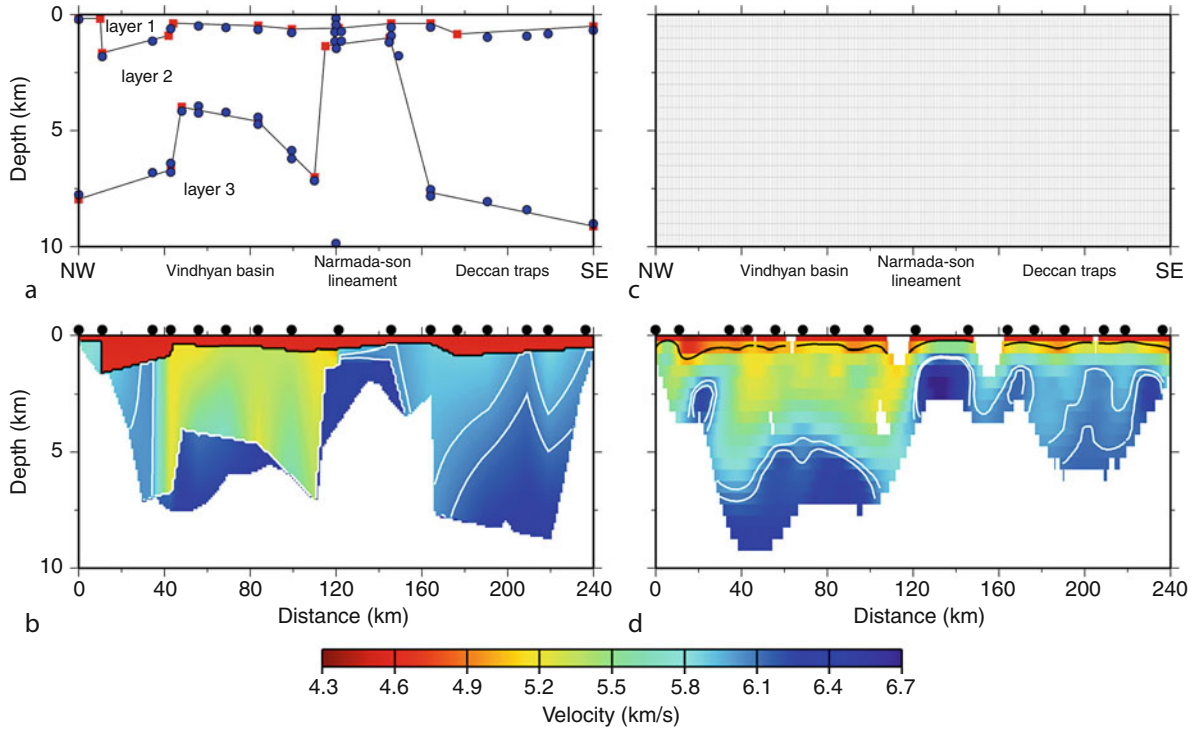


Traveltime Tomography Using Controlled-Source Seismic Data, Figure 6 Example of model assessment for near-surface 3D first-arrival tomography at a groundwater contamination site (Zelt et al., 2006a) using the Zelt and Barton (1998) regularized inversion algorithm seeking a minimum-structure model. Target of the surface was a low-velocity paleochannel cut into a clay layer running roughly north-south through the center of the survey area. The depth to the clay layer was known from extensive well data in the area; depth-to-clay contours (green) from 7 to 11 m overlay the depth slices at 10 m for this comparison of nine different models. Models displayed as perturbations relative to a 1D starting model; the starting velocity at this depth is 1150 ms/s. The preferred final model is in the center. The other models were obtained by trying two different starting models, different values for five free parameters that control the inversion, and by fixing the model above 3 m at the starting model values. Overall, the preferred model has minimum structure; the model in the lower left appears simpler in this depth slice, but it is rougher vertically.



Traveltime Tomography Using Controlled-Source Seismic Data, Figure 7 Example of 2D first-arrival checkerboard test for crustal data from an onshore-offshore experiment across the South American-Caribbean plate boundary at 67°W (Magnani et al., 2009). There were shots and receivers both onshore and offshore. (**a**)–(**e**) show recovered checkerboard anomalies for sizes from 25×2.5 km to 150×15 km. (**f**) Estimated lateral resolution of the velocity model using a 2D version of the method presented by Zelt (1998). Regions with better than 25 km resolution are black; regions with worse than 150 km resolution are white. Pink line indicates the land surface and bathymetry. The coast line is at ~160 km. Black line indicates the Moho from the preferred final model.

The close relationship between seismic velocity and density (Brocher, 2005) lends gravity data to a joint inversion with traveltime data. Nielsen and Jacobsen (2000) used the Zelt and Smith (1992) algorithm as the basis for a simultaneous inversion of refraction and reflection times with gravity data to derive a layered 2D crustal model. Korenaga et al. (2001) carried out a joint inversion of traveltime and gravity data in which error propagation from the velocity model to the predicted gravity anomalies was taken into account.

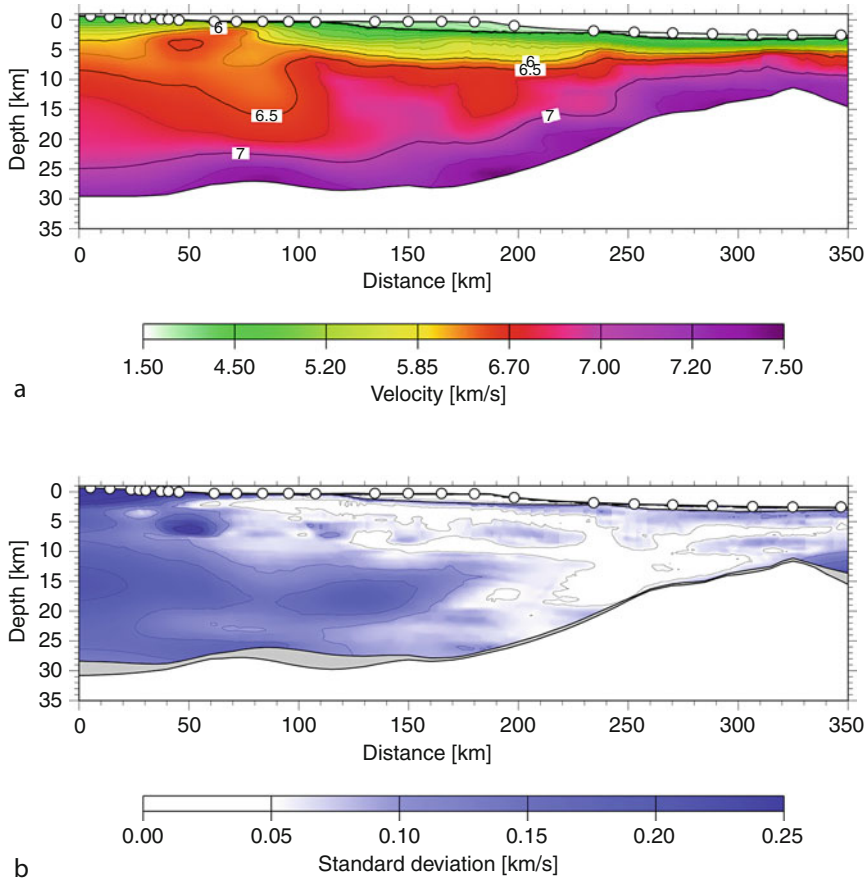


Traveltime Tomography Using Controlled-Source Seismic Data, Figure 8 Example from central India across the Narmada-Son lineament of applying two different traveltime inversion algorithms to the same refraction/wide-angle reflection data (Zelt et al., 2003) to compare the "geological" model with the minimum-structure model. (**a**) Layered model parameterization using the Zelt and Smith (1992) algorithm; there are 43 velocity nodes (blue dots) and 21 interface nodes (red squares). (**b**) Velocity model corresponding to parameterization in (**a**); both first arrivals and reflections from the top of the third layer were used. (**c**) Fine-grid model parameterization using the Zelt and Barton (1998) algorithm to seek a smooth minimum-structure model; there are 7018 slowness cells. (**d**) Velocity model corresponding to parameterization in (**c**); only first arrivals were used. Only those portions of the models sampled by rays are shown. Isovelocity contours of 5.0 km/s (black) and 6.0 and 6.1 km/s (white) indicated. Shot point locations (black dots) and geologic features labeled above the models. An interpretation of the main features of the geological model in (**b**) is supported by the minimum-structure model in (**d**) since it contains the same features.

Model assessment

Once a model to explain a particular set of data is developed, model assessment should be used to explore the non-uniqueness, resolution, and errors associated with the model. This is an attempt to quantify the robustness of the model, a process that is not straightforward for nonlinear inverse problems, and different assessment techniques will be appropriate depending on the data, model parameterization, geologic target, and objectives of the experiment. There are two classes of assessment methods: linear and nonlinear. The former methods assume it is valid to consider only a local neighborhood of the final model within which the model-data relationship is linear. The latter methods involve additional nonlinear inversions of the real data or synthetic data using the true source-receiver geometry of the experiment. Linear methods are quick and easy to apply, although potentially less reliable depending on the nonlinearity of the problem. Nonlinear methods are computational intensive, often more so than what was needed to derive the final model, and may require a specialized inversion algorithm. However, nonlinear methods are the only way to properly estimate the final model's robustness. Zelt (1999) and Rawlinson and Sambridge (2003a) describe many of the assessment techniques in use today, while Zelt et al. (2006a) apply several different techniques to the same 3D dataset (Figure 6).

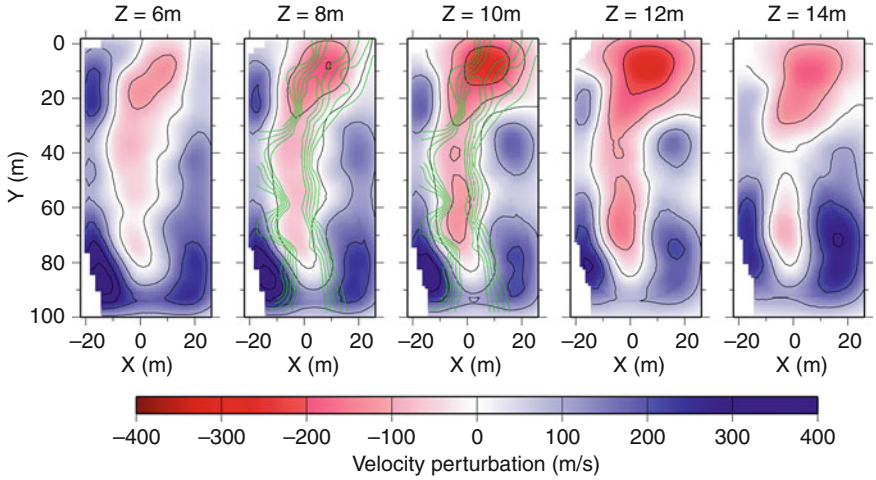
The simplest form of model assessment is to examine the ray coverage in the model using ray plots or "hit counts" (the number of rays sampling each model parameter). These can be misleading since a limited distribution of ray angles may provide less independent model constraint than expected from the number of rays alone. For example, rays in the uppermost mantle will tend to be sub-horizontal, and therefore capable of providing little lateral resolution. Toomey et al. (1994) used a measure called derivative weight sum (DWS) that equals the sum of the partial derivatives for each model parameter. This is somewhat more meaningful than hit count since it weights



Traveltime Tomography Using Controlled-Source Seismic Data, Figure 9 Example of model assessment using a nonlinear Monte Carlo procedure to estimate model variance (Korenaga et al., 2000). (**a**) Velocity model for the southeast Greenland continental margin obtained by joint traveltime inversion of first arrivals and wide-angle Moho reflections recorded at 18 OBS's and eight land stations (white circles). This model was obtained by averaging 100 Monte Carlo ensembles. (**b**) Corresponding standard deviation for velocity and depth nodes; the standard deviation of the Moho is indicated by the width of the grey zone at the base of the crust.

the contribution from each ray according to the parameter's sensitivity to that ray. However, it also fails to account for the degree of independence within the ray set. Other common linear assessment measures include an examination of the diagonals of the posterior resolution and covariance matrices (e.g., Zelt and Smith, 1992). These are more precise measures since they account for the independence of the ray set, but they are best used in a relative as opposed to absolute sense, since they do not account for the nonlinearity of the problem. Rows of the resolution matrix, known as resolution kernels, provide a spatial sense of the averaging of the true structure by the model (Zelt, 1999). Probably the most precise linear assessment method uses singular value decomposition (SVD) since it is able to quantify the nature and degree of model constraint through construction of an orthogonal set of basis vectors that span model space, each with a specified weight in the model reconstruction (e.g., White, 1989; Scales et al., 1990).

A nonlinear assessment of the spatial resolution or uncertainty of a single model parameter, or set of parameters, is possible in which the real data and synthetic data are inverted in the same way that the final model was derived, allowing the full nonlinearity of the problem to be accounted for, including trade-offs between model parameters (Zelt and White, 1995; Zelt, 1999; Christeson et al., 1999). For spatial resolution, the value of a model parameter is perturbed enough to yield a significant traveltime anomaly with respect to the pick uncertainties. Rays are traced through the perturbed model to calculate a set of perturbed traveltimes. The perturbed data are then inverted using the final model as the starting model. The spatial resolution about the selected parameter will be indicated by the amount that the values of adjacent parameters differ from their corresponding value in the original final model. If the model is poorly resolved about the selected parameter, then the parameter's perturbation will be smeared into adjacent parameters, perhaps both



Traveltime Tomography Using Controlled-Source Seismic Data, Figure 10 Example of near-surface 3D first-arrival tomography at a groundwater contamination site (Zelt et al., 2006a) using the Zelt and Barton (1998) regularized inversion seeking a minimum-structure model. The Horizontal depth slices from 6 to 14 m displayed as perturbations relative to a 1D starting model. Black contour interval is 100 m/s. Target of the surface was a low-velocity paleochannel cut into a clay layer running roughly north-south through the center of the survey area. The depth to the clay layer was known from extensive well data in the area; depth-to-clay contours (green) from 7 to 11 m overlay the 8 and 10 m depth slices. This experiment consisted of ~600 shots recorded by ~600 receivers yielding ~360,000 traces, providing 187,877 useable picks for the inversion.

velocities and interface depths if both parameter types are involved, and the extent of the smearing indicates the spatial resolution. Both positive and negative parameter perturbations should be tested, and it will likely be sufficient to examine only one or two representative velocity and interface parameters for each layer or region of the model.

A parameter-selective algorithm such as the Zelt and Smith (1992) approach allows for a nonlinear estimate of a single parameter's absolute uncertainty (Zelt, 1999). The value of the model parameter is slightly perturbed from its value in the final model and held fixed while inverting the observed data involving all other model parameters. The size of the perturbation is increased until the recovered model is unable to fit the observed data as well as the preferred final model based on an F test comparing the traveltime fits of the preferred and perturbed final models. The maximum parameter perturbation that allows a comparable fit to the observed data is an estimate of its absolute uncertainty. Again, both positive and negative parameter perturbations should be tested, and it will likely be sufficient to examine only one or two representative velocity and interface nodes for each layer or region of the model. A fine grid tomographic approach can also be used to perform tests like these. For example, Zelt and Barton (1998) examined one region of a model where there was significant lateral structure even though the constraint from the ray coverage was known to be low in this area. They added regularization to the inversion to force that region to remain laterally homogeneous and thereby establish the required trade-offs elsewhere in the model.

One of the most common forms of nonlinear model assessment to estimate spatial model resolution is the checkerboard test (e.g., Magnani et al., 2009; Figure 7). In these tests an alternating pattern of high and low anomalies is superimposed on the starting model from the inversion of the real data. Synthetic data are calculated for the "checkerboard" model, and then inverted using the same starting model and source-receiver geometry as the real data. The recovered model will closely resemble the checkerboard pattern in regions of good constraint, but will otherwise not resemble the checkerboard model. The resolution at different length scales can be estimated by testing anomaly patterns of different sizes. Anomaly patterns with different polarity, registration, and orientation should be tested to average out the effects of changing ray coverage due to the nonlinearity of the problem (Zelt, 1998).

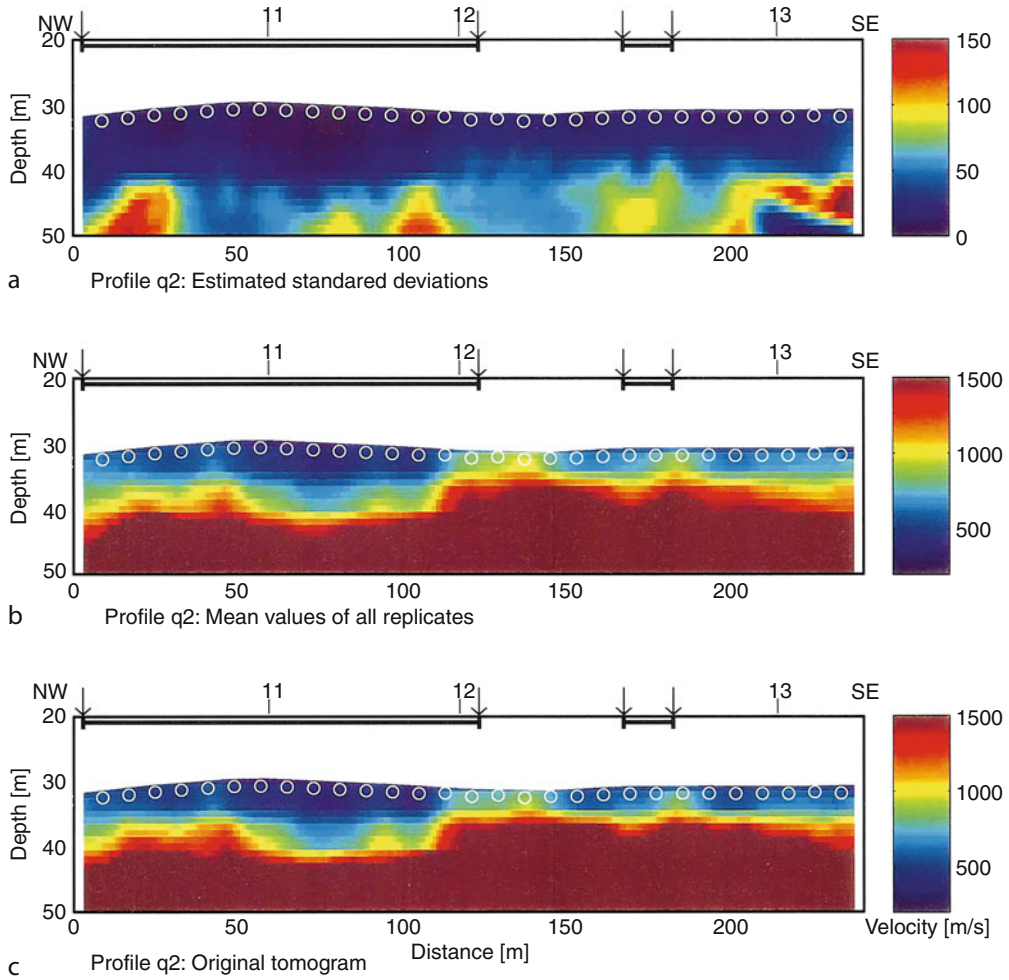
Other nonlinear assessment techniques include trying different starting models, different values of the free parameters in the objective function, and exclusion of subsets of the data considered less reliable (e.g., Zelt et al., 2006a; Figure 6). Using different model parameterizations and different inversion algorithms can be effective, especially when one inversion algorithm seeks a model that satisfies all notions of what is geological reasonable, and one algorithm seeks the minimum-structure model (Figure 8; Zelt et al., 2003). In this way, it is possible to determine what model structure is consistent with the data to facilitate hypothesis testing, while at the same time establishing what model structure is required by the data.

Perhaps the most complete and computationally intensive form of model assessment involves a nonlinear Monte Carlo procedure. Many inversions are performed in which randomized data and/or randomized starting models are used. The resulting models are used to compute posterior model covariance and resolution estimates. Zhang et al. (1998) and Korenaga et al. (2000) applied this approach to 2D inversions of first arrivals and reflections (Figure 9).

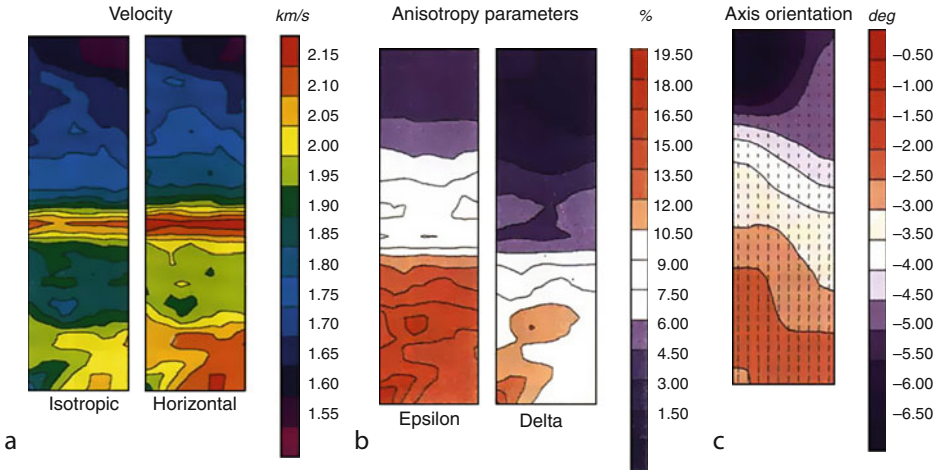
More examples

The examples in Figures 1–9 demonstrate the types of models that can result from a wide range of controlled-source experiments. The remaining examples presented in this section are intended to round out the range of applications to which traveltime tomography and inversion are applied in controlled-source studies, and the types of model structure that can result. The main purpose of the examples is to show what is possible in terms of tomographic imaging of different kinds of geologic features in the near-surface, upper crust, lower crust, and uppermost mantle using traveltime data. All examples shown in this article use only P-wave traveltime data.

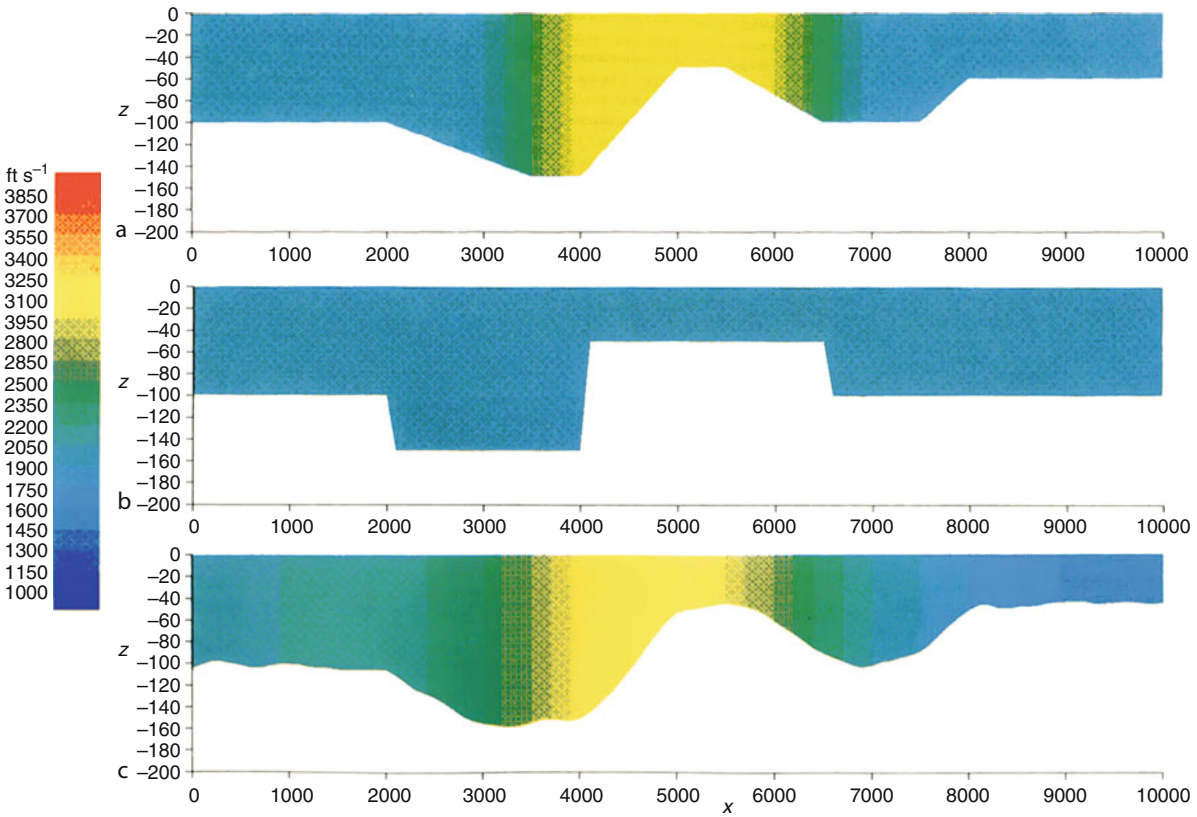
Zelt et al. (2006a) applied the regularized 3D first-arrival tomography algorithm of Zelt and Barton (1998) to a dense 3D dataset from a groundwater contamination site (Figure 10). Lanz et al. (1998) applied 2D first-arrival tomography and a Monte Carlo scheme to estimate a smooth model with standard deviations for a buried waste disposal site (Figure 11). Pratt and Chapman (1992) applied 2D first-arrival tomography to crosswell data to simultaneously determine velocity and anisotropy



Traveltime Tomography Using Controlled-Source Seismic Data, Figure 11 Example of near-surface first-arrival tomography over a buried waste disposal site seeking a smooth model (Lanz et al., 1998). (**a**) Estimated standard deviation of velocities computed using 200 randomized datasets. (**b**) Mean velocity model from 200 randomized datasets. (**c**) Preferred final model from inversion of original picked data. Open circles are shot locations. Arrows indicate landfill borders defined by several other geophysical techniques.



Traveltime Tomography Using Controlled-Source Seismic Data, Figure 12 Example of crosswell tomography in which transverse isotropy is assumed and regularization was used to seek a smooth model (Pratt and Chapman, 1992). The velocity and anisotropic parameters were solved for simultaneously. (**a**) Comparison of isotropic velocity model and horizontal velocity model from anisotropic inversion. The model is 20 m across and 60 m deep. (**b**) Anisotropy parameters. The amount of anisotropy generally increases with depth, reaching nearly 20%. (**c**) Symmetry axis orientation. The length of the markers is proportional to one of the anisotropy parameters (epsilon).

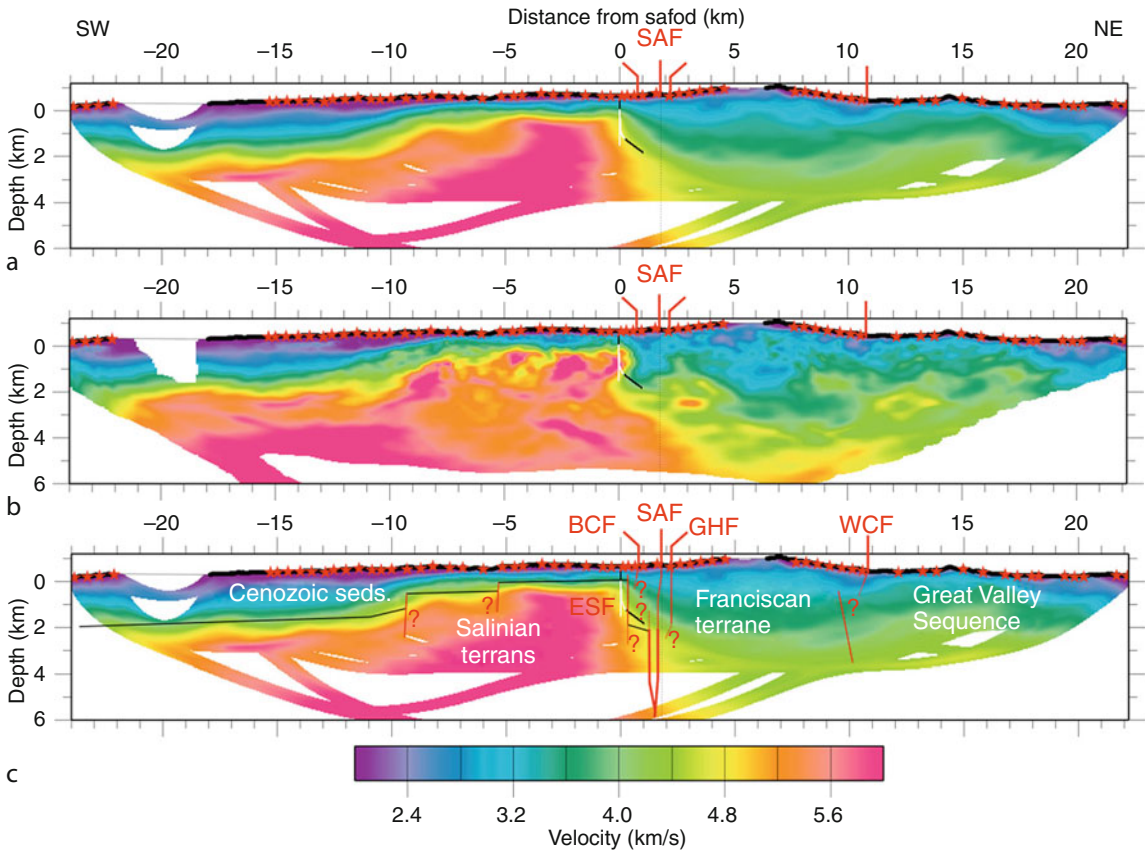


Traveltime Tomography Using Controlled-Source Seismic Data, Figure 13 Example of imaging the near-surface weathering layer (i.e. the refraction statics problem) using smoothing regularization (Scales et al., 1990). Shots were placed at either end of the model and there were 100 receivers between them. (**a**) True model. Refractor velocity is 5000 ft/s. (**b**) Starting model with blocky thickness variations in the weathering layer. Refractor velocity is 4000 ft/s. (**c**) Final model. Refractor velocity is 5001 ft/s. The inversion solved simultaneous for a laterally-varying weathering-layer velocity and refractor, and a constant refractor velocity.

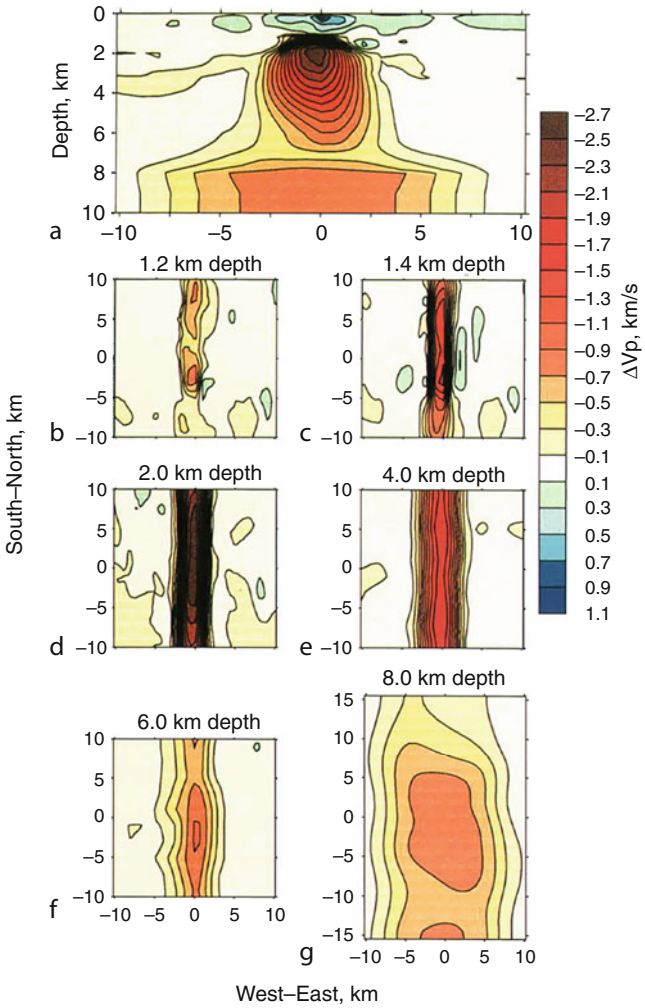
parameters (Figure 12). Scales et al. (1990) imaged a near-surface weathering layer using 2D first-arrival times, solving simultaneously for lateral variations in the velocity and thickness of the layer (Figure 13). Hole et al. (2006) applied 2D first-arrival tomography to data recorded across the San Andreas Fault using two different algorithms (Figure 14). Dunn et al. (2000) applied 3D refraction and reflection tomography incorporating anisotropy to data from the East Pacific Rise (Figure 15). Van Avendonk et al. (1998) inverted refraction and reflection times for a smooth 2D velocity model and an assumed flat Moho across the Clipperton transform fault (Figure 16). Clark et al. (2008) inverted refractions and reflections from an onshore–offshore survey across the South American-Caribbean plate boundary using a smooth tomography approach for the upper crust, and a layer-based inversion for the lower crust, Moho, and upper mantle (Figure 17). Zelt et al. (2003) compared layer-based simultaneous refraction and reflection inversion with independent smooth refraction and reflection tomography for three different 2D crustal datasets (Figure 18).

Future

Full 3D seismic experiments are likely to become more commonplace, necessitating the need for some of the popular 2D tomography and inversion algorithms to be extended to 3D. It is likely that three-component, S-wave, and converted-wave studies will become more widespread, especially for near-surface environmental and engineering studies, necessitating the need for algorithms that can perform coupled P-wave, S-wave, or Poisson's ratio inversions. The desire to incorporate or determine anisotropy will likely become more commonplace in the



Traveltime Tomography Using Controlled-Source Seismic Data, Figure 14 Example of first-arrival tomography across the San Andreas Fault comparing models from two different algorithms (Hole et al., 2006). There were 63 shots (red stars) and 912 receivers (Thick black line). (**a**) Velocity model from application of the Hole (1992) back-projection method. (**b**) Velocity model from application of Zelt and Barton (1998) smoothing-regularized method. Both models provide roughly the same level of fit to the picks. Regions without ray coverage are white. Well-log observations at SAFOD (San Andreas Fault Observatory at Depth) are white for granite and black for sedimentary rock. Note how high velocities (>5 km/s) better match the known position of granite in the well for the model in (**b**) compared to the model (**a**), suggesting that some of the additional structure in the model in (**b**) may be true. (**c**) Geology and fault interpretation overlaid on the model in (**a**). Black lines are top of basement; red lines are faults.

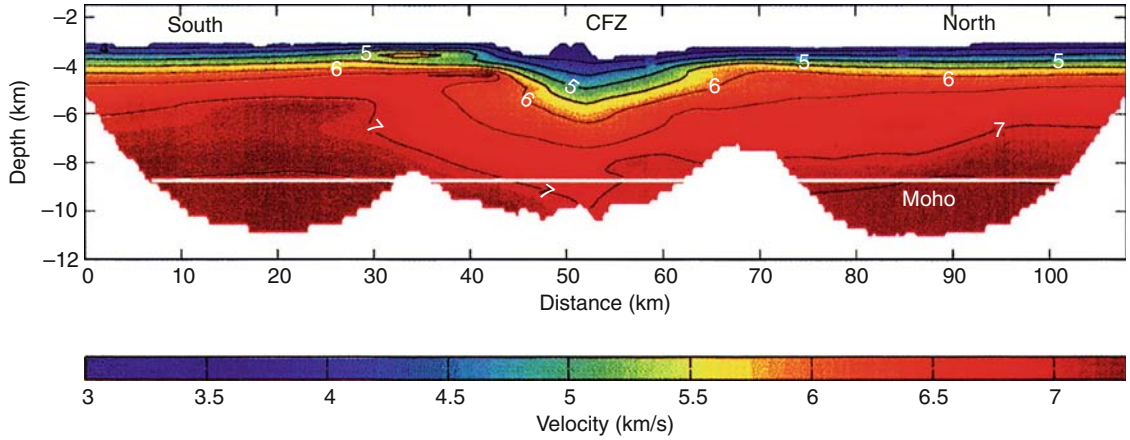


Traveltime Tomography Using Controlled-Source Seismic Data, Figure 15 Example of 3D refraction and reflection tomography across the East pacific Rise at 9°30′N for crust and uppermost mantle structure (Dunn et al., 2000). There were 15 OBS's and 480 explosive sources. Refracted arrivals from the crust and upper mantle and reflections from the Moho were simultaneously inverted. Anisotropy of 4% in the crust and 7% in the upper mantle was included in the modeling. Final velocity model as perturbations relative to a 1D model are shown and are contoured at 0.2 km/s intervals. (**a**) vertical section through the center of the model. (**b**–**g**) Depth slices from 1.2–8.0 km beneath the seafloor. The low-velocity body is interpreted to represent melt distribution within a magmatic system.

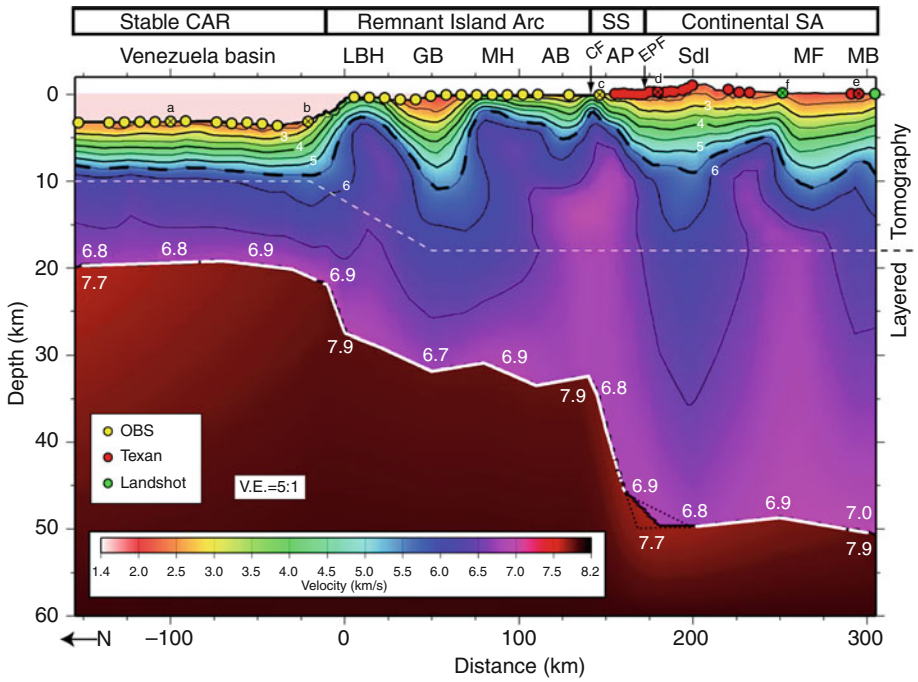
future, and there are currently only a few traveltime algorithms that include any form of anisotropy (e.g., Dunn et al., 2000).

For all scales of seismic experiments, the need for flexible, general-purpose traveltime tomography and inversion algorithms will remain, both as the primary modeling tool and to provide long-wavelength starting models for higher-resolution full waveform inversion techniques (e.g., Brenders and Pratt, 2007). The Zelt and Smith (1992) traveltime inversion algorithm has been widely used for many years for several reasons: (1) an irregular, layered model parameterization adaptable to the data coverage and geologic structures, (2) parameter-selective inversion, (3) any type of body-wave data can be modeled, and (4) it is easy to incorporate prior information. As such it stands apart from most other popular algorithms that use a uniform, fine grid parameterization, and consider only first arrivals, and in some cases reflections. However, the Zelt and Smith (1992) algorithm was primarily intended for relatively sparse 2D data. There will likely be a niche for a new version of the Zelt and Smith algorithm if it maintains all of its current advantages, but is extended to 3D and includes regularization to handle dense data, namely second-order derivatives to seek a smooth, minimum-structure model.

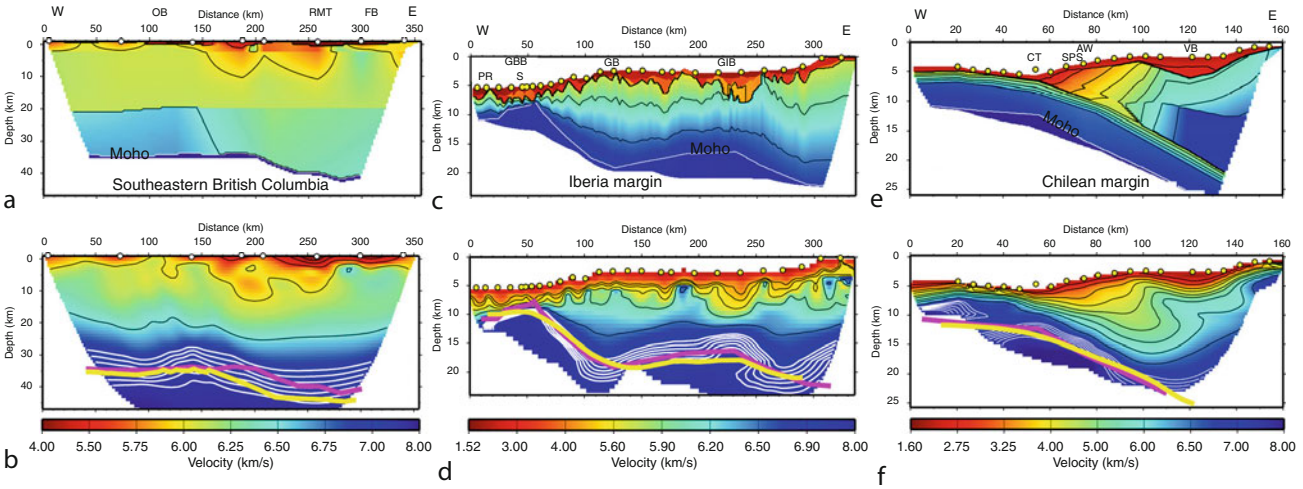
About 10 years ago, a theory for traveltime tomography was developed in global seismology to take into account the finite-frequency nature of seismic data (Dahlen et al., 2000). This method is known as finite-frequency traveltime tomography and it yields what are popularly known as banana-doughnut sensitivity kernels. By taking finite-frequency effects into account, this form of traveltime tomography should, theoretically, yield a more accurate estimation of velocity anomalies in terms of their magnitude and spatial resolution because the physics of wave propagation is treated more accurately. The theory developed for global seismology uses as input data delay times relative to a reference Earth model, and as such the theory is generally not applicable to controlled-source data. This is because for controlled-source data there is no requisite reference velocity model known in advance that is capable of yielding realistic synthetic waveforms that are close enough to the recorded seismograms to yield a meaningful delay time through cross correlation of the waveforms because the crust and near-surface are much more laterally heterogeneous than the mantle. As a result, a nonlinear inverse method must be used, requiring a starting model, an iterative method and recalculated travel paths at each iteration. In addition, the method must be capable of calculating a frequency-dependent traveltime along the total path, as opposed to a delay time. Zelt (2009) presented a finite-frequency traveltime tomography method specifically designed for controlled-source data, that is, a nonlinear inversion of total traveltimes, with frequency taken into account for both the forward calculation of traveltimes, and in the inverse step by calculating the appropriate frequency-dependent sensitivity kernels. The results show it is possible to achieve more accurate velocity estimation than in the equivalent infinite-frequency-derived models. And in addition, applications without any regularization are possible, and therefore finite-frequency traveltime tomography has the potential to allow the data alone to determine the model structure in a robust manner. These characteristics of finite-frequency traveltime tomography



Traveltime Tomography Using Controlled-Source Seismic Data, Figure 16 Example of 2D tomography across the Clipperton transform fault (Van Avendonk et al., 1998). There were only five OBS's. Crust and upper mantle refracted arrivals and Moho reflections were inverted simultaneously for a smooth velocity model. A flat Moho at 8.6 km depth was assumed. The model shows reduced velocities in the upper crust as far as 10 km south and 18 km north of the transform fault (CFZ).



Traveltime Tomography Using Controlled-Source Seismic Data, Figure 17 Example of crustal velocity model from onshore-offshore survey across the South American-Caribbean plate boundary at 64°W (Clark et al., 2008). There were airguns offshore, two shots on land (green dots), OBS's offshore (yellow dots), and seismometers on land (red dots). The model above the dashed white line was obtained by first-arrival tomography using the Zelt and Barton (1998) smoothing-regularized tomography algorithm; the model below the dashed line was derived using the Zelt and Smith (1992) layer-based inversion using upper mantle refractions and Moho reflections to determine a sharp velocity contrast for the Moho. The strike-slip system near the coastline at ~150 km accommodates relative motion between the two plates and from this model is interpreted to extend near vertically through the entire crust and offset the Moho.



Traveltime Tomography Using Controlled-Source Seismic Data, Figure 18 A comparison of three sets of 2D crustal velocity models obtained using two different inversion algorithms (Zelt et al., 2003). The study areas are southeastern British Columbia, the Iberia margin and the Chilean margin. The top row of "geologic" models were obtained using the Zelt and Smith (1992) layer-based algorithm simultaneously inverting crust and upper mantle refractions and Moho reflections. The bottom row of "minimum-structure" models were obtained using the Zelt and Barton (1998) smoothing-regularized first-arrival tomography algorithm. Open circles are shot locations; yellow circles are OBS's. Black contours in the crust have an interval of 0.5 km/s; white contours representing the Moho in the smooth tomographic models have an interval of 0.1 km/s from 7.0–7.6 km/s. Pink lines are the Mohos from the layered models overlain on the tomographic models for comparison. Yellow lines are the Mohos from independent reflection tomography in which the tomographic velocity models were held fixed. Interpretations of the main features of the geological models are supported by the corresponding minimum-structure models since they contain the same features.

suggest that it will play an important role in the future. This is especially true for near-surface studies where infinite-frequency (ray) theory is least likely to be valid because the length scales of heterogeneities are often comparable or smaller than the seismic wavelengths.

Bibliography

Bevington, P. R., 1969. *Data Reduction and Error Analysis for the Physical Sciences*. New York: McGraw-Hill.

Bosch, M., Barton, P., Singh, S. C., and Trinks, I., 2005. Inversion of traveltime data under a statistical model for seismic velocities and layer interfaces. *Geophysics*, **70**, R33–R43.

Brenders, A. J., and Pratt, R. G., 2007. Full waveform-tomography for lithospheric imaging: results from a blind test in a realistic crustal model. *Geophysical Journal International*, **168**, 133–151.

Brocher, T. M., 2005. Empirical relations between elastic wavespeeds and density in the Earth's crust. *Bulletin. Seismological Society of America*, **95**, 2081–2092.

Cerveny, V., Molotkov, I., and Psencik, I., 1977. *Ray Method in Seismology*. Prague: University of Karlova.

Chambers, G., Levander, A., Zelt, C. A., and Dugan, B., 2009. Seismic waveform tomography with multicomponent data at a groundwater contamination site. *SEG Expanded Abstracts*, **28**, 3994.

Chapman, C. H., and Drummond, R., 1982. Body-wave seismograms in inhomogeneous-media using Maslov asymptotic theory. *Bulletin. Seismological Society of America*, **72**, S277–S317.

Christeson, G. L., McIntosh, K. D., Shipley, T. H., Flueh, E. R., and Goedde, H., 1999. Structure of the Costa Rica convergent margin, offshore Nicoya Peninsula. *Journal of Geophysical Research*, **104**, 25,443–25,468.

Clark, S. A., Zelt, C. A., Magnani, M. B., and Levander, A., 2008. Characterizing the Caribbean-South American plate boundary at 64 W using wide-angle seismic data. *Journal of Geophysical Research*, **113**, B07401, doi:10.1029/2007JB005329.

Clowes, R. M., Zelt, C. A., Amor, J. R., and Ellis, R. M., 1995. Lithospheric structure in the southern Canadian Cordillera from a network of seismic refraction lines. *Canadian Journal of Earth Sciences*, **32**, 1485–1513.

Constable, S. C., Parker, R. L., and Constable, C. G., 1987. Occam's inversion: a practical algorithm for generating smooth models from electromagnetic sounding data. *Geophysics*, **52**, 289–300.

Dahlen, F. A., Hung, S. H., and Nolet, G., 2000. Frechet kernels for finite-frequency traveltimes – I. Theory. *Geophysical Journal International*, **141**, 157–174.

Dunn, R. A., Toomey, D. R., and Solomon, S. C., 2000. Three-dimensional seismic structure and physical properties of the crust and shallow mantle beneath the East Pacific Rise at 9°30'N. *Journal of Geophysical Research*, **105**, 23,537–23,555.

Dziewonski, A. M., and Anderson, D. L., 1984. Seismic tomography of the earths interior. *American Scientist*, **72**, 483–494.

Firbas, P., 1981. Inversion of travel-time data for laterally heterogeneous velocity structure – linearization approach. *Geophysical Journal of the Royal Astronomical Society*, **67**, 189–198.

Hammer, P. T. C., Dorman, L. M., Hildebrand, J. A., and Cornuelle, B. D., 1994. Jasper Seamount structure: seafloor seismic refraction tomography. *Journal of Geophysical Research*, **99**, 6731–6752.

Hobro, J. W. D., Singh, S. C., and Minshull, T. A., 2003. Three-dimensional tomographic inversion of combined reflection and

refraction seismic traveltime data. *Geophysical Journal International*, **152**, 79–93.

Hobro, J. W. D., Minshull, T. A., Singh, S. C., and Chand, S., 2005. A three-dimensional seismic tomographic study of the gas hydrate stability zone, offshore Vancouver Island. *Journal of Geophysical Research*, **110**, B09102, doi:10.1029/2004JB003477.

Hole, J. A., 1992. Nonlinear high-resolution three-dimensional seismic travel time tomography. *Journal of Geophysical Research*, **97**, 6553–6562.

Hole, J. A., and Zelt, B. C., 1995. Three-dimensional finite-difference reflection travel times. *Geophysical Journal International*, **121**, 427–434.

Hole, J. A., Clowes, R. M., and Ellis, R. M., 1992. Interface inversion using broadside seismic refraction data and 3-dimensional travel time calculations. *Journal of Geophysical Research*, **97**, 3417–3429.

Hole, J. A., Ryberg, T., Fuis, G. S., Bleibinhaus, F., and Sharma, A. K., 2006. Structure of the San Andreas fault zone at SAFOD from a seismic refraction survey. *Geophysical Research Letters*, **33**, L07312, doi:10.1029/2005GL025194.

Jaiswal, P., Zelt, C. A., and Pecher, I. A., 2006. Seismic characterization of a gas hydrate system in the Gulf of Mexico using wide-aperture data. *Geophysical Journal International*, **165**, 108–120.

Korenaga, J., Holbrook, W. S., Kent, G. M., Kelemen, P. B., Detrick, R. S., Larsen, H. C., Hopper, J. R., and Dahl-Jensen, T., 2000. Crustal structure of the southeast Greenland margin from joint refraction and reflection seismic tomography. *Journal of Geophysical Research*, **105**, 21,591–21,614.

Korenaga, J., Holbrook, W. S., Detrick, R. S., and Kelemen, P. B., 2001. Gravity anomalies and crustal structure at the southeast Greenland margin. *Journal of Geophysical Research*, **106**, 8853–8870.

Lailly, P., and Sinoquet, D., 1996. Smooth velocity models in reflection tomography for imaging complex geological structures. *Geophysical Journal International*, **124**, 349–362.

Lanz, E., Maurer, H., and Green, A. G., 1998. Refraction tomography over a buried waste disposal site. *Geophysics*, **63**, 1414–1433.

Levander, A., Zelt, C. A., and Symes, W. W., 2007. Active source studies of crust and lithospheric structure. In Romanowicz, B., and Dziewonski, A., (eds.), *Treatise on Geophysics, Seismology and Structure of the Earth*. Amsterdam, Boston: Elsevier, Vol. 1, pp. 247–288.

Lutter, W. J., and Nowack, R. L., 1990. Inversion for crustal structure using reflections from the PASSCAL Ouachita experiment. *Journal of Geophysical Research*, **95**, 4633–4646.

Lutter, W. J., Nowack, R. L., and Braile, L. W., 1990. Seismic imaging of upper crustal structure using travel times from the PASSCAL Ouachita experiment. *Journal of Geophysical Research*, **95**, 4621–4631.

Magnani, M. B., Zelt, C. A., Levander, A., and Schmitz, M., 2009. Crustal structure of the South American-Caribbean plate boundary at 67°W from controlled source seismic data. *Journal of Geophysical Research*, **114**, B02312, doi:10.1029/2008JB005817.

McCaughey, M., and Singh, S. C., 1997. Simultaneous velocity and interface tomography of normal-incidence and wide-aperture traveltime data. *Geophysical Journal International*, **131**, 87–99.

McMechan, G. A., and Mooney, W. D., 1980. Asymptotic ray theory and synthetic seismograms for laterally varying structures: theory and application to the Imperial Valley, California. *Bulletin. Seismological Society of America*, **70**, 2021–2035.

Menke, W., 1989. *Geophysical Data Analysis: Discrete Inverse Theory.* San Diego: Academic.

Moser, T. J., 1991. Shortest path calculation of seismic rays. *Geophysics*, **56**, 59–67.

Nakanishi, I., and Yamaguchi, K., 1986. A numerical experiment on nonlinear image reconstruction from first-arrival times for two-dimensional island arc structure. *Journal of Physics of the Earth*, **34**, 195–201.

Nielsen, L., and Jacobsen, B. H., 2000. Integrated gravity and wide-angle seismic inversion for 2-D crustal modelling. *Geophysical Journal International*, **140**, 222–232.

Nowack, R. L., and Braile, L. W., 1993. Refraction and wide-angle reflection tomography: theory and results. In Iyer, H. M., and Hirahara, K. (eds.), *Seismic Tomography: Theory and Practice*. London: Chapman and Hall, pp. 733–765.

Paige, C. C., and Saunders, M. A., 1982. LSQR: an algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software*, **8**, 43–71.

Podvin, P., and Lecomte, I., 1991. Finite difference computation of traveltimes in very contrasted velocity models: a massively parallel approach and its associated tools. *Geophysical Journal International*, **105**, 271–284.

Pratt, R. G., and Chapman, C. H., 1992. Traveltime tomography in anisotropic media–II. Application. *Geophysical Journal International*, **109**, 20–37.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P., 1992. *Numerical Recipes in Fortran: The Art of Scientific Computing*, 2nd edn. Cambridge: Cambridge University Press.

Ramachandran, K., Dosso, S. E., Spence, G. D., Hyndman, R. D., and Brocher, T. M., 2005. Forearc structure beneath southwestern British Columbia: a three-dimensional tomographic velocity model. *Journal of Geophysical Research*, **110**, B02303, doi:10.1029/2004JB003258.

Rawlinson, N., and Sambridge, M., 2003a. Seismic traveltime tomography of the crust and lithosphere. *Advances in Geophysics*, **46**, 181–198.

Rawlinson, N., and Sambridge, M., 2003b. Irregular interface parametrization in 3-D wide-angle seismic traveltime tomography. *Geophysical Journal International*, **155**, 79–92.

Rawlinson, N., Houseman, G. A., and Sambridge, M., 2001. Inversion of seismic refraction and wide-angle reflection traveltimes for 3-D layered crustal structure. *Geophysical Journal International*, **145**, 381–401.

Scales, J. A., Docherty, P., and Gersztenkorn, A., 1990. Regularisation of nonlinear inverse problems: imaging the near-surface weathering layer. *Inverse Problems*, **6**, 115–131.

Schmelzbach, C., Zelt, C. A., Juhlin, C., and Carbonell, R., 2008. P- and Sv-velocity structure of the South Portuguese Zone fold-and-thrust belt, SW Iberia, from traveltime tomography. *Geophysical Journal International*, **175**, 689–712.

Sethian, J. A., and Popovici, A. M., 1999. 3-D traveltime computation using the fast marching method. *Geophysics*, **64**, 516–523.

Spence, G. D., Whittall, K. P., and Clowes, R. M., 1984. Practical synthetic seismograms for laterally varying media calculated by asymptotic ray theory. *Bulletin. Seismological Society of America*, **74**, 1209–1223.

Spence, G. D., Clowes, R. M., and Ellis, R. M., 1985. Seismic structure across the active subduction zone of western Canada. *Journal of Geophysical Research*, **90**, 6754–6772.

Toomey, D. R., Solomon, S. C., and Purdy, G. M., 1994. Tomographic imaging of the shallow crustal structure of the East Pacific Rise at 9°30'N. *Journal of Geophysical Research*, **99**, 24,135–24,157.

Trinks, I., Singh, S. C., Chapman, C. H., Barton, P. J., Bosch, M., and Cherrett, A., 2005. Adaptive traveltime tomography of densely sampled seismic data. *Geophysical Journal International*, **160**, 925–938.

Um, J., and Thurber, C., 1987. A fast algorithm for two-point seismic ray tracing. *Bulletin. Seismological Society of America*, **77**, 972–986.

Van Avendonk, H. J. A., Harding, A. J., Orcutt, J. A., and McClain, J. S., 1998. A two-dimensional tomographic study of the Clipperton transform, fault. *Journal of Geophysical Research*, **103**, 17,885–17,899.
Van Avendonk, H. J. A., Harding, A. J., Orcutt, J. A., and Holbrook, W. S., 2001. Hybrid shortest path and ray bending method for traveltime and raypath calculations. *Geophysics*, **66**, 648–653.
Van Avendonk, H. J. A., Shillington, D. J., Holbrook, W. S., and Hornbach, M. J., 2004. Inferring crustal structure in the Aleutian arc from a sparse wide-angle seismic data set. *Geochemistry, Geophysics, Geosystems*, **5**, Q08008, doi:10.1029/2003GC000664.
Vidale, J. E., 1988. Finite-difference calculation of traveltimes. *Bulletin. Seismological Society of America*, **78**, 2062–2076.
Vidale, J. E., 1990. Finite-difference calculation of traveltimes in three dimensions. *Geophysics*, **55**, 521–526.
White, D. J., 1989. Two-dimensional seismic refraction tomography. *Geophysical Journal International*, **97**, 223–245.
Zelt, C. A., 1998. Lateral velocity resolution from three-dimensional seismic refraction data. *Geophysical Journal International*, **135**, 1101–1112.
Zelt, C. A., 1999. Modeling strategies and model assessment for wide-angle seismic traveltime data. *Geophysical Journal International*, **139**, 183–204.
Zelt, C.A., 2009. Frequency-dependent traveltime tomography for controlled-source, near-surface seismic data. EOS Trans. AGU, 90(52), Fall Meet. Suppl., Abstract NS31A–1155.
Zelt, C. A., and Barton, P. J., 1998. Three-dimensional seismic refraction tomography: a comparison of two methods applied to data from the Faeroe Basin. *Journal of Geophysical Research*, **103**, 7187–7210.
Zelt, C. A., and Ellis, R. M., 1988. Practical and efficient ray tracing in two-dimensional media for rapid traveltime and amplitude forward modeling. *Canadian Journal of Exploration Geophysics*, **24**, 16–31.
Zelt, C. A., and Smith, R. B., 1992. Seismic traveltime inversion for 2-D crustal velocity structure. *Geophysical Journal International*, **108**, 16–34.
Zelt, C. A., and White, D. J., 1995. Crustal structure and tectonics of the southeastern Canadian Cordillera. *Journal of Geophysical Research*, **100**, 24,255–24,273.
Zelt, C. A., Hojka, A. M., Flueh, E. R., and McIntosh, K. D., 1999. 3D simultaneous seismic refraction and reflection tomography of wide-angle data from the central Chilean margin. *Geophysical Research Letters*, **26**, 2577–2580.
Zelt, C. A., Sain, K., Naumenko, J. V., and Sawyer, D. S., 2003. Assessment of crustal velocity models using seismic refraction and reflection tomography. *Geophysical Journal International*, **153**, 609–626.
Zelt, C. A., Azaria, A., and Levander, A., 2006a. 3D seismic refraction traveltime tomography at a groundwater contamination site. *Geophysics*, **71**, H67–H78.
Zelt, C. A., Ellis, R. M., and Zelt, B. C., 2006b. 3-D structure across the Tintina strike-slip fault, northern Canadian Cordillera, from seismic refraction and reflection tomography. *Geophysical Journal International*, **167**, 1292–1308.
Zhang, J., ten Brink, U. S., and Toksöz, M. N., 1998. Nonlinear refraction and reflection traveltime tomography. *Journal of Geophysical Research*, **103**, 29,743–29,757.

Cross-references

Deep Seismic Reflection and Refraction Profiling
Earth's Structure, Continental Crust
Ocean Bottom Seismics
Seismic Anisotropy
Seismic Structure at Mid-Ocean Ridges
Seismic, Ray Theory
Seismic, Waveform Modeling and Tomography

TSUNAMI

Steven N. Ward
Institute of Geophysics and Planetary Physics, University of California at Santa Cruz, Santa Cruz, CA, USA

Definition

Dispersive. Characteristic of waves whose velocity of propagation depends on wave frequency. The shape of a dispersive wave packet changes as it moves along.
Eigenfunction. Functional shape of the horizontal and vertical components of wave motion versus depth in the ocean for a specific wave frequency.
Geometrical spreading. Process of amplitude reduction resulting from the progressive expansion of a wave from its source.
Run-up. Final phase of tsunami life starting when the wave shoals to a size equal to the water depth and begins to break, and ending when the water runs over land and reaches its highest level.
Shoal. Process of waves coming into shallow water. Shoaling waves slow, shorten their wavelength, and grow in size.
Wavenumber. Wavenumber k equals 2π divided by wavelength λ. Large wavenumbers associate with short waves and small wavenumbers associate with long waves.

Tsunami are gravity-driven water waves. They belong to the same family as common sea waves that we see every day; however, tsunami are distinct in their mode of generation and in their physical traits. Unlike common sea waves that evolve from persistent winds, most tsunami spring from sudden shifts of the ocean floor. These sudden shifts can originate from undersea landslides and volcanoes, but mostly, submarine earthquakes parent tsunami. Compared to wind-driven waves, tsunami waves have periods, velocities, and wavelengths ten or a hundred times larger and present profoundly different shoreline consequences than do their common cousins.

Tsunami = killer wave?

In the years since the 2004 Sumatra earthquake, everyone has seen disturbing videos of tsunami-caused destruction. Certainly the prospect of a "killer wave" born from some far off earthquake is frightening. More so is the thought that after traveling with stealth great distances, that wave might suddenly rise up without warning at your own doorstep. Could it be me desperately floating by in the next tsunami video?

Understandably, worst-case scenarios of natural hazards come to mind, but it is important to keep perspective.

Tsunami over 2 m high are not common. It takes a submarine earthquake greater than magnitude M8 to source a wave of this size. On a global average, only one M8+ earthquake occurs per year. Of these, maybe one in five strikes under the ocean with an orientation favorable for tsunami excitation. True, tsunami decay in transit and have shorter run-up at distant shores, so wave damage tends to localize within 1,000 km of that one in five quakes. On these accounts, tsunami that induce widespread damage and casualties number only about one per decade. Even at that modest recurrence rate, your personal risk is mitigated increasingly through education and technology. Today, ocean bottom pressure sensors detect stealthy tsunami of a few centimeters height in the open sea. Moreover, with advances in understanding, communication, and implementation since the 2004 Sumatra earthquake, scientists are better, quicker, and more specific at prediction than they were just a decade past. There is a good chance now that you will be warned and take action even if that rare wave does come your way.

Tsunami characteristics

Tsunami period, velocity, and wavelength

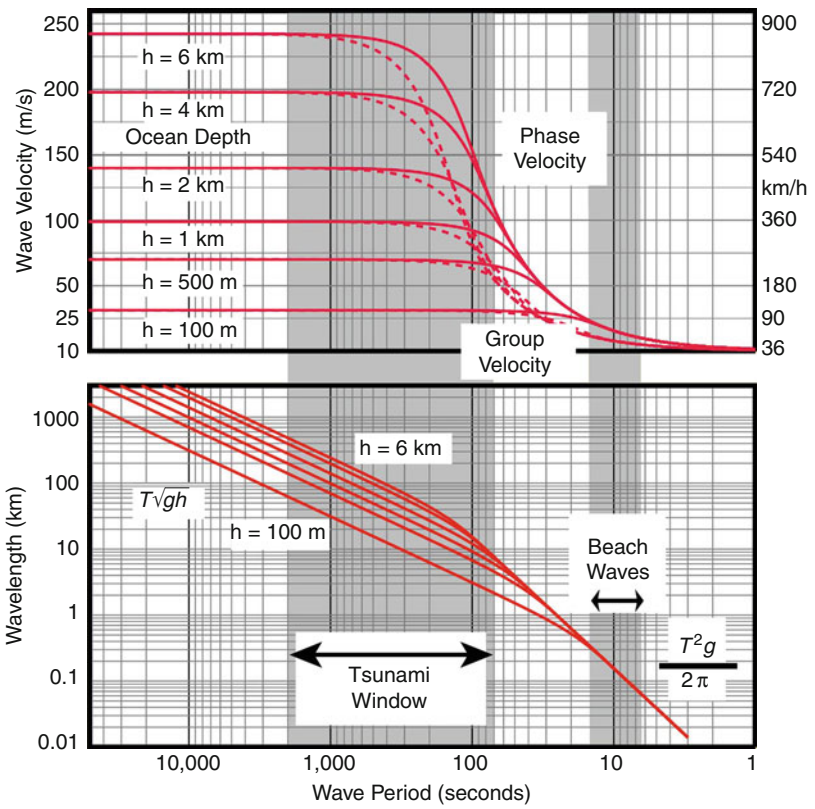
When discussing tsunami, I like to contrast them with something that we all have experienced – waves at the beach. Tsunami belong to the same family as these ordinary ocean waves but with major distinctions: principally tsunami period, tsunami velocity, and tsunami wavelength.

Tsunami period. The period (T) of a wave equals the time elapsed between one passing crest and the next. Common beach waves have a period of about 10 s (darkened column right side Figure 1). The period of tsunami depends upon what creates it (earthquake, landslide, asteroid impact), but I can tell you that tsunami have far greater periods than beach waves. The "tsunami window" in Figure 1 covers waves of 70, 200, 500, or even 2,000 s periods. Unlike "splash and dash" beach waves, tsunami arrive and may continue to flow in for several minutes. I like the description – "The ocean turns into a river." Anyone who has seen those 2004 Sumatra videos can testify that tsunami act more flood-like than wave-like.

Tsunami velocity. Under classical theory, the phase c (ω), and group u(ω) velocity of surface gravity waves on a flat ocean of uniform depth h are:

$$c(\omega) = \sqrt{\frac{gh\ \tanh[k(\omega)h]}{k(\omega)h}} \tag{1}$$

and



Tsunami, Figure 1 Phase velocity c(ω) (*solid lines*) and group velocity u(ω) (*dashed lines*) of tsunami waves on a flat earth covered by oceans of 100 m to 6 km depth (*top*). Wavelength associated with each wave period. The "tsunami window" is marked (*bottom*).

$$u(\omega) = c(\omega)\left[\frac{1}{2} + \frac{k(\omega)h}{\sinh[2k(\omega)h]}\right]. \quad (2)$$

Here, g is the acceleration of gravity (9.8 m/s^2) and $k(\omega)$ is the *wavenumber* associated with a sea wave of frequency $\omega = 2\pi/T$. Wavenumber connects to wavelength as $\lambda(\omega) = 2\pi/k(\omega)$ and to phase velocity as $c(\omega) = \omega/k(\omega)$. Wavenumber also satisfies the relation:

$$\omega^2 = gk(\omega)\tanh[k(\omega)h]. \quad (3)$$

For surface gravity waves spanning 1–50,000 s period, Figure 1 (*top*) plots $c(\omega)$ and $u(\omega)$. These velocities vary widely and increase for waves of longer periods in deeper water. The speed of ordinary sea waves is about 15 m/s (50 km/h) – the speed of a moped. Because of their longer period, waves in the tsunami window travel much more rapidly, reaching 160–250 m/s (600–900 km/h) in the open ocean. It is said that deep-water tsunami travel at the speed of a jet airliner. While true, be aware that wave energy is the quantity that runs ahead at jet speeds. The physical bits of water in a deep-water tsunami move less than a meter per second (see next section). Waves whose velocity varies with frequency are called *dispersive*. During propagation, dispersion "pulls apart" originally pulse-like waves into their component frequencies. Dispersion is strongest for waves whose period falls on the steepest slopes of the group velocity curves in Figure 1.

Tsunami wavelength. Wavelength measures the distance between one wave crest and the next. Wavelength equals the product of wave period times phase velocity $\lambda(\omega) = Tc(\omega)$. Common beach waves have a wavelength of about 100 m – the length of a football field (Figure 1, *bottom*). Tsunami, with their longer period and higher velocity, have much longer wavelengths than beach waves. In the deep ocean, tsunami span 10, 30, even 100 km between crests. If you could stand on one tsunami crest, the next one might be over the horizon. With wavelengths this large, tsunami slopes are very small even if the wave has large amplitude. For ships at sea, tsunami pass completely unnoticed.

Short wave versus long wave: Discussions of waves of length λ in oceans of depth h sometimes include two simplifications: a long wave approximation ($\lambda >> h$, $1/k >> h$) and a short wave approximation ($\lambda << h$, $1/k << h$). Under a long wave approximation [$kh \to 0$, $\tanh(kh) \to$ kh, $\sinh(2kh) \to 2kh$] Equations 1–3 predict nondispersive wave propagation with $c(\omega) = u(\omega) = \sqrt{gh}$. Long wave theory holds for the flat part of the curves in Figure 1 (*top*). Under a short wave approximation, [$kh \to \infty$, $\tanh(kh) \to 1$, $\sinh(kh) \to \infty$] the equations predict dispersive propagation with $c(\omega) = 2u(\omega) = gT/2\pi$. Short wave theory holds to the right in Figure 1 (*top*) where all the curves lie atop each other. Waves in the tsunami window have intermediate character, behaving like shallow water waves at their longest periods and like deep-water waves at their shortest periods. Neither the long nor the short wave simplification serves adequately in tsunami studies. A rigorous treatment requires an approach that works for waves of all lengths.

Tsunami eigenfunctions

Tsunami also differ from waves at the beach in the way they move the ocean. Tsunami *eigenfunctions* describe wave motion in a tsunami mode of a particular frequency. Consider coordinate system (x, y, z) where $\hat{\mathbf{x}}$ points north, $\hat{\mathbf{y}}$ east, and $\hat{\mathbf{z}}$ down. Vertical (u_z) and horizontal (u_x) components of tsunami eigenfunctions normalized to vertical amplitude A_z at the sea surface are:

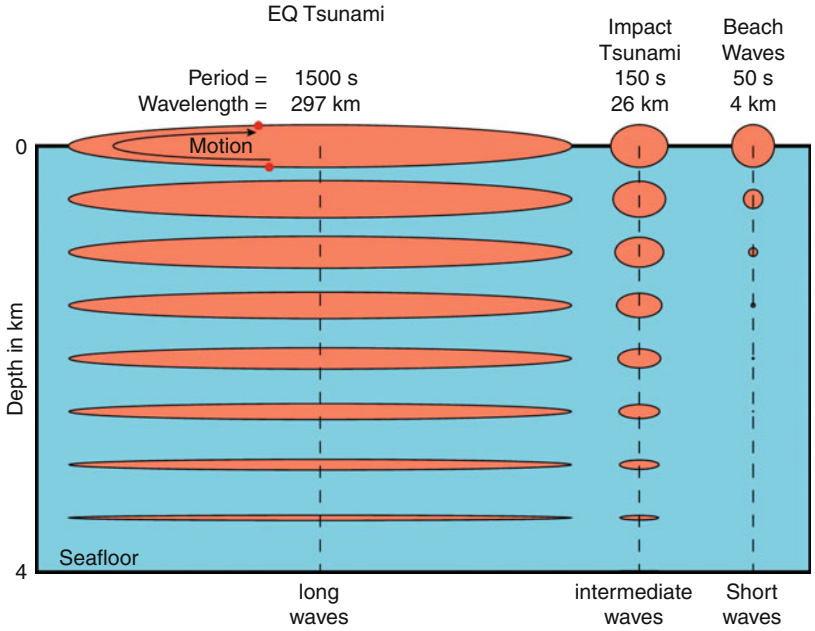
$$u_z(\omega, z) = A_z \frac{k(\omega)g}{\omega^2} \frac{\sinh[k(\omega)(h-z)]}{\cosh[k(\omega)h]} e^{i[k(\omega)x - \omega t]} \quad (4)$$

$$u_x(\omega, z) = A_z \frac{-ik(\omega)g}{\omega^2} \frac{\cosh[k(\omega)(h-z)]}{\cosh[k(\omega)h]} e^{i[k(\omega)x - \omega t]}$$

Figure 2 plots tsunami eigenfunctions versus depth in a 4-km deep ocean at long (1,500 s), intermediate (150 s), and short (50 s) periods. The ellipses trace the path of a water particle as a wave of frequency ω passes. At 50-s period (Figure 2, *right*), ordinary ocean waves have deep-water behavior. Water particles move in circles that decay exponentially from the surface. Sensibly, because the eigenfunctions of short waves do not reach to the seafloor, their velocity is independent of ocean depth [$c(\omega) = 2u(\omega) = gT/2\pi$, Figure 1, *top right*]. The failure of short waves to "feel" the seafloor also means that they cannot be excited by deformations of it. The only means to excite ordinary ocean waves is to disturb the surface.

At 1,500 s period (Figure 2, *left*), the tsunami has a wavelength of λ = 297 km and it acts like a long wave. Although vertical displacement peaks at the ocean surface and drops to zero at the seafloor, horizontal displacement persists undiminished through the ocean column. Unlike ordinary waves that are confined near the sea surface, the energy of a tsunami spreads through the entire depth of the sea. You cannot out dive a tsunami. This fact explains why tsunami can be detected by pressure sensors on the seafloor. A second distinction of tsunami versus ordinary sea waves is that their horizontal motion far exceeds their vertical motion. Every meter of up and down in a 1,500-s tsunami involves ≈10 m of back and forth. If you were to build a shore-based or space-based "tsunami detector," the large horizontal motions might make a better target than the vertical motions. Applications of Doppler radar come to mind.

Toward the short period side of the tsunami window at 150 s (Figure 2, *middle*), λ decreases to 26 km. For these tsunami waves, long wave characteristics begin to break down, and horizontal and vertical motions more closely agree in amplitude.



Tsunami, Figure 2 Tsunami eigenfunctions in a 4-km deep ocean at periods 1,500, 150, and 50 s. Vertical displacements at the ocean surface has been normalized to A_z = 1 m in each case.

From (4), it is easy to deduce the peak velocity of surface water in waves of amplitude A_z as:

$$\left|v_x^{\max}(\omega, z=0)\right| = \frac{A_z k(\omega) g}{\omega} = \frac{A_z g}{c(\omega)}. \tag{5}$$

For long waves $c(\omega)$ equals$\sqrt{gh}$, and peak surface water velocity in a 1-m amplitude tsunami in 4,000 m of waver would be just 4.9 cm/s. Peak water velocity at the sea floor

$$\begin{aligned}\left|v_x^{\max}(\omega, z=h)\right| &= \frac{A_z\, k(\omega) g}{\omega \cosh[k(\omega)h]} \\ &= \frac{A_z g}{c(\omega)\cosh[k(\omega)h]}\end{aligned} \tag{6}$$

is even less. As mentioned, although tsunami travel at several hundred meters per second, the water itself moves at a tiny fraction of this.

Tsunami excitation

Tsunami get started in many ways. Suppose that the seafloor at positions $\mathbf{r}_0$ uplift instantaneously by an amount $u_z^{\text{bot}}(\mathbf{r}_0)$ at time $\tau(\mathbf{r}_0)$. Under classical tsunami theory in a uniform ocean of depth h, this sea bottom disturbance produces surface tsunami waveforms (vertical component) at observation point $\mathbf{r} = x\hat{\mathbf{x}} + y\hat{\mathbf{y}}$ and time t of

$$u_z^{\text{surf}}(\mathbf{r}, t) = \text{Re} \int_{\mathbf{k}} d\mathbf{k}\, \frac{e^{i[\mathbf{k}\cdot\mathbf{r} - \omega(k)t]}}{4\pi^2 \cosh(kh)} F(\mathbf{k}). \tag{7a}$$

with

$$F(\mathbf{k}) = \int_{\mathbf{r}_0} d\mathbf{r}_0\, u_z^{\text{bot}}(\mathbf{r}_0) e^{-i[\mathbf{k}\cdot\mathbf{r}_0 - \omega(k)\tau(\mathbf{r}_0)]} \tag{7b}$$

$k = |\mathbf{k}|$, and $\omega^2(k) = gk\tanh(kh)$. The integrals in (7a, 7b) cover all wavenumber space and locations $\mathbf{r}_0$ where the seafloor disturbance $u_z^{\text{bot}}(\mathbf{r}_0) \neq 0$.

Equation 7a looks scary but it has three identifiable pieces:

(a) The $F(\mathbf{k})$ term is the wavenumber spectrum of the seafloor uplift. This number relates to the amplitude, spatial and time distribution of the uplift. Tsunami trains are dominated by wavenumbers in the span where $F(\mathbf{k})$ is the greatest. The peak of $F(\mathbf{k})$ corresponds to the characteristic dimension of the uplift. Large-dimensioned uplifts produce longer wavelength, lower frequency tsunami than small-dimensioned sources.

(b) The $1/\cosh(kh)$ term comes from the tsunami eigenfunction shapes (4) and it acts to low-pass filter the source spectrum $F(\mathbf{k})$. Because $1/\cosh(kh) \to 1$ when $kh \to 0$, and $1/\cosh(kh) \to 0$ when $kh \to \infty$, the filter favors long waves. Due to the low-pass filter effect of the ocean layer, only wavelengths of the uplift that exceed three times the ocean depth (i.e., $kh = 2\pi h/\lambda <\approx 2$) contribute much to tsunami.

(c) The exponential term in (7a) contains all of the propagation information including travel time, *geometrical spreading*, and frequency dispersion.

By rearranging Equation 7a, 7b, vertical tsunami motions at **r** can also be written as:

$$u_z^{\text{surf}}(\mathbf{r},t) = \text{Re}\int_0^\infty \frac{k\,\text{d}k e^{-i\omega(k)t}}{2\pi\cosh(kh)} \sum_{n=-\infty}^{\infty} \text{J}_n(kr)e^{in\theta} F_n(k)$$

with (8)

$$F_n(k) = \int_{\mathbf{r}_0} \text{d}\mathbf{r}_0\, u_z^{\text{bot}}(\mathbf{r}_0)\, J_n(k\text{r}_0)\, e^{i(\omega(k)\tau(\mathbf{r}_0)-n\theta_0)}$$

Here θ marks azimuth from north (the $\hat{\mathbf{x}}$ direction) of the observation point **r** from the coordinate origin. The $J_n(x)$ are cylindrical Bessel functions. For simply distributed uplift sources, (8) might be easier to evaluate than (7a, 7b).

Tsunami excitation by earthquakes

Earthquakes produce 80–90% of tsunami. Not surprisingly, earthquake features determine many sea wave characteristics. Earthquakes result from slip on faults and three primary parameters describe the process – moment, mechanism, and depth.

Moment measures earthquake strength. Moment M_o is the product of rigidity μ of the source region's rocks, fault area A, and average fault slip Δu. Earthquake moment and earthquake magnitude tie through a number of empirical formulae. One formula defines moment magnitude M_w as $M_w = (2/3)(\log M_o - 9.05)$. Earthquake moment varies by 2×10^4 within the magnitude range $6.5 \le M_w \le 9.5$ (Table 1). Even without a detailed understanding of tsunami generation, it is safe to suppose that the larger the earthquake moment, the larger the tsunami, all else fixed.

Mechanism specifies the orientation of the earthquake fault and the direction of slip on it. Usually, faults are idealized as plane rectangles with normal $\hat{\mathbf{n}}$. Three angles then, summarize earthquake mechanisms – the strike and dip of the fault and the angle of slip vector $\hat{\mathbf{a}}$ measured from the horizontal in the plane of the fault. (Seismologists call this angle the "rake.") The role of fault mechanism on tsunami production is not as obvious as the influence of moment; however, one might suspect that earthquakes that affect large vertical displacements of the seafloor would be more effective than faults that make large horizontal displacements.

Earthquake *depth* needs no explanation. Because seafloor shifts cause tsunami, the distance of the fault from the seafloor should be important. Presumably, deep earthquakes produce less potent tsunami than similar shallow earthquakes.

Numerical, or synthetic waveforms quantify the roles of earthquake parameters on tsunami generation. For illustration, insert into (7a, 7b) the surface uplift pattern $u_z^{\text{bot}}(\mathbf{r}_0)$ of a small earthquake fault (point source really) placed at depth d in a halfspace. Further assume that the uplift occurs instantly with $\tau(\mathbf{r}_0) = 0$. (Actually, real earthquakes uplift the seafloor over several, or several tens of, seconds. This distinction is not a big issue because tsunami waves have periods of many hundreds of seconds. Uplifts taking a few dozen seconds to develop look "instantaneous" to tsunamis.) Equation 7a, 7b becomes

$$u_z^{\text{surf}}(\mathbf{r},t) = \int_0^\infty k\,\text{d}k \frac{\cos\omega(k)t}{2\pi\cosh(kh)}\left[A\Delta u \mathbf{M}_{ij}\varepsilon_{ij}\right], \qquad (9)$$

where

$$\begin{aligned}
\varepsilon_{xx} &= -\frac{1}{4}\left(\frac{\mu}{\lambda+\mu} - kd\right)[J_0(kr) - J_2(kr)\cos 2\theta]e^{-kd}\\
\varepsilon_{yy} &= -\frac{1}{4}\left(\frac{\mu}{\lambda+\mu} - kd\right)[J_0(kr) + J_2(kr)\cos 2\theta]e^{-kd}\\
\varepsilon_{xy} = \varepsilon_{yx} &= \frac{1}{4}\left(\frac{\mu}{\lambda+\mu} - kd\right)[J_2(kr)\sin 2\theta]e^{-kd}\\
\varepsilon_{zz} &= -\frac{1}{2}\left(\frac{\mu}{\lambda+\mu} + kd\right)[J_0(kr)]e^{-kd}\\
\varepsilon_{xz} = \varepsilon_{xz} &= \frac{kd}{2}[J_1(kr)\cos\theta]e^{-kd}\\
\varepsilon_{yz} = \varepsilon_{zy} &= \frac{kd}{2}[\text{J}_1(kr)\sin\theta]e^{-kd}
\end{aligned} \qquad (10)$$

The six elements of symmetric tensor

$$M_{jk} = \left(\hat{a}_j\hat{n}_k + \hat{n}_j\hat{a}_k\right) \qquad (11)$$

capsulize the mechanism of the earthquake. In (11), $\hat{\mathbf{n}}$, $\hat{\mathbf{a}}$ are the fault normal and slip vectors introduced above. A pure dip slip earthquake on a vertical north–south

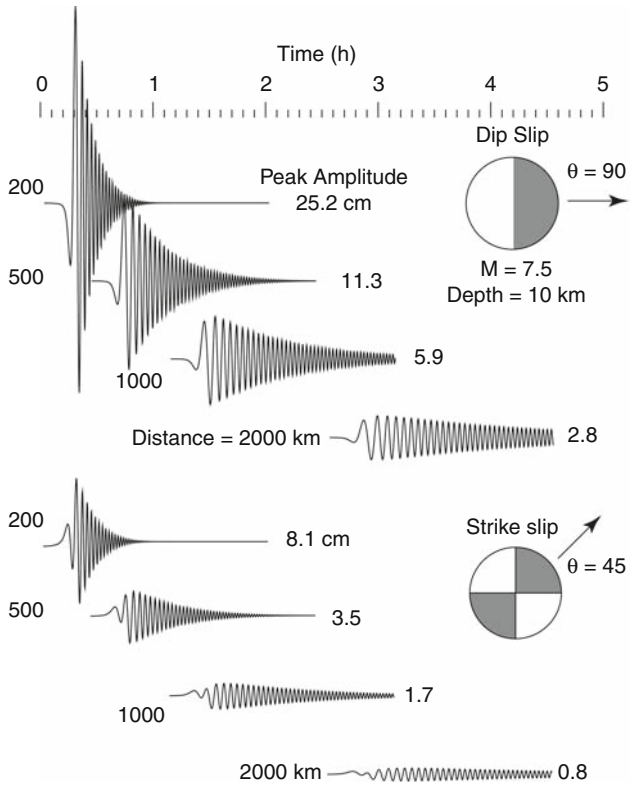
Tsunami, Table 1 Relationship between earthquake magnitude and moment with values of fault area, length, and mean slip for typical tsunami-generating earthquakes. This paper assumes $\log(L) = 0.5\,M_w - 1.8$, $\Delta u = 2 \times 10^{-5}\,L$, and $\lambda = \mu = 5 \times 10^{10}$ Pa

Magnitude M_w	Moment M_0 (Nm)	Area A (km^2)	Length L (km)	Width W (km)	Slip Δu (m)
6.5	6.3×10^{18}	224	28	8	0.56
7.0	3.5×10^{19}	708	50	14	1.00
7.5	2.0×10^{20}	2,239	89	25	1.78
8.0	1.1×10^{21}	7,079	158	45	3.17
8.5	6.3×10^{21}	22,387	282	79	5.66
9.0	3.5×10^{22}	70,794	501	141	10.0
9.5	2.0×10^{23}	223,872	891	251	17.8

trending fault for instance, has $\hat{\mathbf{n}} = \hat{\mathbf{y}}$ and $\hat{\mathbf{a}} = \hat{\mathbf{z}}$, so $M_{yz} = M_{zy} = 1$ and $M_{xx} = M_{yy} = M_{zz} = M_{xz} = M_{zx} = M_{xy} = M_{yx} = 0$.

The bracketed terms in (9) contain all of the relationships between earthquake parameters and tsunami features. Some relationships are easy to spot: tsunami amplitudes from earthquakes are proportional to the product of fault area and average slip ($A\Delta u$); tsunami amplitudes decrease with earthquake depth via the e^{-kd} terms. The ε_{ij} provide the dependence of tsunami amplitude and azimuthal radiation pattern on source type. Equation 10 says that tsunami from point sources radiate in azimuthal patterns no more intricate than $\sin 2\theta$ or $\cos 2\theta$.

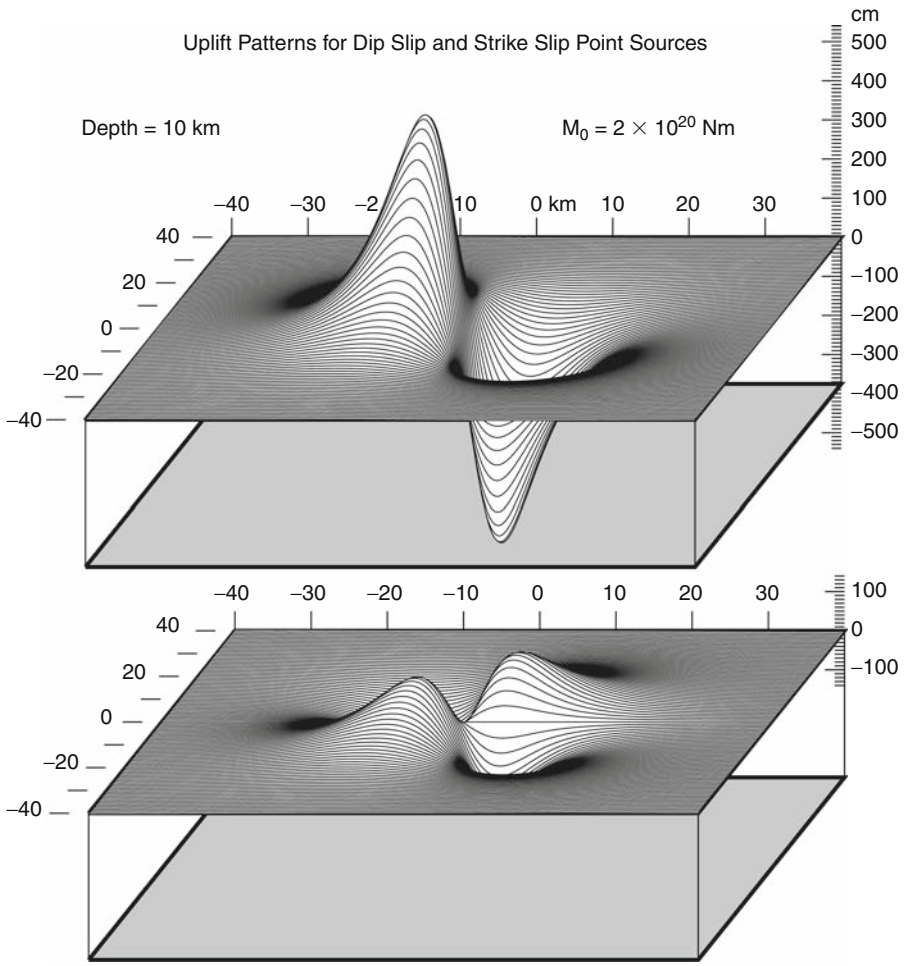
Figure 3 shows 5 h of tsunami waveforms calculated from (9) at distances of r = 200, 500, 1,000, 2,000 km from dip slip ($M_{yz} = M_{zy} = 1$) and strike slip ($M_{xy} = M_{yx} = 1$) point sources of magnitude $M_w = 7.5$ ($\Delta uA = 3.98 \times 10^6$ m^3. See Table 1) buried at 10 km depth. Sea waves from these sources have radiation patterns of $\sin\theta$ and $\sin 2\theta$, respectively. I compute the waveforms in Figure 3 at the azimuth of maximum strength, $\theta = 90°$ and $\theta = 45°$. Frequency dispersion, with the long periods arriving first, is the most conspicuous feature of the waveforms. Tsunami onset rapidly. They reach maximum height in the first few cycles and then decay slowly over an hour or more. Even for this large earthquake, tsunami beyond 500 km distance reach just a few centimeters – hardly killer waves. Note that if the observation direction was west versus east (Figure 3, *top*), or northwest versus northeast (Figure 3, *bottom*), the tsunami waveforms would be inverted. Whether tsunami onset in a withdrawal or an inundation at a particular shoreline strictly depends on the style of faulting and the relative positions of the shore and the fault.



Tsunami, Figure 3 Synthetic record sections of vertical tsunami motions at distances of 200, 500, 1,000, and 2,000 km from point dip slip (*top*) and strike slip (*bottom*) earthquakes of magnitude $M_w = 7.5$ and depth 10 km. Time runs for 5 h and the peak amplitude of each trace is given in centimeters at the right. The lower half of the focal sphere and azimuth of observation θ are shown toward the right. For other directions, the waveforms should be scaled by $\sin\theta$ and $\sin 2\theta$, respectively.

Figure 3 demonstrates that for point sources, dip slip earthquakes produce three or four times larger tsunami than strike slip earthquakes of equal moment. The differences in generation efficiency are understood most easily by considering directly the seafloor deformation patterns $u_z^{\text{bot}}(\mathbf{r}_0)$. I find these by setting $t = 0$ and $h = 0$ in (9) so $\cos\omega(k)t/\cosh(kh) = 1$. Figure 4 pictures the uplift patterns for the two faults of Figure 3 where two ($\sin\theta$) and four-lobed ($\sin 2\theta$) deformations spread over a region 40 km wide. The most striking contrast in the fields is maximum vertical displacement – 1.4 m for the strike slip versus 5.4 m for the dip slip. It is no coincidence that the ratio of maximum uplift for these two faults replicates the ratio of tsunami heights in Figure 3. After all, vertical seafloor deformation drives tsunami and vertical deformation is controlled largely by the rake of the slip vector $\hat{\mathbf{a}}$. Strike slip faults have a rake of 0° or 180°. Dip slip faults have rake equal ±90°. Further simulations show that, excepting very shallow, nearly horizontal faults, dip is not a terribly significant factor in tsunami production.

The sea surface cross sections in Figure 5 (*left* and *middle*) chronicle the birth and early life of a sea wave spawned by M7.5 thrust earthquakes on 45° dipping planes. In these figures, I replace the idealized point sources of Figure 3 with faults of typical dimension (L = 89 km, W = 25 km, and Δu = 1.78 m; see Table 1). In Figure 5 (*left*) the fault reaches to the sea floor. In Figure 5 (*middle*) the fault stops 30 km down. Soon after the earthquake, the sea surface forms "dimples" similar to those on the deformed sea floor. The sea surface dimples, however, are smoother and a bit lower in amplitude because of the $1/\cosh(kh)$ low-pass filtering effect of the ocean layer. After a time roughly equal to the dimension of the uplift divided by tsunami speed $\sqrt{gh}$, the leading edges of the wave organize and begin to propagate outward as expanding rings (Figure 5, *right*). Early on, the wave appears as a single pulse. Characteristic tsunami dispersion begins to be seen only after 10 or 20 min. Consequently, for shorelines close to tsunami sources, seismic sea waves arrive mostly as a single pulse. For distant shorelines, sea waves arrive with many oscillations, dispersion having spread out the initial pulse.



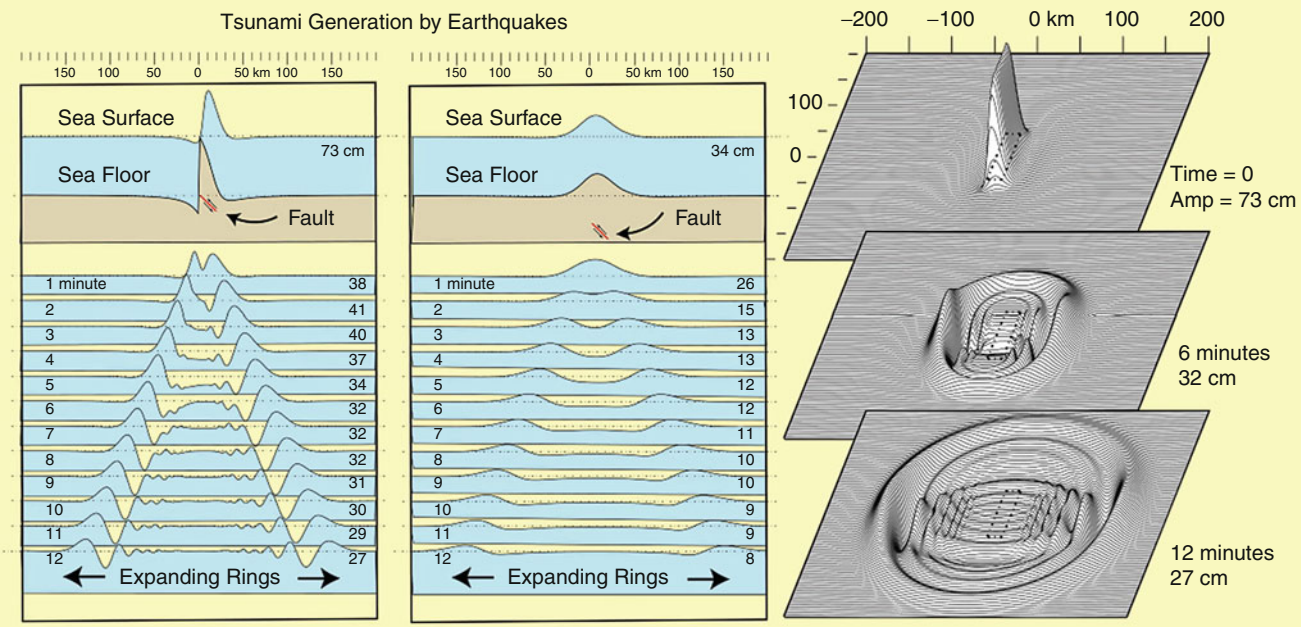
Tsunami, Figure 4 Static vertical displacements of the seafloor for the dip slip (*top*) and strike slip (*bottom*) earthquake point sources that generated the tsunami of Figure 3. Maximum excursions of the seafloor are 5.4 and 1.4 m, respectively.

The e^{-kd} terms in the tsunami excitation functions (10) let shallow earthquakes excite higher frequency tsunami than deep earthquakes (compare Figure 5 *left*, with Figure 5 *middle*). Higher frequency waves travel more slowly than longer period waves, so high frequency waves contribute to peak tsunami height only while it propagates as a single pulse. After a few hundred kilometers of travel, high frequency waves drift to the back of the wave train (Figure 5, *left*) and no longer add to the tsunami maximum. At 200 km distance, the shallow earthquake generates a wave about 3 times larger than the 30-km deep event. If you track the waves out to 2,000 km, however, you would find that the extra high frequencies in the shallow event will have fallen behind and that the maximum wave heights for the two events would be nearly equal. Beyond 2,000 km distance, any earthquake depth less than 30 km appears to be equally efficient in tsunami genesis.

Faults of finite size, like those in Figure 5, radiate tsunami in distributions more complex than the $\sin\theta$ and $\sin 2\theta$ patterns from point sources. The largest earthquakes have fault lengths of several hundred kilometers. Simulations show that long earthquake faults preferentially emit tsunami in a tight beam perpendicular to the fault strike, regardless of the focal mechanism. This preferential beaming simplifies tsunami forecasting because it tells us which direction to look for the biggest waves.

Tsunami excitation from submarine landslides

Earthquakes parent most tsunami, but other things do too. For instance, earthquake shaking often triggers landslides. If the slide happens under the sea, then tsunami may form. Consider a seafloor landslide confined in a rectangle of length L and width W. Let a constant uplift u_0 start along one width of the rectangle and run down its length ($\hat{\mathbf{x}}$ direction, say) at velocity v_r, that is, $\tau(\mathbf{r}_0) = x/v_r$. Placing this $u_z^{\text{bot}}(\mathbf{r}_0)$ and $\tau(\mathbf{r}_0)$ into (7a, 7b), I find the tsunami from this uplift source at observation point $\mathbf{r}$ and time t to be



Tsunami, Figure 5 *Left*. Cross sections of expanding tsunami rings from a M7.5 thrust earthquake. The fault strikes north–south (into the page) and the sections are taken east–west. Elapsed time in minutes and maximum amplitude in centimeters are given at the left and right sides. *Middle*. Cross sections of expanding tsunami rings from the same M7.5 thrust earthquake now buried 30 km. Deeper earthquakes make smaller and longer wavelength tsunami. *Right*. Views of the expanding tsunami rings from the shallow earthquake case at t = 6 and 12 min. The dashed rectangle in the center traces the surface projection of the fault. For large earthquakes, nearly all tsunami energy beams perpendicular to the strike of the fault (*toward the left and right in this picture*).

$$u_z^{\text{surf}}(\mathbf{r},t) = \frac{u_0 LW}{4\pi^2} \operatorname{Re}\int_k d\mathbf{k}\, \frac{e^{i(\mathbf{k}\cdot\mathbf{r}-\omega(k)t)}e^{-iX(\mathbf{k})}}{\cosh(kh)}\,\frac{\sin X(\mathbf{k})}{X(\mathbf{k})}\,\frac{\sin Y(\mathbf{k})}{Y(\mathbf{k})} \tag{12}$$

where

$$X(\mathbf{k}) = \frac{kL}{2}\left(\hat{\mathbf{k}}\cdot\hat{\mathbf{x}} - c(k)/v_r\right) \quad \text{and} \quad \mathrm{Y}(\mathbf{k}) = \frac{\mathrm{kW}}{2}\left(\hat{\mathbf{k}}\cdot\hat{\mathbf{y}}\right).$$

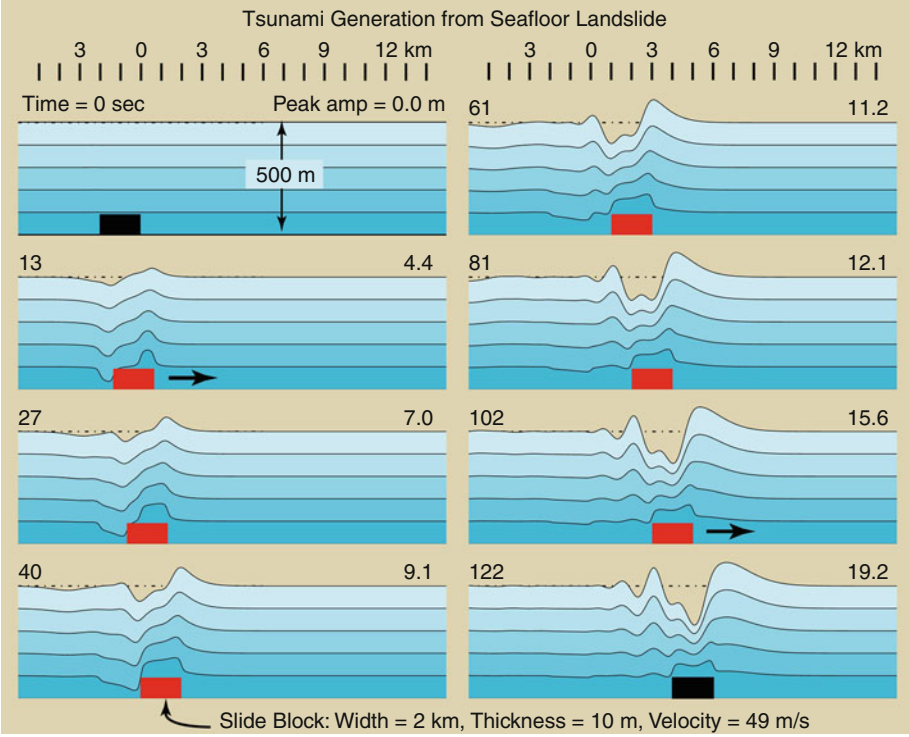
The $X(\mathbf{k})$ and $Y(\mathbf{k})$ factors, because they depend on the relative positions of the observation point and the landslide source, instill radiation patterns to the tsunami much like the ε_{ij} do for earthquakes.

By adding a trailing excavation source to the uplift source (12), I simulate tsunami from a sliding submarine block. Figure 6 pictures in cross section the waves stirred from a 2-km wide block, 10-m thick sliding at 49 m/s for 6 km along the bottom of an ocean 500 m deep. As the block moves, the water must go around the obstruction. Water gets pushed up in front and drawn down behind the slide block. Experiments reveal that, depending on the aspect ratio of the block and the ratio of slide velocity to the tsunami phase velocity, significant beaming and amplification of the tsunami are possible. In particular, when the slide velocity v_r approaches the tsunami speed $c(k) \approx \sqrt{gh}$, $X(\mathbf{k})$ is nearly zero for waves traveling in the slide direction. In this direction, the waves from different parts of the landslide arrive nearly "in phase" and constructively build. Figures 6 and 7 highlight the beaming and amplification effects. In both cases, witness the large tsunami pulse sent off in the direction of the slide. (Quicktime movies of Figures 6 and 7 and many other Figures in this paper are available for viewing; see Table 2.) Submarine landslides are prime suspects in the creation of "surprise tsunami" from small or distant quakes. Surprise landslide tsunami might initiate well outside of the earthquake uplift area or be far larger than expected given the magnitude of the quake. Historical examples of surprise landslide tsunami include the 1929 Grand Banks and the 1998 Papua New Guinea waves.

Tsunami excitation from impacts

In addition to tsunami from earthquakes and landslides, another class holds interest – those generated from surface detonations, explosive volcanoes, and asteroid strikes. I call waves from these sources "impact tsunami." Imagine that the initial stage of cratering by an impact excavates a radially symmetric, parabolic cavity of depth D_C, and radius R_C:

$$u_z^{\text{impact}}(\mathbf{r}) = D_C(1 - r^2/R_C^2) \qquad r \le \sqrt{2}R_C = R_D$$



Tsunami, Figure 6 Tsunami produced by a submarine landslide. The slide block is 10 m thick and 2 km wide. The slide starts to the left and runs to the right at 49 m/s. Note the amplification of the wave in the slide direction.

$$u_z^{\text{impact}}(\mathbf{r}) = 0 \qquad r > \sqrt{2}R_C = R_D$$

Based on Equation 8, the impact tsunami at observation point **r** and time t is

$$u_z^{\text{surf}}(\mathbf{r}, t) = \int_0^\infty \frac{k\,dk}{2\pi} \cos[\omega(k)t] J_0(kr) F_0(k), \qquad (13)$$

where

$$F_0(k) = \int_{\mathbf{r}_0} r\,d\mathbf{r}_0\, u_z^{\text{impact}}(\mathbf{r}_0) J_0(kr_0) = \frac{2\pi D_C R_D}{k} J_3(kR_D).$$

The principal distinction between (8) and (13) is the absence in the latter of the $1/\cosh(kh)$ low-pass ocean filter. Impacts crater the surface of the ocean not the sea floor, so this filter does not come into play.

If we restrict ourselves to asteroid-induced tsunami and suppose that the depth of the impact cavities equals 1/3 their diameter $d_c = 2R_C$, then D_c relates scales with the density, velocity, and radius of the impacting asteroid as

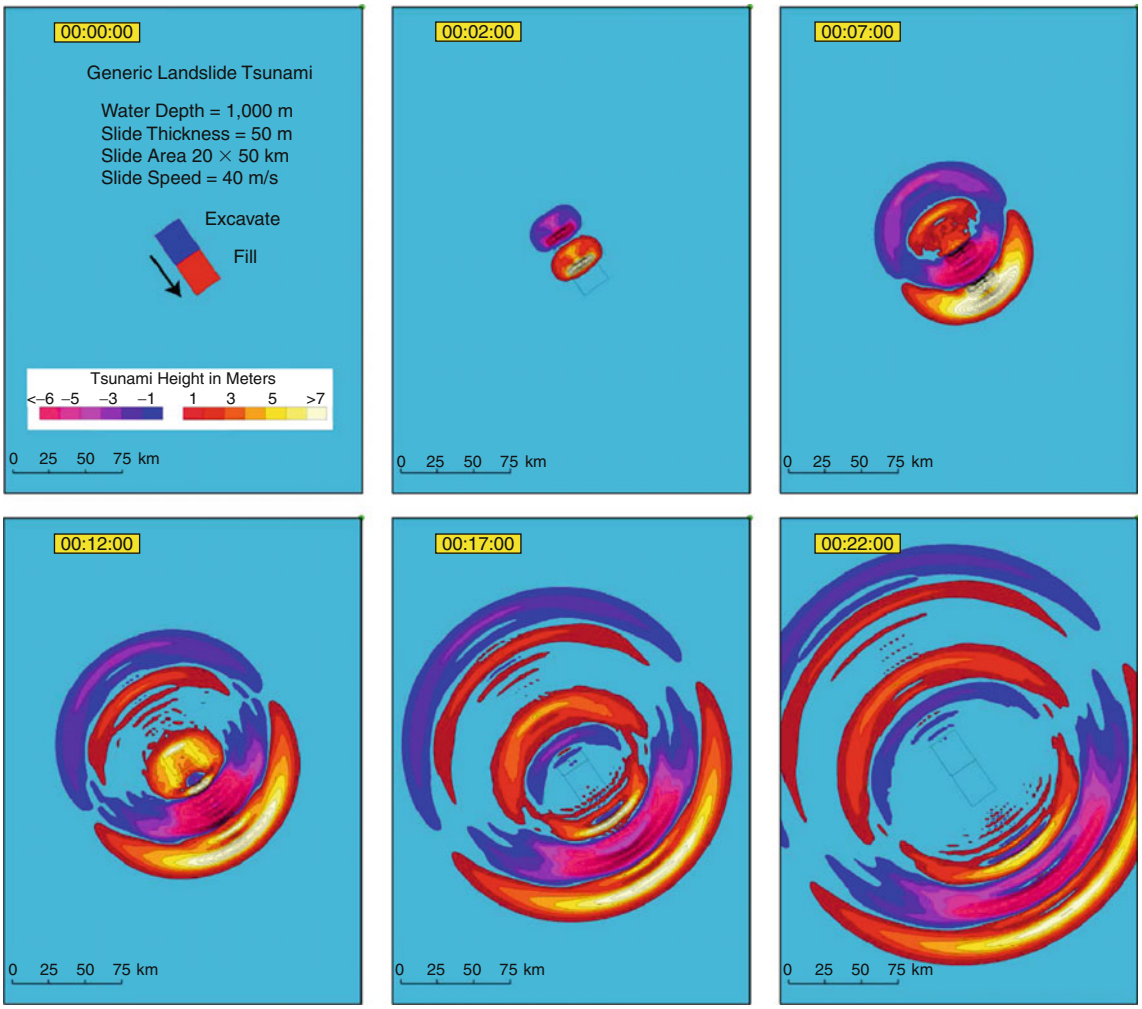
$$D_C = d_c/3 = (8\varepsilon\rho_I V^2{}_I/9\rho_w g)^{1/4} R_I^{3/4} \qquad (14)$$

In (14), ε is the fraction of the kinetic energy of the asteroid that goes into the tsunami wave. High velocity impact experiments suggest that $\varepsilon \approx 0.15$. With $\rho_I = 3$ gm/cm^3, $\rho_w = 1$ gm/cm^3, and $V_I = 20$ km/s, (14) returns crater depths of 1,195, 2,010, and 4,000 m for asteroids of radius $R_I = 50$, 100, and 250 m.

Figure 8 plots cross sections of the expanding rings of tsunami waves induced by the impact of a 200-m diameter asteroid at 20 km/s as computed by Equation 13. Within 100 km of ground zero, tsunami from moderate size (100–250 m) asteroids have heights of many hundreds of meters and dwarf the largest (10–15 m) waves parented by earthquakes. Fortunately, impact tsunami fall on the steep-sloped part of the group velocity curve (Figure 1) so they decay faster with distance than earthquake tsunami (more like r^{-1} versus $r^{-3/4}$ for earthquake tsunami). Figure 9 illustrates how dispersion quickly pulls apart the initial impulse into dozens of waves. At 1,000 and 3,000 km, the tsunami in Figure 8 would decay to 6 and 2 m amplitude, respectively – still a concern, but not catastrophic. For perspective, asteroids with diameters >200 m impact Earth about every 10,000 years, far less frequently than great M9 earthquakes that strike the planet once in 25 years or so.

Tsunami propagation

In uniform depth oceans, tsunami propagate out from their source in circular rings (e.g., Figure 9) with ray paths that look like spokes on a wheel. In real oceans, tsunami



Tsunami, Figure 7 Map view of a tsunami from a landslide block similar to Figure 6. Note the dipole form of the wave in its early stages and the beaming in the southeast direction of the slide.

speeds vary place to place (even at a fixed frequency) so tsunami ray paths refract and become bent. Consequently, in real oceans, both tsunami travel time and amplitude have to be adjusted relative to their values in uniform depth ones. To propagate tsunami in real oceans, I find it best to keep tsunami "mode-like" versus depth but "ray-like" horizontally. I first transform the various integrals over wavenumber (7, 8, 9, etc.) to integrals over frequency because wave frequency, not wave number, is conserved throughout. Using the relations $u(\omega) = \mathrm{d}\omega/\mathrm{d}k$ and $c(\omega) = \omega/k(\omega)$, I find that tsunami vertical motions from (13) for instance, are to a good approximation:

$$u_z^{\mathrm{surf}}(\mathbf{r},t) = \int_0^\infty \frac{\cos(\omega t)\mathrm{d}\omega}{2\pi k(\omega)u(\omega)} J_0(\omega T(\mathbf{r},\omega))F_0(k(\omega)) G(\mathbf{r})S_{\mathrm{L}}(\omega,\mathbf{r}) \quad (15)$$

In (15) the travel time of waves of frequency ω has been changed from $r/c(\omega)$ to

$$T(\mathbf{r},\omega) = \int_{\mathrm{raypath}} \mathrm{d}r/c(\omega,h(\mathbf{r})), \quad (16)$$

where the integral path traces the tsunami ray from the source to the observation point. Equation 15 also incorporates a new shoaling factor $S_{\mathrm{L}}(\omega,\mathbf{r})$ that accounts for wave height changes due to water depth and new ray geometrical spreading factor $G(\mathbf{r}) \leq 1$ that takes into consideration the reduction of wave amplitudes into shadow zones.

Real geometrical rays for tsunami can be very messy (*top left*, Figure 10). Rather, I employ "network rays" that are simply minimum time paths between source and receiver (*top right*, Figure 10). Unlike geometrical rays, network rays possess neither caustics nor shadow zones.

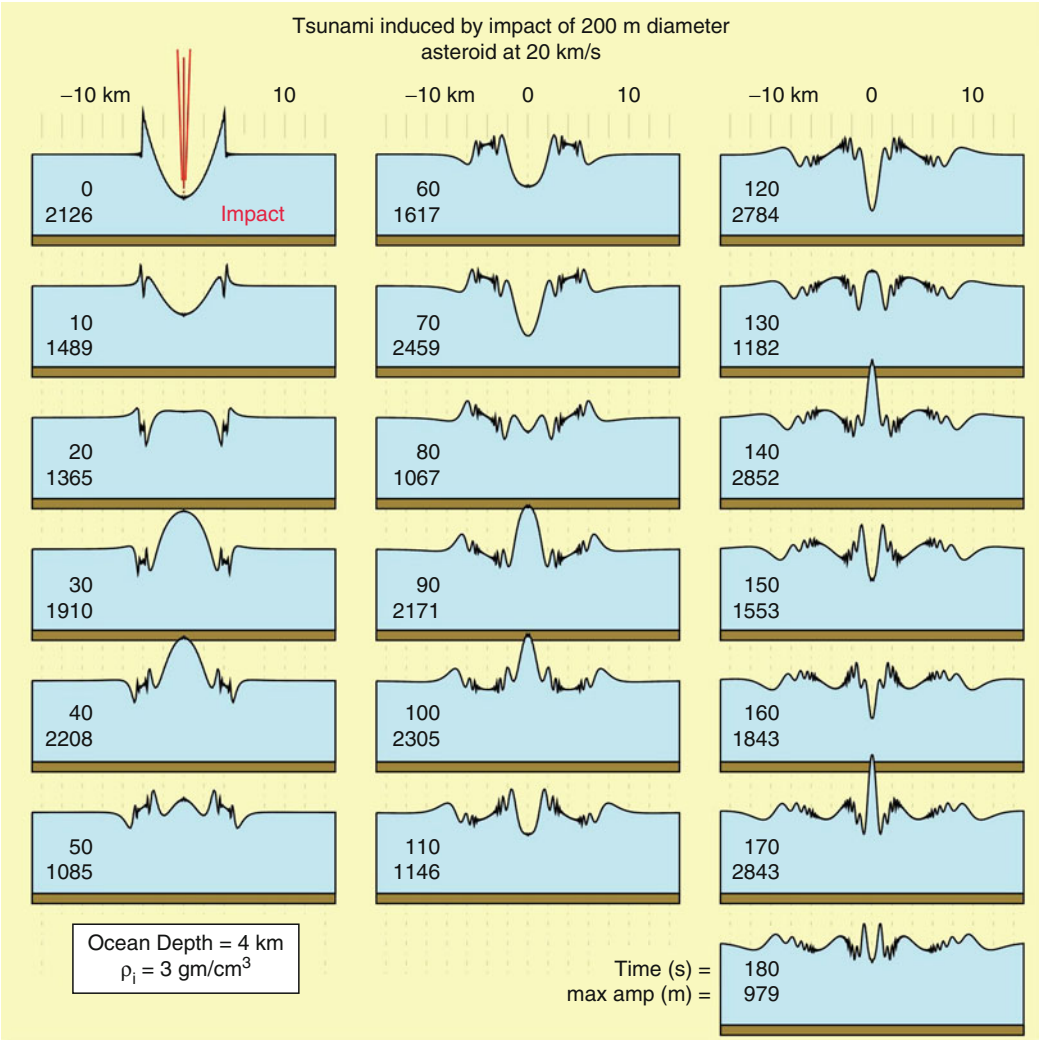
Tsunami, Table 2 Quicktime movie links to many of the figures in this article, plus a few others. If you want your own copy: Play the Movie in a Browser, then click "Save to Disk."

Figure 6	http://es.ucsc.edu/~ward/landslide(1_big).mov
Figure 7	http://es.ucsc.edu/~ward/generic-slide.mov
Figure 8	http://es.ucsc.edu/~ward/imp-parab1.mov
Figure 9	http://es.ucsc.edu/~ward/impact-example-3d.mov
Figure 13	http://es.ucsc.edu/~ward/RU(1200s-20m-0.5).mov
Figure 14	http://es.ucsc.edu/~ward/galveston-2m-close.mov
	http://es.ucsc.edu/~ward/galveston-2m.mov
Figure 15	http://es.ucsc.edu/~ward/indo-3D3.mov
	http://es.ucsc.edu/~ward/indo-3D2.mov
Figure 16	http://es.ucsc.edu/~ward/LP-3D.mov
Figure 17	http://es.ucsc.edu/~ward/guinea(270).mov
	http://es.ucsc.edu/~ward/ire-nn.mov
Figure 18	http://es.ucsc.edu/~ward/samoa-92909(c).mov

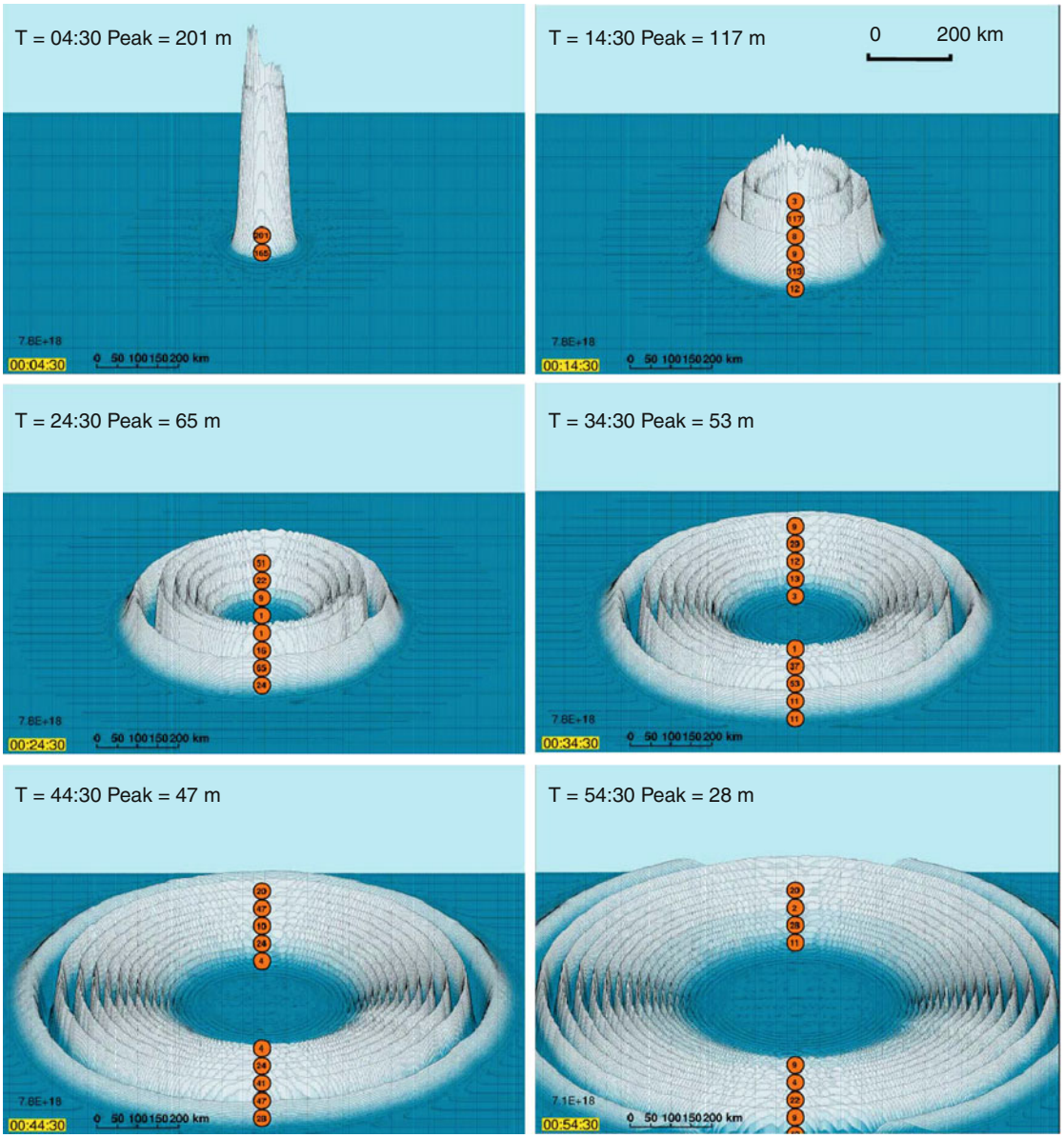
To compute the ray geometrical spreading factor $G(\mathbf{r})$, I replace the curvature of the wave front with curvature of the ray in the path integral (Figure 10, *bottom row*):

$$G(\mathbf{r}) = \exp\left[-\frac{1}{2}\int_{\text{raypath}} \frac{\partial\alpha(s)}{\mathrm{d}w}\mathrm{d}s\right] \Rightarrow \exp\left[-\frac{1}{2}\int_{\text{raypath}} \left[\frac{\partial\theta(s)}{\mathrm{d}s}\right]\mathrm{d}s\right]. \tag{17}$$

In this approach the distinction between wave refraction and diffraction blurs.



Tsunami, Figure 8 Computed tsunami induced by the impact of a 200-m diameter asteroid at 20 km/s. The waveforms (shown at 10 s intervals) trace the surface of the ocean over a 30 km cross section that cuts rings of tsunami waves expanding from the impact site at $x = 0$. Maximum amplitude in meters is listed to the left.



Tsunami, Figure 9 Expanding rings of an impact tsunami in a 3D-like style. Note the strong wave spreading due to dispersion.

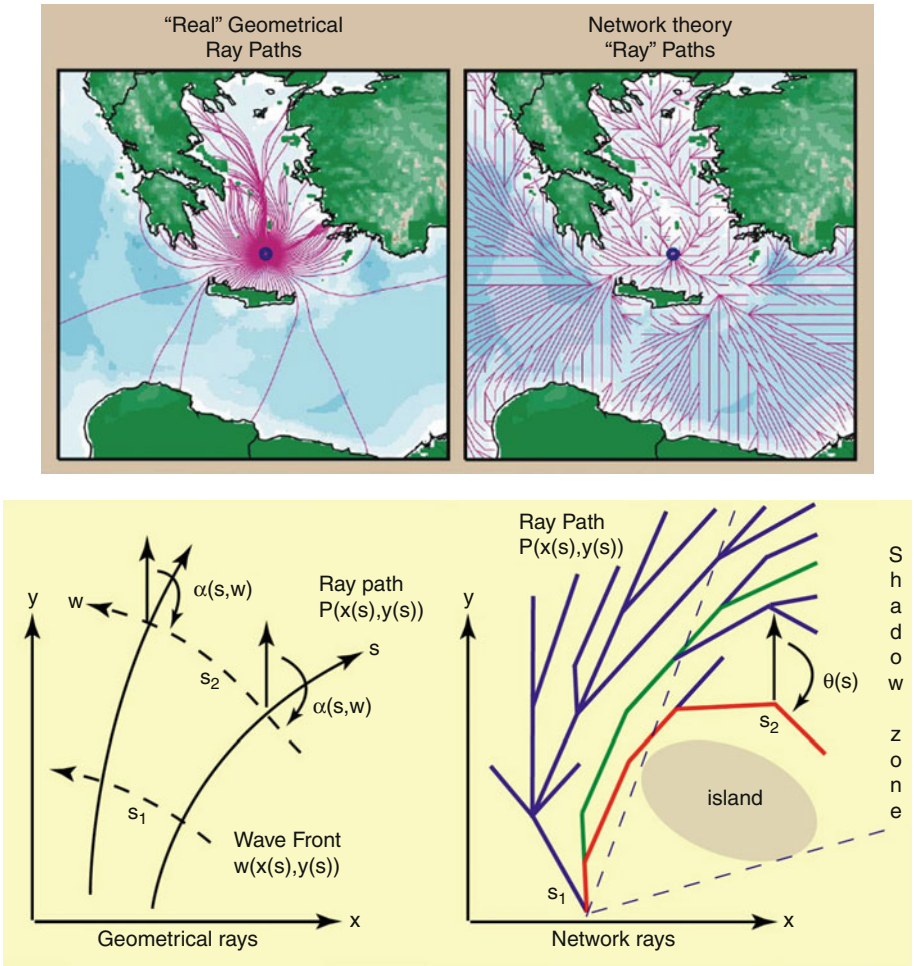
Tsunami shoaling and run-up

Shoaling. Toward shore, real oceans shallow and the waves carried on them amplify in a process called *shoaling*. Shoaling is easy to understand. The tsunami velocity depends on the ocean depth so as water shallows, tsunami waves slow down. Because their frequency is fixed, the wavelength of a slowing tsunami decreases (Figure 11, *inset*) and its energy compresses horizontally. Secondly, because a tsunami occupies the entire water column, as it enters shallow water its energy also gets compressed vertically. The only way for the compressing wave to maintain the same energy flux is for it to grow in amplitude. For the shoaling factor in (15), linear theory gives:

$$S_{\mathrm{L}}(\omega, \mathbf{r}) = \sqrt{\frac{u(\omega,h_0)}{u(\omega,h(\mathbf{r}))}}. \tag{18}$$

Shoaling amplification depends on the ratio of group velocity at the nucleation-site and the coast-site (ocean depths h_0 and $h(\mathbf{r})$, respectively). As does $G(\mathbf{r})$, S_{L} reverts to one in oceans of uniform depth.

Figure 11 pictures a shoaling tsunami wave of 150 s period. Initially, a unit height wave comes ashore from

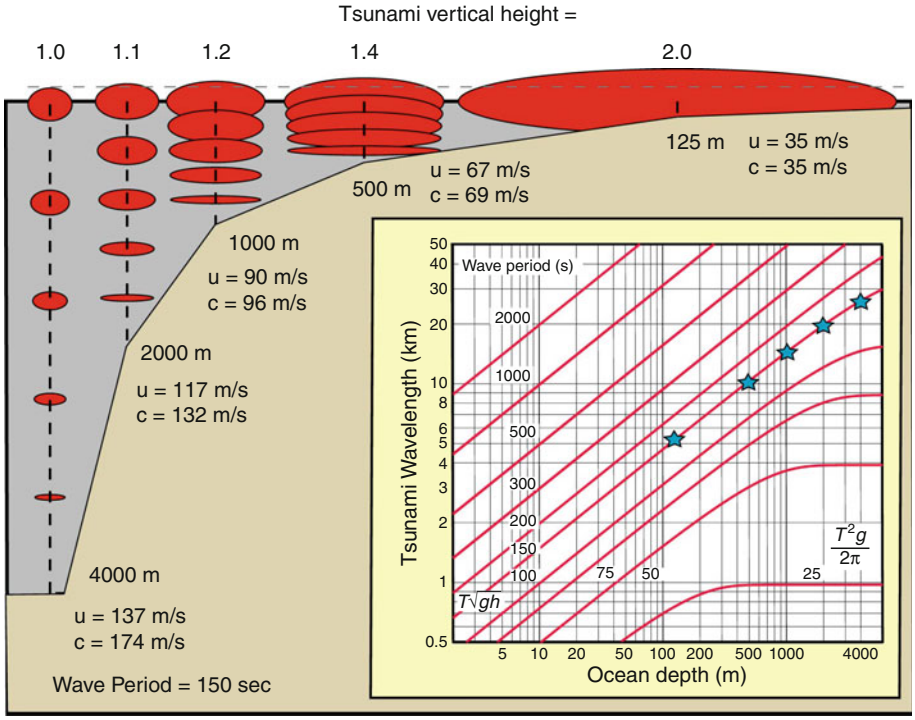


Tsunami, Figure 10 Real tsunami rays (*top left*) versus network rays (*top, right*). I track the curvature of the ray (*right, bottom*) rather than the curvature of the wave front (*left, bottom*) to account for wave size losses as tsunami propagate around obstacles.

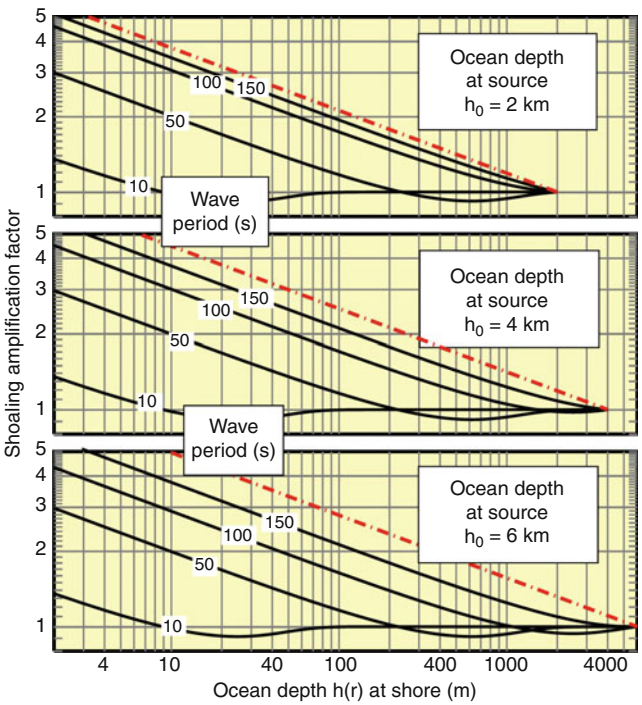
4,000 m of water at the left. As the water shallows, the velocity of the wave decreases and the wave grows in amplitude. By the time it reaches 125 m depth it has slowed from 137 m/s group velocity to 35 m/s and grown in height by a factor of 2. Figure 12 plots (18) as a function of coast-site depth for sea-wave periods from 10 s, and ocean depths of 2, 4, and 6 km. Because their energy does not occupy the entire water column, beach waves at 10 s period do not amplify much (perhaps 50%) in shoaling. Tsunami waves (100–2,000 s period) experience much stronger shoaling amplification – about 3–6 over a wide range of conditions. For waves of period greater than 250 s, $u(\omega,h_0) = (gh_0)^{1/2}$, $u(\omega,h(r)) = (gh(\mathrm{i}))^{1/2}$, and the shoaling factor reduces to Green's Law, $S_L = (h_0/h(r))^{1/4}$ (dashed red line in Figure 12).

Run-Up. Run-up is the final phase of tsunami life. The run-up phase begins when the approaching tsunami shoals to an amplitude roughly equal to the water depth and the wave begins to break. Run-up also covers the inundation phase where the water runs over land and reaches its maximum excursion above sea level. Run-up is the most complicated phase of tsunami life; nevertheless, "Will I get wet?" is the question everyone wants answered, so tsunami scientists have to take their best shot at carrying the waves to the last mile.

One approach to run-up is wave tank or computer experiments like Figures 13 and 14, where waves of various sizes and periods are sent onto model beaches. These experiments reveal a wide range of behaviors depending on wave period, wave size, wave direction, number of waves, beach slope, and beach friction. In real-world situations, this information is not known well, so the direct applications of the experimental results are limited. I follow an alternative approach that takes a broad brush view of run-up, but allows for considerable random uncertainty in the outcome. Using the methods above, I propagate tsunami to a position close to shore in shallow water where it can be considered a long wave, but not so



Tsunami, Figure 11 Effect of shoaling on tsunami eigenfunctions. The shallowing ocean near shore concentrates wave energy into smaller and smaller volumes and tsunami grow in response.



Tsunami, Figure 12 Shoaling amplification factor for ocean waves of various frequencies and source depths. The *dashed red line* is Green's Law.

shallow that its amplitude $A(h)$ exceeds water depth h there. I estimate run-up height η, on shore by:

$$\eta = A(h)^{4/5} h^{1/5} \tag{19}$$

Because $A(h) < h$, run-up always exceeds the offshore wave amplitude. Considering the large uncertainties involved, I view run-up at any field site to be a random variable drawn from a distribution function with (19) being its mean and its spread ranging from 1/2 to 2 times the mean. It does not matter at which depth h is used to evaluate (19), provided the water is shallow enough for the tsunami to act like a long wave. According to Green's Law, a long wave with amplitude $A(h_1)$ in water of depth h_1, has an amplitude $A(h_2) = A(h_1)\,(h_1/h_2)^{1/4}$ in water of depth h_2. Run-up estimate at η_2 at depth h_2

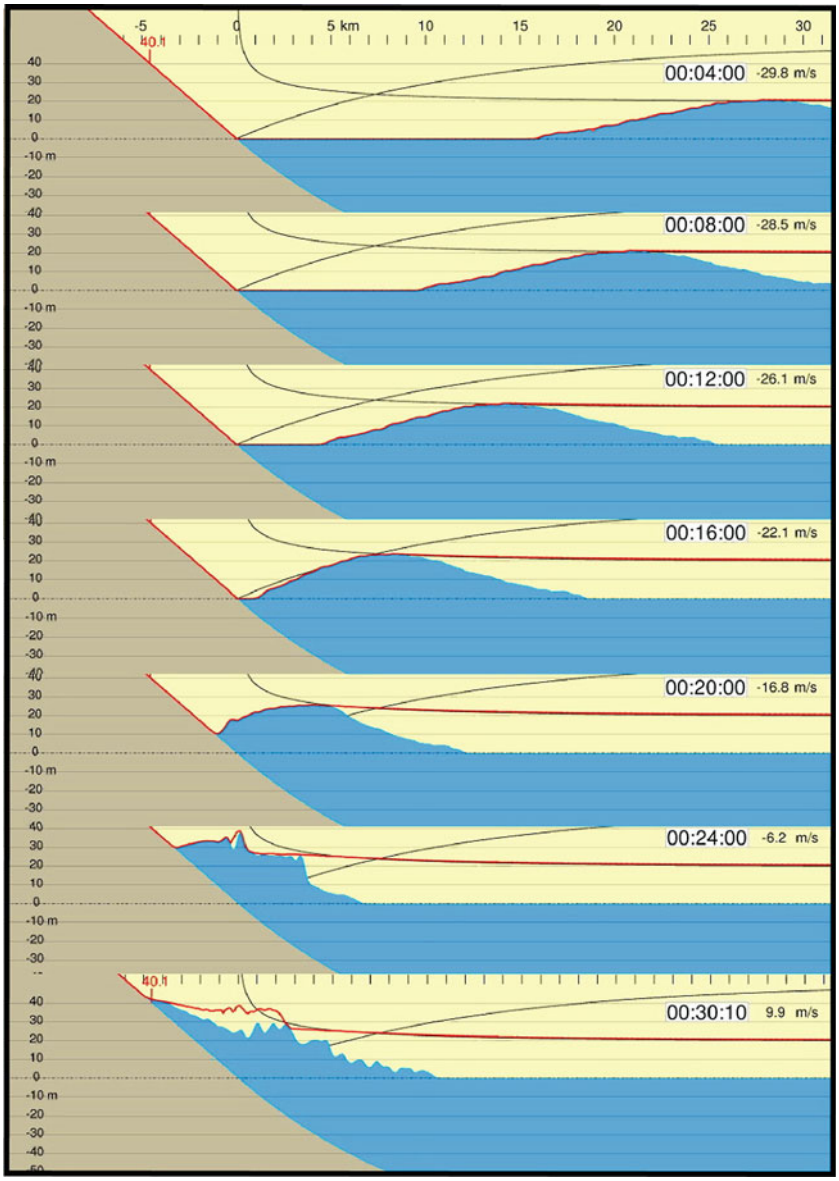
$$\eta_2 = A(h_2)^{4/5} h_2^{1/5} = \left[A(h_1) \left[\frac{h_1}{h_2} \right]^{1/4} \right]^{4/5} h_2^{1/5}$$

$$= A^{4/5}(h_1) {h_1}^{1/5} = \eta_1$$

equals estimate η_1 at depth h_1.

Tsunami samples

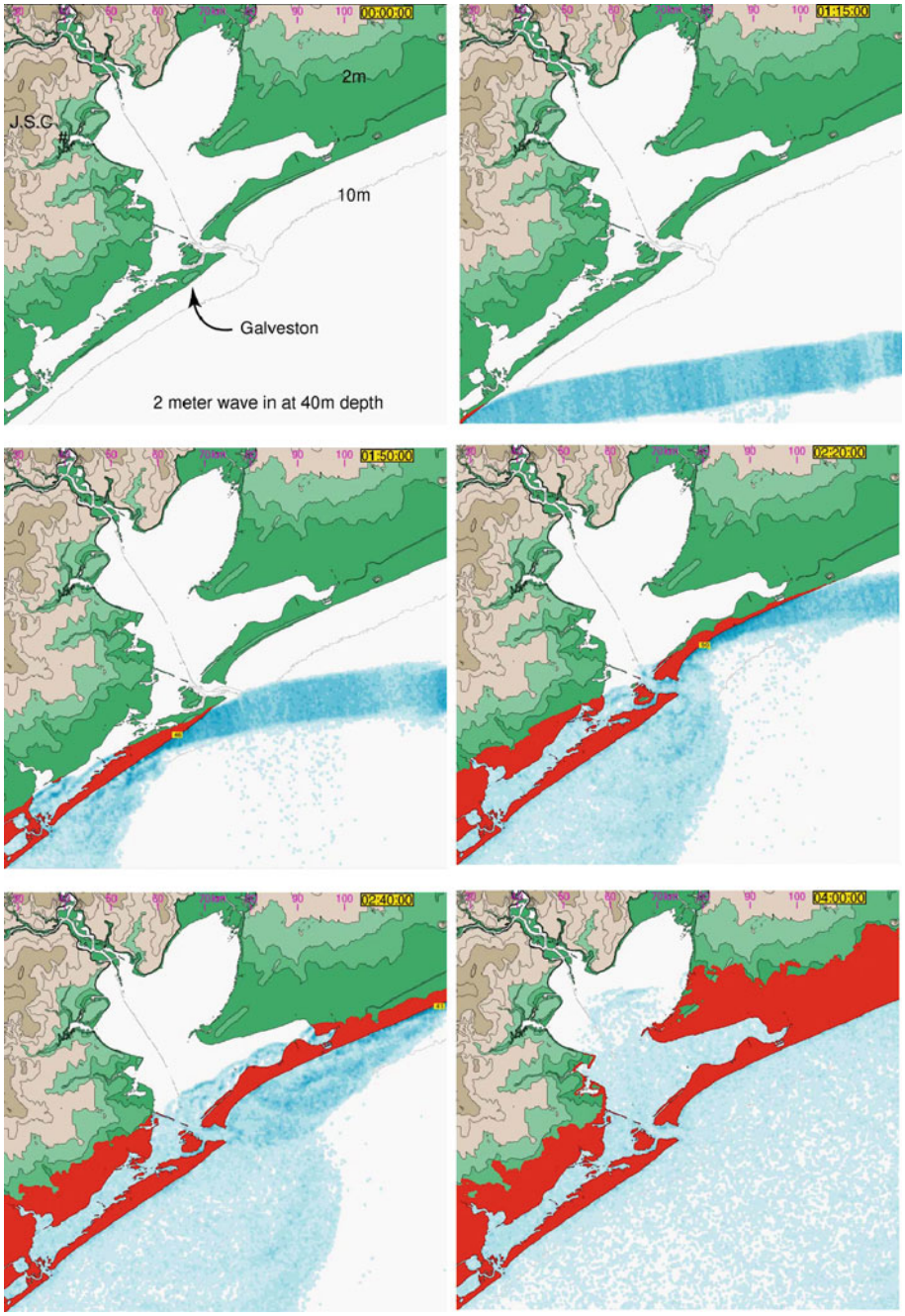
Sumatra earthquake 12/26/2004. Figure 15 snapshots a simulation of the 2004 Sumatra tsunami. As modeled



Tsunami, Figure 13 Run-up and inundation of a 20 m high, 1,200 s duration wave on a steep 0.5° slope. On steep beaches breaking has little time to operate so run-up amplifications are large.

by several rectangular patches, the megathrust fault that sourced the wave trends north–south for several 100 km (Figure 15 *top left*). As the example in Figure 5 predicted, nearly all of the wave energy in the 2004 tsunami emits perpendicular to the fault – Sumatra and Thailand toward the east, and India and Sri Lanka toward the west. Not surprisingly, these locations suffered the lion's share of tsunami damage. I believe that the very large run-up near Banda Aceh was due to concentrated slip on a thrust ramp splaying up from the megathrust just offshore at that latitude.

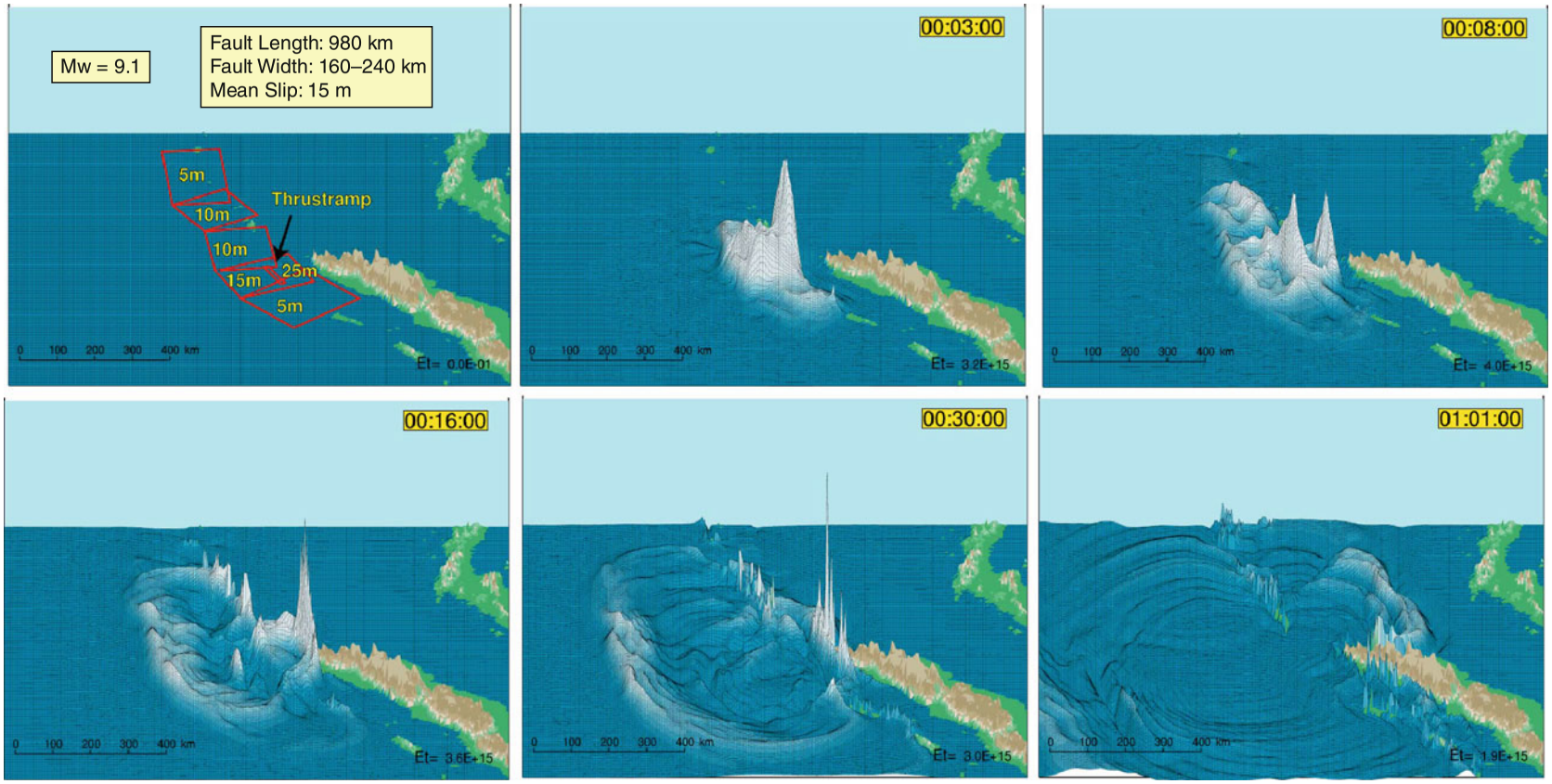
La Palma landslide tsunami. Like its fellows in the Canary Island chain, the volcanic Island of La Palma has experienced cycles of growth and partial collapse. Its flank last failed 550,000 years ago. Since then, new Cumbre Vieja volcano has filled the vacant scar. For the last 125,000 years this volcano grew steadily but in the last 7,000, its structure changed. It now mainly erupts along a north–south trending rift that splits the mountain in half. Moreover, during an eruption in 1949 a fault broke surface along the crest of the volcano and part of its western side slid 5 m down and toward the ocean. The volcano again may be showing initial stages of instability. While certainly collapse is not imminent and it may take many eruptive cycles over the next few thousand years to give it that final shove, the volcano will collapse so it is sensible to consider the consequences (Figure 16). If 500 km^3 of its



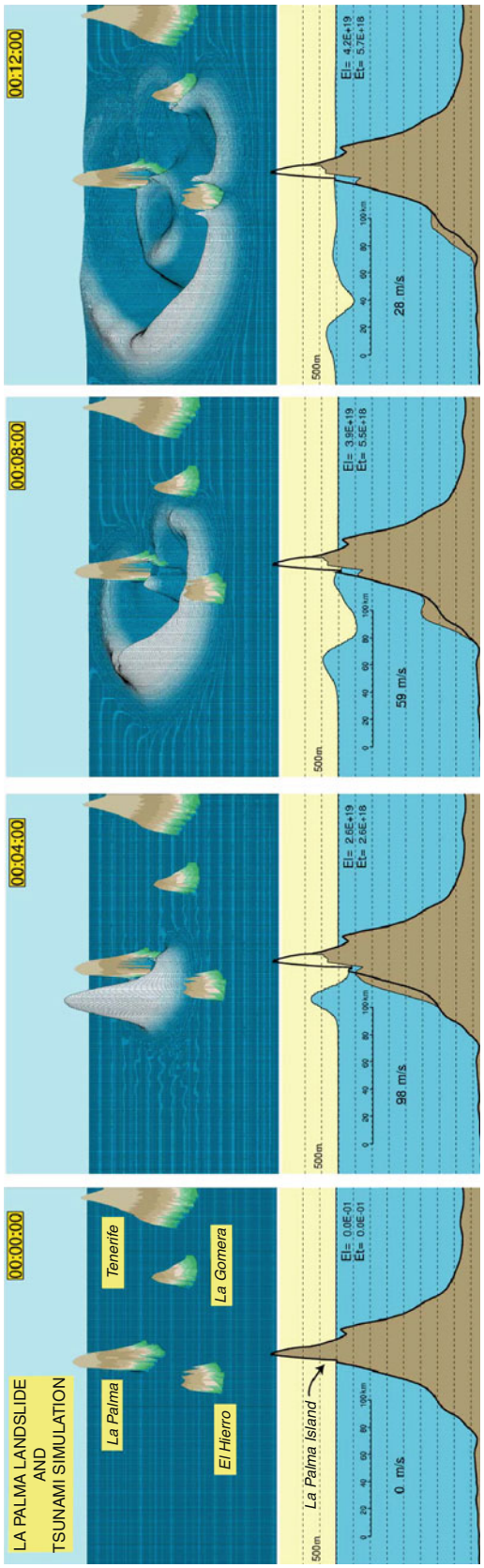
Tsunami, Figure 14 Simulation of 2 m tsunami wave beaching at Galveston Texas. Red area blankets the predicted inundation zone. Note the multiple wave reflections and wave interference that contribute to run-up complexity and randomness.

flank falls into the sea like it did last time, a La Palma collapse will spawn a mega tsunami hundreds of meters high locally that will buffet all coasts of the Atlantic with 10–30 m waves.

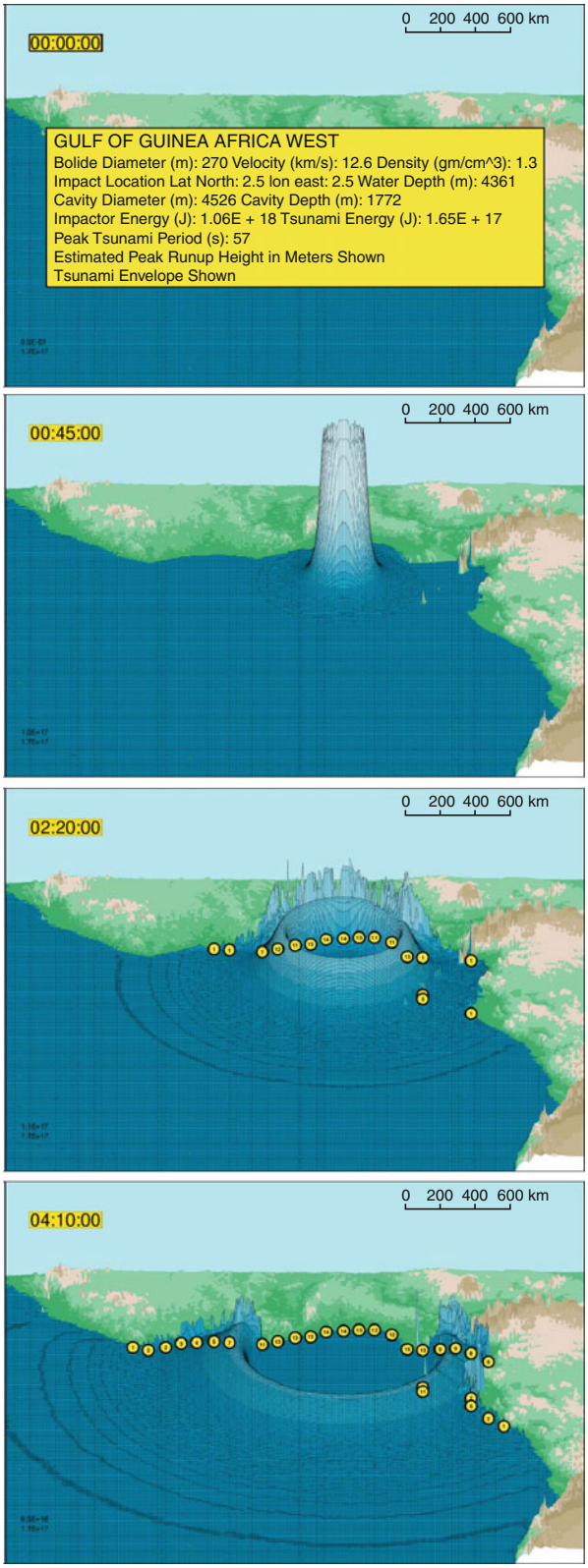
Gulf of Guinea impact tsunami. In a 2008 training exercise, the US Air Force considered the consequences of, and response to, an impact of a small comet into the Atlantic Ocean off the coast of Guinea. In this scenario, pictured in Figure 17, tsunami run-ups exceed 10 m along a 1,200 km stretch of coastline. This wave would knock out coastal shipping and oil production facilities for many months and trigger major ripple effects in the global economy. Unlike those Hollywood movies, currently there is no defense against an asteroid impact.



Tsunami, Figure 15 Simulation of the Sumatra Earthquake Tsunami of December 2004. Note the large wave directed toward the northern tip of Sumatra due to the small thrust ramp offshore there (*upper left*).



Tsunami, Figure 16 (*Top row*) First 12 min of the La Palma Tsunami. Initial wave heights exceed 500 m. (*Bottom row*) Landslide and tsunami in cross section. Just like the simple example in Figure 6, water gets pushed up in front of the slide and drawn down behind.



Tsunami, Figure 17 Guinea Impact Tsunami. Four-hour sequence of waves from the impact of a 270 m diameter icy bolide off the west coast of Africa.

Tsunami forecasting

Official approach: Perhaps the ultimate goal of tsunami research is forecasting. A forecast predicts the size of a particular tsunami given the knowledge that a potentially dangerous earthquake has occurred already. Tsunami travel at jet speeds, but for many places there may be several hours between the earthquake and the arrival of the sea waves. This time can be spent analyzing seismograms, estimating earthquake parameters, and forecasting the expected height of the oncoming wave with the aid of computer models and initial field reports. This is precisely the mission of the various warning centers like NOAA's Pacific Tsunami Warning Center (PTWC). The past decade has witnessed considerable progress in tsunami forecasting coupled with faster and more geographically selective warnings. By and large, the current tsunami prediction system functions well. Figure 18 pictures a nearly real-time prediction of tsunami run-up from the M8.1 Samoa Earthquake of September 29, 2009. Based on rapid calculations like these, the PTWC issued accurate advanced warnings and watches for this tsunami – at least for locations greater than 2,000 km distance from the quake where sufficient delay time existed.

The biggest uncertainty in the current procedures lay in fixing earthquake parameters quickly. For instance, in the first hours after a quake, it is not uncommon for estimates of earthquake magnitude to vary 0.3 or 0.4 units as more extensive seismic data arrive. Because earthquake magnitude or earthquake moment is the prime driver for tsunami (Equation 9), a 0.4 unit uncertainly makes for a factor of 2 or 3 uncertainly in tsunami height all else aside. Partly on this concern, the DART buoy network that provides real-time measurements of passing tsunami was expanded from 6 to 24 sites after the Sumatra earthquake. Information from the DART sensors provides a real-time "reality check" on computer simulations run rapidly on uncertain earthquake parameters.

Do-it-yourself approach: Apart from official predictions, readers of this article have enough information to produce a Do-It-Yourself tsunami forecast. The three step recipe needs: earthquake magnitude M, the ocean depth at source H_o, distance to the quake R, and mean ocean depth $\overline{H}$ between the earthquake and you.

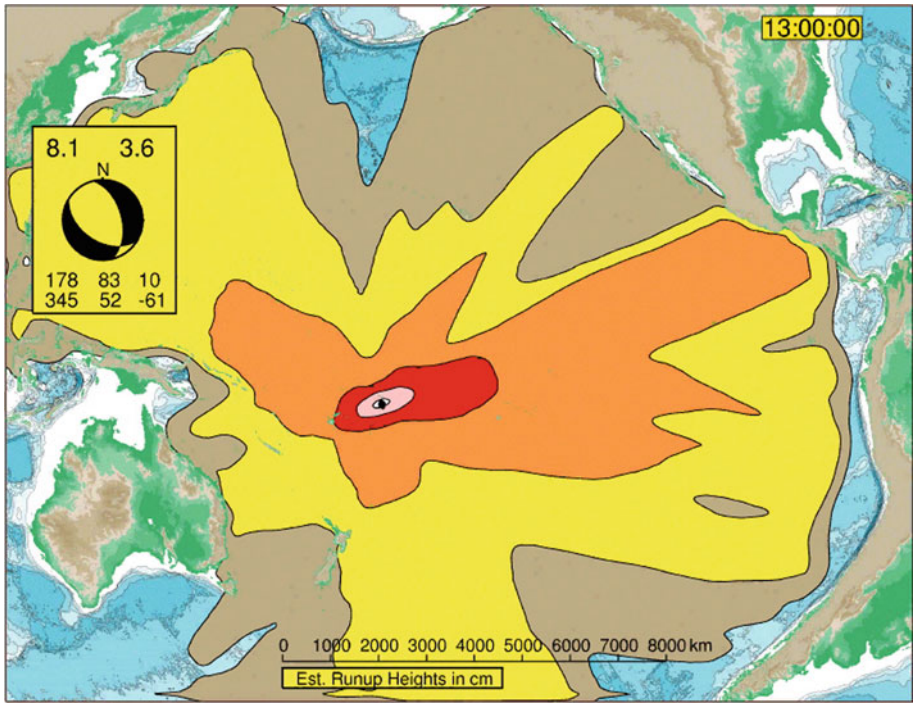
Step 1 – Find Initial Tsunami Amplitude A_0. Let us approximate initial tsunami height by:

$$A_0 = \frac{\alpha \Delta u(M)}{\cosh\left[\frac{4\pi H_0}{W(M)+L(M)}\right]}, \tag{20}$$

where α is the fraction of earthquake slip that transforms into tsunami-making uplift:

$$\alpha = (1 - \theta/180)\sin\theta|\sin\phi|,$$

with θ and ϕ being the fault dip and rake angles in degrees. Fault slip $\Delta u(M)$, width $W(M)$, and length $L(M)$ are functions of magnitude and can be read from Table 1 or computed from the formulas in its caption.



Tsunami, Figure 18 Nearly real-time prediction of run-up height from the 9/29/09 M8.1 Samoa Earthquake. Colors correspond to run-up heights: *Brown* 10–25 cm; *Yellow* 25–50 cm; *Orange* 50–100 cm; *Red* 1–2 m; *Pink* >2 m.

You recognize the 1/cosh as the low-pass filter due to the ocean layer. Unlike Equation 9, Equation 20 makes no assumption about radiation pattern, so in this sense it represents the worst case. The most efficient mechanism (surface breaking, pure dip slip on planes dipping near 45°) for tsunami generation gives $\alpha = 0.55$.

Step 2 – Correct for Propagation Loss *P*.

Tsunami waves decay as they travel due to geometrical spreading and frequency dispersion. For an ocean of constant depth $\overline{H}$, the propagation loss is roughly

$$P = \left(1 + \frac{2R}{L(M)}\right)^{-\psi}, \tag{21}$$

where

$$\psi = 0.5 + 0.575 \exp\left(-0.0175 \frac{L(M)}{\overline{H}}\right). \tag{22}$$

The first term in exponent (22) accounts for square root of distance losses due to geometrical spreading. The second term in (22) accounts for additional losses due to frequency dispersion with larger dimensioned tsunami decaying more slowly. Peak deep-water tsunami amplitude $A_d(R)$ at distance R is

$$A_d(R) = A_0 P = \frac{\alpha \Delta u(M)}{\cosh[4\pi H_0/(W(M) + L(M))]} \left(1 + \frac{2R}{L(M)}\right)^{-\psi} \tag{23}$$

Propagation loss (23) supposes an unobstructed wave path to the observation location. If the waves pass around headlands, squeeze through narrows, or cut into shadow zones, then (23) again overestimates tsunami size.

Step 3 – Correct for Shoaling and Run-Up.

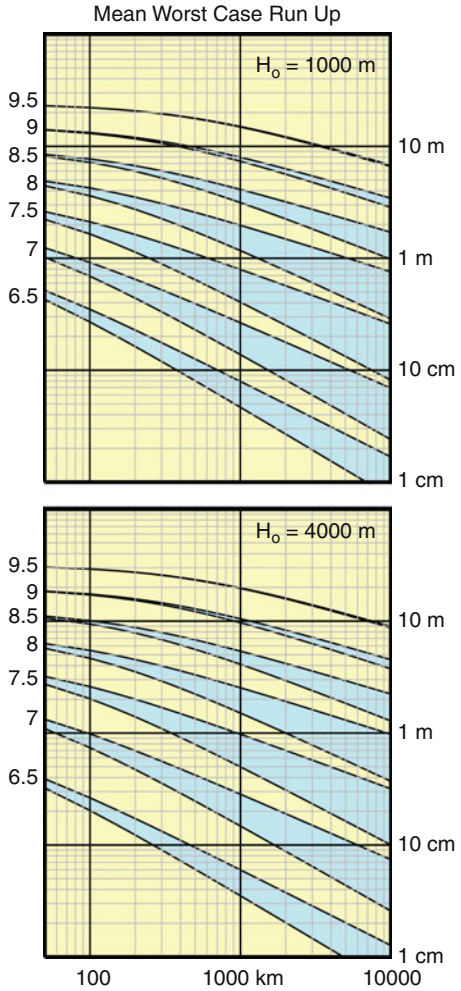
For long waves, the shoaling correction S_L (18) reduces to Green's Law $S_L = (H_0/H)^{1/4}$ and the shoaled tsunami amplitude $A_S(R)$ at distance R is

$$A_S(R) = A_0 P S_L = \frac{\alpha \Delta u(M)}{\cosh\left[\frac{4\pi H_0}{W(M)+L(M)}\right]} \left(1 + \frac{2R}{L(M)}\right)^{-\psi} \left[\frac{H_0}{H}\right]^{1/4}. \tag{24}$$

Applying (19) on (24) returns peak wave run-up height $\eta(R)$

$$\eta(R) = [A_0 P S_L]^{4/5} H^{1/5} = \left[\frac{\alpha \Delta u(M)}{\cosh\left[\frac{4\pi H_0}{W(M)+L(M)}\right]} \left(1 + \frac{2R}{L(M)}\right)^{-\psi}\right]^{4/5} H_0^{1/5}. \tag{25}$$

Figure 19 plots Do-It-Yourself prediction (25) for earthquake magnitudes 6.5–9.5, at distances R to 10,000 km, for $\overline{H}$ between 1,000 and 4,000 m, and $H_0 = 1{,}000$ m (*top*



Tsunami, Figure 19 Mean worst-case tsunami run-up height versus distance from earthquakes of magnitude 6.5–9.5. The blue areas include an allowance for mean ocean depth between 1,000 (*upper limit*) and 4,000 m (*lower limit*). Top and Bottom panels assume 1,000 and 4,000 m water depth at the source.

panel) and H_0 = 4,000 m (*bottom panel*). Because several steps above selected extreme assumptions, these curves represent a mean worst case. The worst worst case might be twice as large still. For $M = 8$, $H_0 = \overline{H} = 4{,}000$ m, Figure 19 (*bottom*) predicts run-ups of 3 m at 300 km distance, 2 m at 600 km distance, 1 m at 2,000 km distance, and 0.5 m at 5,000 km distance. These numbers compare well with the more rigorous calculations in Figure 18.

Summary

Tsunami are gravity-driven water waves caused mostly by undersea earthquakes, but sometimes they are sourced by submarine landslides, surface detonations, explosive volcanoes, and asteroid strikes. Tsunami differ from ordinary beach waves in that they have longer period, larger wavelength, and higher velocity. Because of shoaling effects, these characteristics make tsunami especially damaging because they grow in height proportionally more in approaching shore than do ordinary waves of equal size. This article provides mathematical expressions for tsunami excitation, propagation, and simplified steps to compute wave run-up height from tsunamigenic earthquakes.

Cross-references

Earthquakes and Crustal Deformation
Earthquakes, Source Theory
Seismic, Ray Theory
Surface Waves
Tsunami Watch and Warning Centers
Tsunami: Bay of Bengal

TSUNAMI: BAY OF BENGAL

Vineet Gahalaut
National Geophysical Research Institute (CSIR), Hyderabad, India

Definition

Bay of Bengal: Northeastern part of the Indian Ocean, which is about 2,000 km long and about 1,500 km wide, and is bordered on the west by India and Sri Lanka, on the north by Bangladesh, and on the east by Myanmar and Thailand; the Andaman and Nicobar Islands separate it from the Andaman Sea.
Tsunami: Gravity-driven water waves in the sea caused by submarine earthquakes, landslides, volcanic eruption, and impact meteorites.

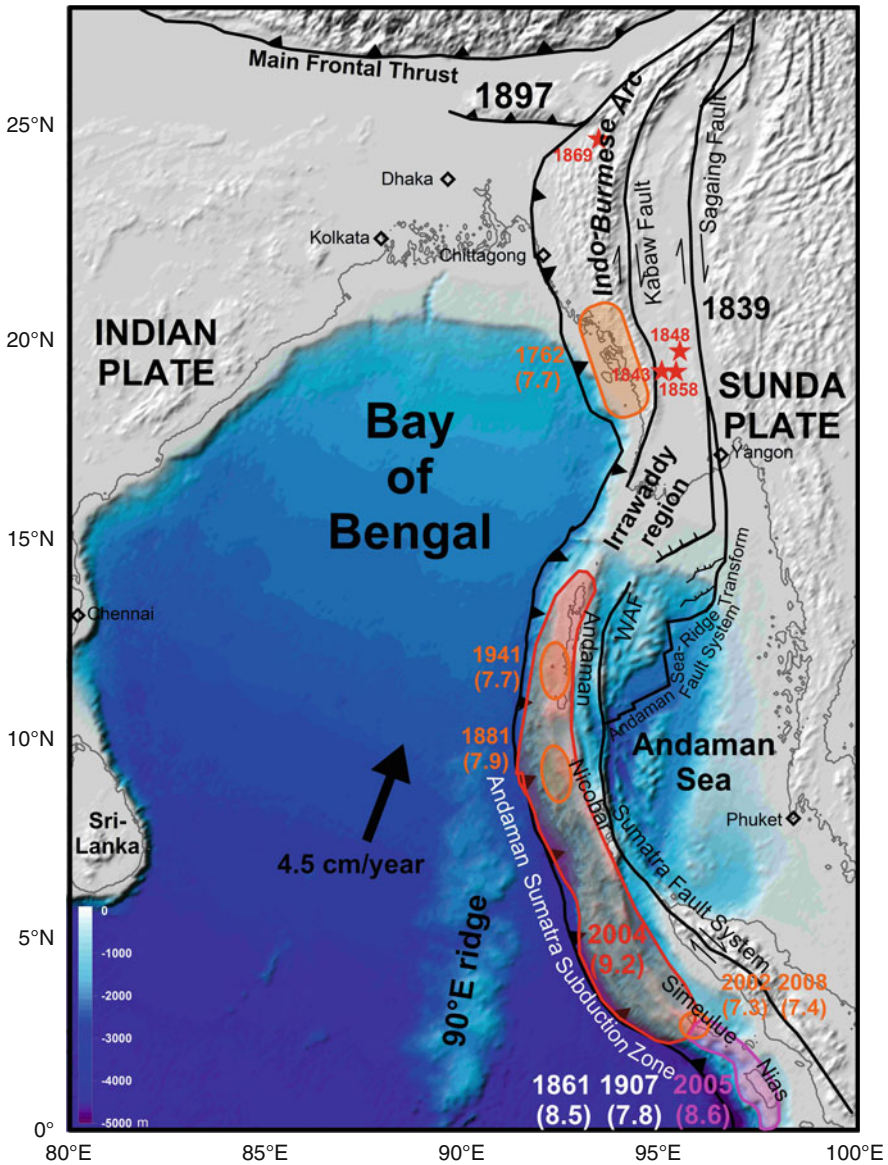
Introduction

The coastal region around the Bay of Bengal is one of the most populated regions of the world with population density of more than 1,000 people/km^2 in Bangladesh and region around it. Natural hazards in this region, namely, floods and cyclones, have turned into disasters making this region very vulnerable. The occurrence of the December 26, 2004, Sumatra-Andaman earthquake (Mw 9.2) and the tsunami caused by it (Lay et al., 2005), which devastated the coastal regions around the Bay of Bengal and killed more than 200,000 people, further exposed the vulnerability of this region due to earthquakes and tsunami. For an objective assessment of the tsunami hazard in this region, it is necessary to understand the tectonics and the mechanism of earthquake occurrence in this region. Other than earthquakes, submarine landslide in the several-kilometer-thick sediments of the Ganga and Brahmaputra rivers deposited in the northern Bay of Bengal may also pose tsunami hazard in the region. There are no historical records to suggest occurrence of such submarine landslides and no studies have been carried out to assess the tsunami hazard due to this. However, some

attempts have been made to assess tsunami hazard due to earthquakes in the region (Okal and Synolakis, 2008; Cummins, 2007).

Tsunami hazard in the Bay of Bengal

The tsunami hazard in the Bay of Bengal region is mainly due to the great and major earthquakes along the northern Sunda arc where the eastern margin of the Indian plate interacts with the Sunda plate (Figure 1). Tsunami caused by great earthquakes in the far-field regions, namely, the Banda arc which is to the east of Sunda arc (i.e., southeast of the equator), the Makran region which is along the southern coast of Pakistan and Iran, and the circum pacific belt, may not affect the Bay of Bengal region due to directivity effect and the presence of intervening land masses during their propagation. The frontal arc of the northern Sunda arc consists of the Indo-Burmese arc in the north, and the Sumatra- Andaman arc in the south with the



Tsunami: Bay of Bengal, Figure 1 Bay of Bengal, general tectonics of the northern Sunda arc and locations of major and great earthquakes along the arc. Stars denote major earthquakes that occurred in the Indo-Burmese arc. These earthquakes were not associated with any tsunami. The great 1897 Shillong Plateau earthquake and the great 1839 earthquake on Sagaing fault are also shown for reference. The 1897 intraplate earthquake is associated with the tectonics of the Indian plate and its interaction with the Himalaya, rather than the tectonics of the Indo-Burmese arc. Major earthquakes of Sagaing fault are not shown here, as they do not pose any tsunami hazard. The 1861 and 1907 earthquakes occurred in the source region of the 2005 Nias earthquake. However, none of the tsunami caused by of these earthquakes, including that of the 2005 Nias earthquake, affected the Bay of Bengal region.

intervening Irrawaddy delta region. In the back arc, the Sagaing fault which runs through Myanmar on the land and is characterized by strike slip motion, the Sumatra Fault system which is characterized by the earthquakes with strike slip motion, and the Andaman Sea ridge transform fault system which is characterized by the normal and strike slip earthquakes of low magnitude ($M < 6.5$), do not pose tsunami hazard to the Bay of Bengal region. The tsunami hazard in the Bay of Bengal due to low-level seismicity in the diffused plate boundary region between India and Australia in the Indian Ocean may be very low.

Plate motion along the northern Sunda arc

In the northern Sunda arc, relative motion between the India and Sunda plates is about 3.7 cm/year toward N15° in the northern part (i.e., Indo-Burmese arc region) and about 4.5 cm/year toward N25° in the southern part (i.e., the Sumatra region). The oblique motion between the India and Sunda plates is accommodated along the arc through slip partitioning in the frontal and back arc (Gahalaut and Gahalaut, 2007; McCaffrey 2009). In the north, the north–south trending Indo-Burmese arc accommodates slip of about 1.6 cm/year through dextral motion (Gahalaut et al., 2010a) and the remaining motion of about 2.0 cm/year is accommodated across the dextral strike slip Sagaing fault in the Myanmar (Maurin et al., 2010). Thus, at present, the underlying gently east dipping Indian plate beneath the accretionary wedge of the Indo-Burmese arc moves toward north. Further south in the Irrawaddy and Andaman region, the motion in the frontal arc becomes oblique. In the back arc, the dextral strike slip motion occurs along the north–south planes and normal motion occurs along the Andaman Sea ridge transform fault system which is consistent with the Andaman sea opening. In the Sumatra region, the motion in the frontal arc is arc normal, leading to thrust motion on the plate boundary interface. The strike slip motion in the back arc is accommodated through dextral motion along the Sumatra Fault System. The Sumatra Fault System joins the Sagaing fault in the north through the Andaman Sea ridge transform fault system.

Seismicity of the northern Sunda arc region and tsunami hazard

For the convenience of discussion, the seismicity of the Indo-Burmese arc, Irrawaddy, and Sumatra-Andaman region is described separately.

Indo-Burmese arc

The plate reconstruction models suggest that subduction probably occurred along the Indo-Burmese arc in the geological past (prior to about 50 Ma) when the arc was predominantly southeast–northwest trending. However, after the collision of the Indian plate with the Eurasian plate, the northern Sunda arc rotated clockwise to become predominantly north–south trending. Several geological and geophysical studies confirm that the subduction occurred across the Indo-Burmese arc between India and Burma plates and there is evidence of the subducted Indian slab, but presently the subduction does not appear to be active (Hall, 1997; Guzman-Speziale and Ni, 1996; Rao and Kumar, 1999; Pesicek et al., 2010). This is further confirmed from the earthquake occurrence and the crustal deformation studies.

The historical records of earthquakes and tsunami in the Indo-Burmese arc are generally very poor. Although there are a few unverifiable reports of earthquake occurrence as early as 1548 in Tripura, Assam (two states in northeastern India), or Bangladesh, there are large uncertainties in the location of historical earthquakes. The catalogue is probably reliable only after 1762. The most notable earthquakes are the April 2, 1762, the August 24, 1858, Arakan earthquakes, and the January 10, 1869, Cachar earthquake (Richter, 1958). Other than these earthquakes, several instances of damage due to earthquakes are reported from Chittagong, Sylhet, Manipur valley, and Cachar regions. However, it is possible that the high damage in these sediment-filled valleys was due to relatively large population and local site effects. The April 2, 1762, Arakan earthquake has been considered as a great tsunamigenic earthquake (Cummins, 2007). It caused extensive damage in the Chittagong region through shaking, liquefaction, damming of channels, seiches, etc. (Gulston, 1763). The reported uplift of Cheduba Island due to this earthquake was not found to be singularly due to this earthquake (Oldham, 1883; Martin and Szeliga, 2010). Halsted (1843) reported the effects of the earthquakes which were mostly found to be highly exaggerated by Gupta and Gahalaut (2009). They found that the highly qualitative observations of the coseismic elevation changes are consistent with a predominant strike slip motion on the earthquake rupture which was located under the land and/or under very shallow water, and it was only a major earthquake (Martin and Szeliga, 2010) which probably did not cause any major tsunami (Gupta and Gahalaut, 2009). The August 24, 1858, Arakan earthquake was felt in many parts of Burma and was severely felt at Kyauk Pyu, Ramree Island, Myanmar. It caused liquefaction and damage to buildings and Pagodas. In this region, the distance between the Sagaing fault and the structurally mapped Arakan trench is less than 200 km, and, hence, based on the scanty reports of damage, it is difficult to judge whether this one and the 1843 and 1848 earthquakes are linked with the Indo-Burmese arc or with the Sagaing fault. But none of these three earthquakes is associated with tsunami. The January 10, 1869, Cachar-Manipur earthquake is referred to as the most severe earthquake in the available 2,000 years of written historical records of Manipur state of India. However, recent examinations of the historical records suggest that even this earthquake was not a great one and the damage was very much confined to the sediment-filled valleys of Imphal and Cachar (Kundu and Gahalaut, 2010b). As this earthquake occurred

on the land, far from the sea, there is no question of tsunami being generated by this earthquake.

Although precise motion on rupture during these major earthquakes is not known, the current seismicity and the results from crustal deformation studies suggest that the historical earthquakes in the Indo-Burmese arc might have occurred through predominant strike slip motion on steep planes, which dips away from the coast. This severely limits the possibility of a tsunamigenic earthquake in the coastal Indo-Burmese arc region. Lack of paleo-tsunami deposits in the coastal Myanmar region (Aung et al., 2008) further supports this argument.

Irrawaddy region

Seismicity in the Irrawaddy frontal arc is extremely low as compared to that in the Andaman Nicobar arc in the south and the Indo-Burmese arc in the north. Earthquakes in the Irrawaddy region occur at depths less than 50 km, whereas in the Indo-Burmese arc and the Andaman Nicobar region, earthquakes occur down to a depth of up to 150–200 km (Guzman-Speziale and Ni, 1996). From the limited earthquake focal mechanisms, it is not possible to constrain sense of motion along this part; however, the regional plate motion models suggest that very oblique motion occurs along the frontal arc (Gahalaut and Gahalaut, 2007). The reason for very low seismicity of the frontal arc at shallow depth (>50 km) could be that the region is either aseismic or accumulating strain. However, absence of seismicity even at depths greater than 50 km suggests that the subducting slab might not be continuing to that depth at all. In the Andaman and Indo-Burmese arc regions, occurrence of earthquakes at intermediate depths confirms the continuity of the Indian plate slab in those regions. The absence of seismicity at intermediate depth level of the Irrawaddy region is possibly due to the presence of tear in the subducting Indian slab, which is has been confirmed from the tomographic studies of the region (Richards et al., 2007). Thus, it is the tear in the slab which is responsible for very low seismicity at shallow depth and no seismicity at deeper depth (Kundu and Gahalaut, 2010a). Presence of a tear and lack of slab in the region may substantially reduce the potential of occurrence of a future great earthquake in the region. Accordingly, it may also reduce tsunami hazard in the nearby regions due to major and great earthquakes in this region. It is possible that this tear hindered the northward rupture propagation of the 2004 Sumatra-Andaman earthquake. However, it is necessary to further confirm the presence of such a tear using marine deep seismic surveys, detailed tomographic studies, and analysis of earthquakes with the deployment of ocean bottom and land seismometers in the region.

Sumatra-Andaman arc region

The state of knowledge about the Sumatra-Andaman region was very poor until the occurrence of the December 26, 2004, Sumatra-Andaman (M 9.2), March 28, 2005, Nias (M 8.6), and several other earthquakes since then. These earthquakes have provided immense information about the tectonics and seismogenesis of this region. These earthquakes suggest that Indian plate subducts obliquely beneath the Andaman arc while the subduction becomes arc normal in the Sumatra arc. Precise hypocentral locations of these earthquakes and their aftershocks have helped in mapping the subsurface plate boundary interface in the frontal arc region and its seismogenic width. The 2004 Sumatra-Andaman earthquake occurred on a 1,400 km long and 100–150 km wide rupture between the northern Sumatra and north Andaman Islands. The large coseismic displacement in the source region (Gahalaut et al., 2006) caused a devastating tsunami which propagated through the Indian Ocean and caused heavy destruction in the countries surrounding the Bay of Bengal, namely, Indonesia, India, Thailand, and Sri Lanka. There were a few remarkable aspects of this earthquake and tsunami caused by it. It is learnt that the rupture in the northern part, i.e., under the Andaman Islands, was slower (Lay et al., 2005) and that it did not reach the surface. Thus this part of the rupture did not contribute in tsunami generation. It has been suggested that the subducting 90°E ridge probably slowed down the rupture under the Andaman Islands (Gahalaut et al., 2010b) and the subducted unconsolidated sediments of Bay of Bengal probably did not allow the rupture to propagate to the surface. These aspects need to be explored further. Earlier, another earthquake occurred here on June 26, 1941, with M 7.7 (Pacheco and Sykes, 1992), for which no seismological data exist to suggest whether slow rupture occurred during that earthquake as well. Though the reports are scanty, it appears that even that earthquake did not generate a tsunami (Rajendran et al., 2007). There may be various reasons for that earthquake for not being tsunamigenic, including slow rupture speed and rupture on blind fault. It may be possible that the Andaman segment may not generate major tsunami during future great earthquakes as well. However, it needs to be explored further by analyzing the frontal arc characteristics through deep seismic imaging and modeling.

An earthquake occurred on December 31, 1881 (M 7.9), in the Nicobar region. It caused a tsunami with less than 1 m height in the Andaman Islands and the coastal region of India. Unfortunately, the history of the Andaman and Nicobar Islands is very poor. The intermittent colonial occupation of the Islands started in the middle of the eighteenth century by the Danish, followed by the British, and continuous colonial occupation started only after 1869 until the independence of India in 1947. Prior to the nineteenth century, we do not know much about the history of the Andaman Nicobar region as the indigenous tribes did not maintain any written records. Thus, prior to the occurrence of earthquakes on October 31, 1847, possibly near Nicobar Islands (Bilham et al., 2005), and the two earthquakes mentioned above, we have no knowledge of historical earthquakes or tsunami in the region. However, it is certain that the great and major earthquakes under the Nicobar and northern Sumatra arc

have produced tsunami and are capable of generating tsunami during future great and major earthquakes.

In the northern Sumatra region, after the occurrence of the 2004 Sumatra-Andaman earthquake, another great earthquake occurred to the south on March 28, 2005, which is referred as the Nias earthquake. Despite its large magnitude and normal rupture speed, the earthquake could not generate a major tsunami. Majority of the slip on the rupture occurred under the shallow and no water and hence it could not displace large amount of water, causing only a localized tsunami whose amplitude faltered quickly as it moved in the open ocean. Nevertheless, this region has the potential to generate a tsunami during great earthquakes, though it may not affect the Bay of Bengal region due to rupture directivity. The arc region located further southeast, though capable of causing tsunami during great earthquakes, may not affect the Bay of Bengal region.

Summary

The Arakan coast of the Indo-Burmese arc can produce a major earthquake. However, it is unlikely that a major tsunami would be generated by the earthquake in this segment due to predominant strike slip motion on the plate boundary interface which is mostly located on the land and under very shallow water, and dips eastward, away from the sea. The Irrawaddy frontal arc region may not produce great or even major earthquakes and hence the tsunami threat from this region is low. The Andaman arc which did not contribute in tsunami generation during the 2004 Sumatra-Andaman earthquake needs to be studied in great detail in view of the slow and blind rupture. The real and definite threat is from the Nicobar and northern Sumatra arc region, which has produced tsunami in the past and is capable of generating tsunami during future great earthquakes.

Bibliography

Aung, T. T., Okamura, Y., Satake, K., Swe, W., Swe, T. L., Saw, H., and Tun, S. T., 2006. Paleoseismological field survey along the western coast of Myanmar. *Annual Report on Active Fault and PaleoEarthquake Researches*, Active Fault Research Center, Geological Survey of Japan, **6**, 171–188.

Aung, T. T., Satake, K., Okamura, Y., Shishikura, M., Swe, W., Saw, H., Swe, T. L., Tun, S. T., and Aung, T., 2008. *Geologic evidence for great Holocene earthquakes off Myanmar*. Phuket, Thailand: International Symposium on the Restoration Program from Giant Earthquakes and Tsunamis, pp. 13–18.

Bilham, R. G., Engdahl, E. R., Feldl, N., and Satyabala, S. P., 2005. Partial and complete rupture of the Indo-Andaman plate boundary 1847–2004. *Seismological Research Letters*, **76**, 299–311.

Cummins, P. R., 2007. The potential for giant tsunamigenic earthquakes in the northern Bay of Bengal. *Nature*, **449**, 75–78.

Gahalaut, V. K., and Gahalaut, K., 2007. Burma plate motion. *Journal of Geophysical Research*, **112**, B10402, doi:10.1029/2007JB004928.

Gahalaut, V. K., Nagrajan, B., Catherine, J. K., and Kumar, S., 2006. Constraints on 2004 Sumatra–Andaman earthquake rupture from GPS measurements in Andaman-Nicobar Islands. *Earth and Planetary Science Letters*, **242**, 365–374, doi:10.10116/j.epsl.2005.11.051.

Gahalaut, V. K., Kundu, B., Singh, S., Kumar, A., Tiwari, R. P., Catherine, J. K., and Ambikapathy, A., 2010a. GPS measurements in the Indo-Burmese arc, (in press).

Gahalaut, V. K., Subrahmanyam, C., Kundu, B., Catherine, J. K., and Ambikapathy, A., 2010b. Slow rupture in Andaman during 2004 Sumatra-Andaman earthquake: consequence of subduction of 90°E ridge. *Geophysical Journal International*, **180**, 1181–1186.

Gulston, E., 1763. An account of an earthquake at Chattigaon: translated from the Persian by Mr. Edward Gulston, in the service of the Honourable East India Company, and communicated by him to the Reverend Mr. Hirst. *Philosophical Transactions*, **53**, 251–256.

Gupta, H., and Gahalaut, V. K., 2009. Is the northern Bay of Bengal tsunamigenic. *Bulletin of the Seismological Society of America*, **99**, 3496–3501.

Guzman-Speziale, M., and Ni, J. F., 1996. Seismicity and active tectonics of the western Sunda Arc. In Yin, A., and Harrison, T. M. (eds.), *The Tectonic Evolution of Asia*. New York: Cambridge University Press, pp. 63–84.

Hall, R., 1997. Cenozoic plate tectonic reconstructions of SE Asia, In Fraser, A. J., Methews, S. J., and Murphy, R. W. (eds.), *Petroleum Geology of Southeast Asia*, Geological Society of London Special publication No 126, pp. 11–23.

Halsted, E. P., 1843. Report on the Island of Chedooba. *Journal of Asiatic Society of Bengal*, **114**, 319–446.

Kundu, B., and Gahalaut, V. K., 2010a. An investigation into the seismic potential of the Irrawaddy region, northern Sunda Arc. *Bulletin of the Seismological Society of America*, **100**, 891–895.

Kundu, B., and Gahalaut, V. K., 2010b. Earthquake occurrence processes in the Indo-Burmese arc and Sagaing fault region. *Tectonics*, (in press).

Lay, T., Kanamori, H., Ammon, C. J., Nettles, M., Ward, S. N., et al., 2005. The great Sumatra-Andaman earthquake of December 26, 2004. *Science*, **308**, 1127–1133.

Martin, S., and Szeliga, W., 2010. A catalog of felt intensity data for 570 earthquakes in India from 1636 to 2009. *Bulletin Seismological Society of America*, **100**, 562–569.

Maurin, T., Masson, F., Claude Rangin, U., Min, T., and Collard, P., 2010. First global positioning system results in northern Myanmar: constant and localized slip rate along the Sagaing fault. *Geology*, **38**, 591–594.

McCaffrey, R., 2009. The tectonic framework of the Sumatran subduction zone. *Annual Review of Earth Planetary Sciences*, **37**, 345–366.

Okal, E. A., and Synolakis, C. E., 2008. Far-field tsunami hazard from mega-thrust earthquakes in the Indian Ocean. *Geophysical Journal International*, **172**, 995–1015.

Oldham, T., 1883. A catalogue of Indian earthquakes. *Memoirs of Geological Survey of India*, **19**, 163–215.

Ortiz, M., and Bilham, R., 2003. Source area and rupture parameters of the 31 December 1881 Mw = 7.9 Car Nicobar earthquake estimated from tsunamis recorded in the Bay of Bengal. *Journal of Geophysical Research*, **108**, 2215.

Pacheco, J. F., and Sykes, L. R., 1992. Seismic moment catalogue of large shallow earthquakes, 1900 to 1989. *Bulletin of the Seismological Society of America*, **82**, 1306–1349.

Pesicek, J. D., Thurber, C. H., Widiyantooro, S., Zhang, H., DeShon, H. R., and Engdahl, E. R., 2010. Sharpening the tomographic image of the subducting slab below Sumatra, the Andaman Islands and Burma. *Geophysical Journal International*, **182**, 433–453.

Rajendran, C. P., Rajendran, K., Anu, R., Earnest, A., Machado, T., Mohan, P. M., and Freymueller, J., 2007. Crustal deformation and seismic history associated with the 2004 Indian Ocean earthquake: a perspective from the Andaman–Nicobar Islands. *Bulletin of the Seismological Society of America*, **97**, S174–S191.

Rao, N. P., and Kumar, M. R., 1999. Evidences for cessation of Indian plate subduction in the Burmese arc region. *Geophysical Research Letters*, **26**, 3149–3152.
Richards, S., Lister, G., and Kennett, B., 2007. A slab in depth: three dimensional geometry and evolution of the Indo-Australian plate. *Geochemistry, Geophysics, Geosystems*, **12**, Q12003, doi:10.1029/2007GC001657.
Richter, C. F., 1958. *Elementary Seismology*. San Francisco: Freeman, p. 768.

Cross-references

Earthquakes and Crustal Deformation
GPS, Data Acquisition and Analysis
Plate Tectonics, Precambrian
Tsunami

TSUNAMI WATCH AND WARNING CENTERS

Shailesh R. Nayak[1], Srinivasa Kumar Tummala[2]
[1]Ministry of Earth Sciences, Government of India, New Delhi, India
[2]Indian National Centre for Ocean Information Services, Hyderabad, India

Synonyms

Maremoto

Definition

Tsunami. Japanese word literally meaning harbor wave (*tsu+nami*). Tsunamis are series of ocean gravity waves caused mainly due to undersea earthquakes and undersea volcanic activity, rarely due to submarine landslides and very infrequently due to meteor impacts.
Tsunamigenic earthquake. Large earthquakes occurring mostly in subduction zones capable of generating tsunami.
Tsunami warning center (TWC). A 24 × 7 operational center with a mission to advise and warn coastal populace on tsunami.
Tsunami Bulletin. A real-time situation report of the event.
National tsunami warning center (NTWC). An early warning system responsible for collating and disseminating tsunami warning advisories information within a country.
Regional tsunami watch provider (RTWP). A NTWC capable of detecting and analyzing tsunami events and informing other NTWC within an oceanic region.
Service level. Level of operational capability in accurately advising/warning target populace.

Introduction

The 2004 Sumatra-Andaman earthquake and subsequent tsunami in the Indian Ocean were a grim reminder of nature's fury. The event prompted many countries in the region to set up tsunami warning centers (TWC), which were further integrated into a collaborative network of regional tsunami watch providers (RTWP) (UNESCO-IOC, 2007). Within India, an interim warning center was established in 2005 (Gupta, 2005; and Nayak and Kumar, 2008a) at Indian National Center for Ocean Information Services (INCOIS) and by the end of 2007, it became a full-fledged operational warning center.

An important fact that is well known about tsunami is that they travel very fast in open ocean (velocity of the order of several hundreds of kilometers per hour) and cannot be discernible easily by wave heights (of the order of a few centimeters), but slow down as they approach the coast (of the order of several tens of kilometers per hour) with wave heights increasing to disastrous proportions (of the order of several meters). Hence, to forewarn ourselves, we need to have a seismic network to detect tsunamigenic earthquakes and a network of sea-level observations to detect any water level changes. A forewarning system is essential to mitigate the effects of tsunami. Few centuries ago, traditional knowledge indicated that as soon the water withdraws from the beach, there is an impending tsunami. However, with the advancement of science (especially modeling, geospatial technologies, and communication infrastructure) and rapid increase in coastal population, the inherent risk in terms of loss of lives and economic damages (socioeconomic parameters) too has increased many fold, but now it is possible to mitigate damage by scientifically confirming/canceling such warnings derived from traditional knowledge and recommending the right advice to the populace at the right time. Such a scientific establishment is known as TWC. The mission of a TWC is to provide early warnings (known as tsunami advisories) on potentially destructive tsunamis to emergency/disaster management officials, and as appropriate, directly to the public to prevent loss of life and property. This article briefly describes TWC, its components, the advisory products that are generated, and level of operational service that these centers are designed for.

Types of tsunamis

Broadly, the Intergovernmental Oceanographic Commission (IOC) of United Nations Educational, Scientific, and Cultural Organization (UNESCO) has classified tsunamis into nine types: Historical tsunami, Local tsunami, Maremoto, Meteorological tsunami (meteotsunami), Microtsunami, Ocean-wide tsunami, Paleotsunami, Regional tsunami, and Teletsunami (Distant tsunami). However, from the early warning perspective, three types are considered most relevant, based on the distance from the source and the tsunami travel time (UNESCO-IOC, 2006). They are:

1. *Local Tsunami* originates from a nearby source for which its destructive effects are confined to coasts within 100 km or less than 1 h tsunami travel time from its source. These can often be the most dangerous because there is often little warning between the causing event and the arrival of the tsunami.

2. *Regional Tsunami* is capable of creating destruction in a particular geographic region, generally within 1,000 km or 1–3 h tsunami travel time from its source.
3. *Distant Tsunami* (also called an ocean-wide, distant, tele or far-field tsunami) is a tsunami that originates from a far away source, which is generally more than 1,000 km away from the area of interest or more than 3 h tsunami travel time from its source. A distant tsunami is capable of causing widespread destruction, not only in the immediate region of its generation, but across an entire ocean. All ocean-wide tsunamis have been generated by major earthquakes in the subduction regions.

Tsunami warning centers in the world

There are seven major tsunami warning centers operating in the world today in terms of independent services. They are: National Oceanic and Atmospheric Administration's (NOAA) Pacific Tsunami Warning Center (PTWC), Japan Meteorological Agency (JMA), NOAA's West Coast and Alaska Tsunami Warning Center, Sakhalin Tsunami Warning Center (STWC), German Indonesian Tsunami Early Warning System (GITEWS), Joint Australian Tsunami Warning Centre (JATWC), and the Indian National Tsunami Early Warning System (INTEWS). Each of these systems is conceptualized as the National Tsunami Warning Centres (NTWC), where they are the sole advisors in terms of area of responsibility regarding tsunami warning. In the event of distant tsunamis, they will act as regional tsunami watch providers (RTWP) to those regions that could be immediately affected and do not have their own NTWC or to other NTWCs operating in the region, as "heads-up" information. In addition to the above, few systems have been conceived as system of systems like International Tsunami Information Center (ITIC) and Indian Ocean Tsunami Warning System (IOTWS), where information from different major tsunami advisory providers is integrated as shown in Figure 1. The latter, IOTWS, has been an initiative from the US and UNESCO to enhance the warning capability in the region. In addition, critical seismic data providers like USGS (Earthworm seismic data processing system) and Global Seismic Network (GSN) are few of the specialized information providers essential for tsunami early warning.

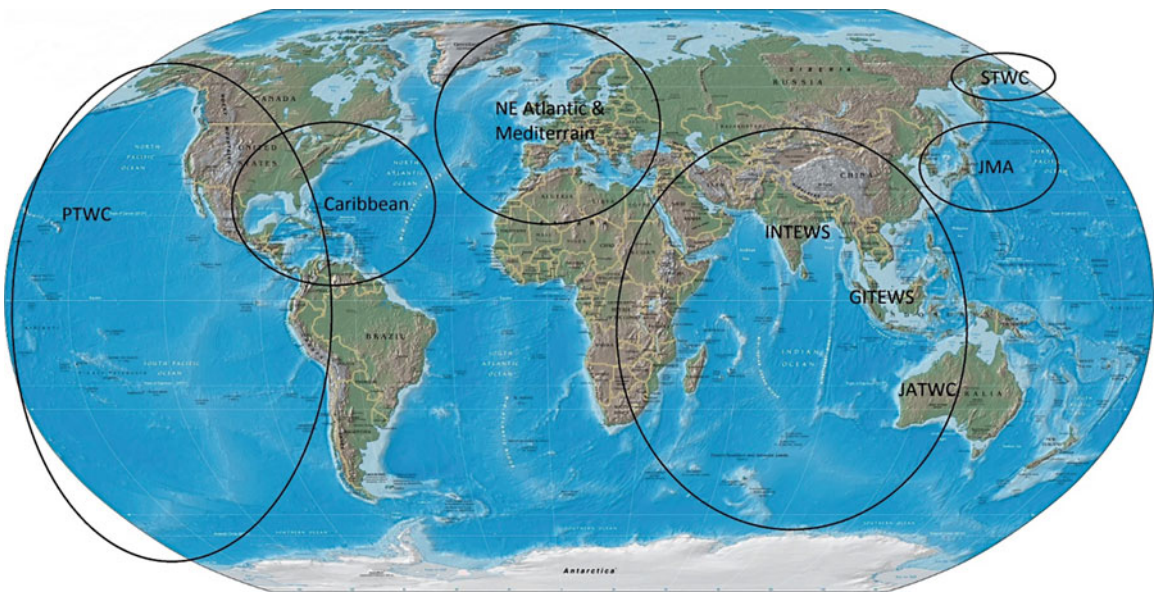
Key components

In order to carry out its mission, the TWC must at least consist of five important components: a network of sensors (seismic and sea level) to detect earthquakes and evaluate its potential for tsunamis, tsunami modeling capability (generation, propagation, and inundation models), high-resolution database on bathymetry and coastal topography, communication infrastructure to issue timely alarms to permit evacuation of coastal areas, and a 24 × 7 operational facility housing the infrastructure, decision support system (DSS) based on standard operating procedures (SOP), and experienced staff (Nayak and Kumar, 2008a). A typical outlay of TWC is shown in Figure 2.

Network of sensors

Seismic network

TWCs use local and global seismographic networks transmitting seismograms in real time to continuously monitor seismicity, in order to locate and estimate the size of potentially tsunamigenic earthquakes. To produce accurate

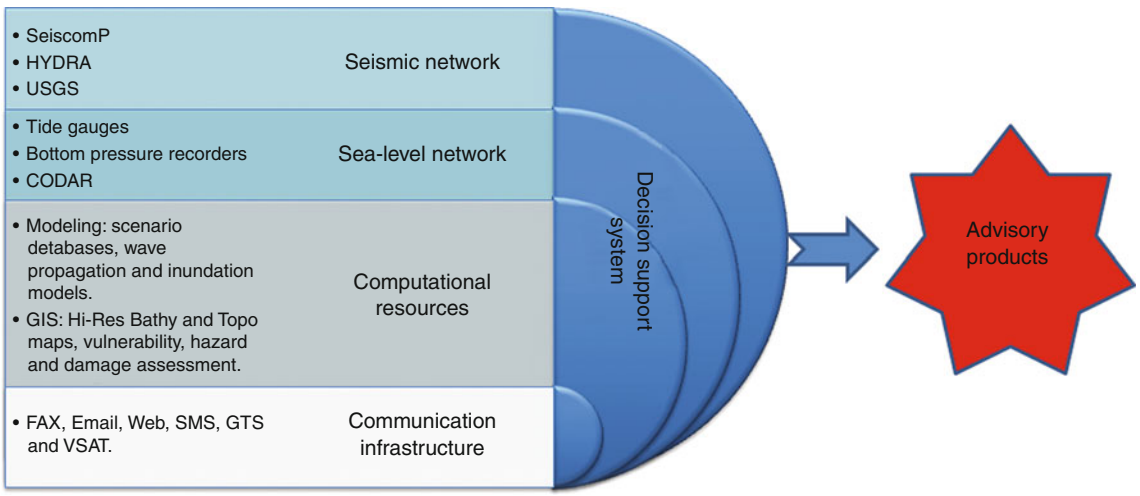


Tsunami Watch and Warning Centers, Figure 1 Tsunami warning systems presently operating in the world.

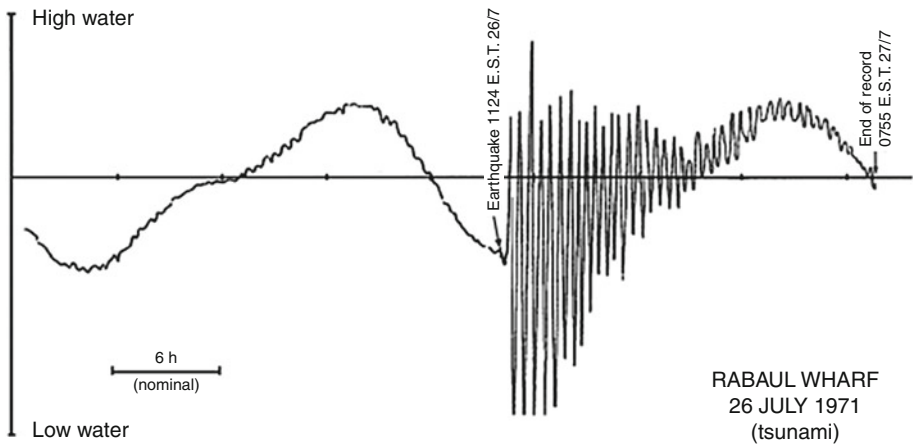
moment magnitude, NTWCs and RTWPs require reliable, broad-frequency, low-noise, high dynamic-range, digital seismic data in real time. The timeliness of the data is crucial in issuing an initial bulletin within 5 min of an earthquake. This is especially important for centers with local tsunami sources. In terms of network density, it is recommended that at least eight stations must be operating in a radius of 900 km, with data latency not exceeding 30 s.

Sea-level network

In order to confirm whether the earthquake has actually caused a tsunami, it is essential to measure the change in water level as near to the fault zone as possible with high accuracy. A TWC uses sea-level networks (tide gauge, bottom pressure recorders, and coastal ocean dynamics applications radar, CODAR) reporting data in real and near real time to verify the generation and evaluate the severity of a tsunami. Numerous international sea-level monitoring networks provide essential real-time data that can be accessed by NTWCs and RTWPs. Many of these networks are coordinated by the IOC of the UNESCO. The most extensive and notable is the Global Sea Level Observing System (GLOSS), conducted under the auspices of the Joint Technical Commission for Oceanography and Marine Meteorology (JCOMM) of the World Meteorological Organization (WMO) and the IOC. A tsunami creates a signal component that is not normally present in the observed sea level in usual times. Fortunately, the signature of tsunami waves at sea-level gages



Tsunami Watch and Warning Centers, Figure 2 Typical components of an end-to-end tsunami early warning system. (Courtesy: INCOIS.)



Tsunami Watch and Warning Centers, Figure 3 Typical tsunami signal found in tide gauge sea-level data. (UNESCO Manual of Sea Level Interpretation Vol. 4, 2006.)

is generally quite distinctive and thus recognizable (see Figure 3) and can be differentiated from a signal due to storm surge (UNESCO-IOC Manual and Guides, 2006).

Modeling capability

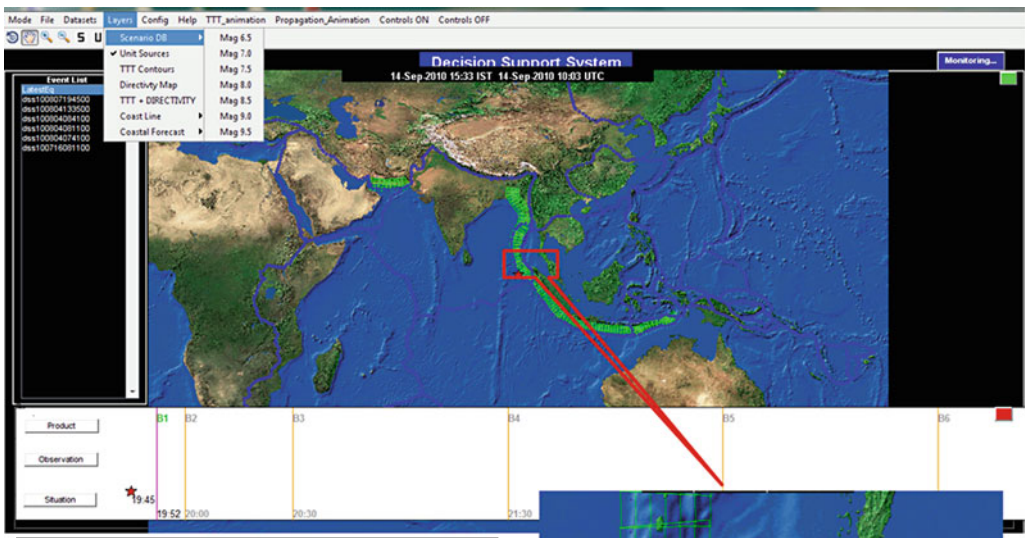
Guesstimating the generation mechanism of tsunami from the earthquake information, identifying the wave propagation scheme, forecasting the travel time, and mapping inundation zones are critical requirements in tsunami early warning, which can be accomplished using numerical models.

Tsunami modeling

Two of the most commonly used tsunami models presently are: the Method of Splitting Tsunami (MOST) developed by Titov and Gonzalez (1997) and the Tohoku University's Numerical Analysis Model (TUNAMI) developed by Imamura et al. (1995). Both these models are capable of simulating the three processes of a tsunami: earthquake generation, transoceanic propagation, and inundation of land. These models basically take the seismic deformation as input to predict the run-up heights and inundation levels at coastal regions for a given tsunamigenic earthquake (Imamura et al., 1995). The seismic deformation for an earthquake is computed using Mansinha and Smylie (1971) formulation using the earthquake parameters like epicenter, focal depth, strike, dip and rake of the rupture plane, length, width of the rupture, and slip magnitude on the rupture plane.

Scenario database

The above models cannot be run in real-time situations because of the computational overheads and resources. Hence, a pre-run scenario database or look-up-table (LUT) approach is most suitable in operational conditions. The forecast strategy is based on a unit-source function methodology, whereby the model runs are individually scaled and combined to produce arbitrary tsunami scenarios. Each unit-source function is equivalent to a tsunami generated by a Mw 7.5 earthquake with a rectangular fault 100 km by 50 km in size and 1 m slip (Gica et al., 2008; Greenslade and Titov, 2008; and Nayak and Kumar, 2008b, c). The faults of the unit functions are placed abutting each other. When the functions are linearly combined, the resultant source function is equivalent to a tsunami generated by a combined fault rupture with assigned slip at each sub-fault. These unit sources are located along the known fault zones for the entire Pacific Basin, Caribbean for the Atlantic region, and Indian Ocean. Figure 4 shows how the unit sources are set up in the Indian Ocean. A database of pre-run scenarios has been created for 1,000 unit sources covering all the tsunamigenic sources in the



Magnitude (M_w)	Width (W) (km)	Length (L) (km)	Slip (u_0) (m)
7.5	50	100	1
8.0	75	150	2.5
8.5	100	300	5
9.0	150	600	10
9.5	300	1200	20

Tsunami Watch and Warning Centers, Figure 4 Modeling using "unit-source" approach and the resulting scenario database. (Courtesy: INCOIS.)

Indian Ocean region (Kuwayama, 2006). At the time of earthquake occurrence, based on the location and magnitude, the basic unit-source open-ocean propagation scenarios are selected from the pre-run scenario database. These scenarios are then merged and scaled up/down using scaling relations (Nayak and Kumar, 2008c) to reconstruct a scenario representing the actual magnitude.

Geospatial integration: inundation modeling and mapping

The resulting open-ocean wave from the scenario database, described in previous section, has to be extended over to the shallow waters and to the coastal area to simulate inundation over land. Knowing the coastal inundation zone is the first step in disaster management and mitigation, as it directly tells us the extent of population to be evacuated and the estimated economic loss due to an event. To perform accurate modeling of tsunami inundation and run-up, a detailed model of near-shore bathymetry and coastal topography is necessary. This is crucial as the accuracy of model predictions is directly related to the quality of the data used to create the bathymetry and topography of the model area. Coastal Bathymetry is the prime determinant of the height of the tsunami wave or storm surge as it approaches the coast. Within the Indian Ocean context, topography of the region at 1:25,000-scale with contours at intervals of 0.5–2 m up to 20 m has been acquired. Preliminary maps have been prepared of coastal topography using satellite stereo data (CARTOSAT-1) for the Indian coast. These datasets, along with airborne **Li**ght **D**etection and **R**anging (LIDAR) data, were further used to create coastal vulnerability index (CVI) maps and multi-hazard vulnerability maps (MHVM) of coastline of India (Nayak and Kumar, 2008b). Similar high-resolution geospatial database is being leveraged in the Atlantic and Pacific Oceans (Gica et al., 2008).

Inundation mapping is not the only area where geospatial technologies are used, the inundation levels estimated and the subsequent threat status has to be accurately linked and conveyed to the respective administrative block within the coastal zones or within a province or a country. To implement such mechanism, the concept of coastal forecast points (CFP) and coastal forecast boxes (CFB) was developed (ICG/IOTWS) and implemented within the Indian Ocean region. Another major application of evolving geospatial trends is the application of 3D modeling or "videogrammetric" approach to visualize inundation and to identify evacuation routes and safe houses during inundation.

Communication infrastructure

Communication infrastructure is the most crucial and tangible aspect of any warning system for tsunami or any hazard. The ability to report and warn of an event is what makes and defines the warning system. Communication in terms of receiving all scientific data and disseminating information to the target audience within the required time needs a reliable communication infrastructure. During this process, TWC must respond fast, be as accurate as possible, and be reliable in order to be effective. All tsunami warning systems feature multiple lines of communications (such as SMS, email, fax, radio, texting, and telex, often using hardened dedicated systems), enabling emergency messages to be sent to the emergency services and armed forces, as well as to population alerting systems.

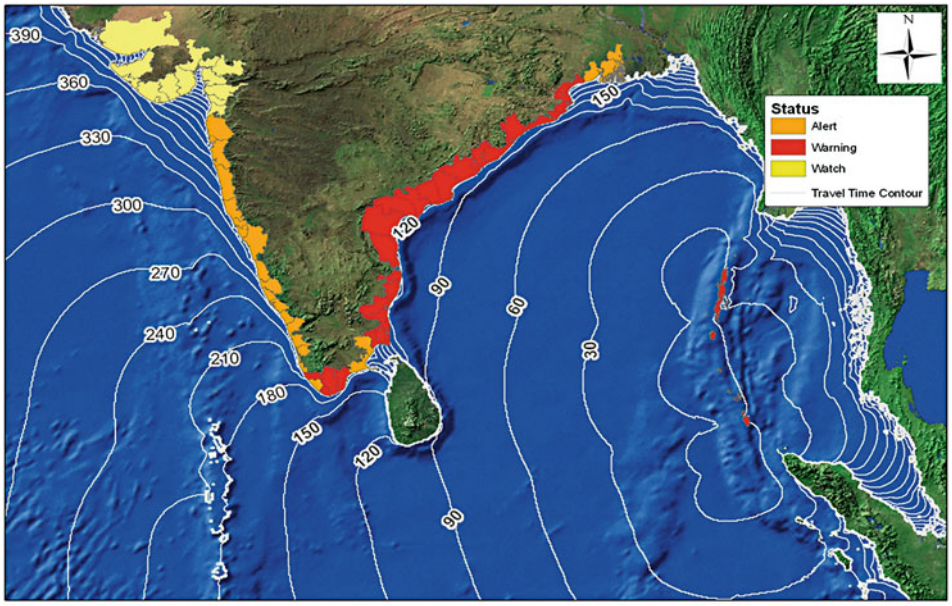
Decision support system and 24 × 7 operational center

In simple words, when the standard operating procedures are coded into a program that can support organizational decision making, then such a program can be called as a decision support system. It helps decision makers to compile useful information from a combination of raw real-time data, documents, personal knowledge, and tsunami scenarios models to accurately disseminate warning advisories to the needed area at the right time. The DSS encapsulates and foresees all the operations from the beginning of the event till the passage of threat by issuing timely advisories (Figure 2). The set of currently practiced advisories are earthquake information, warning, watch, and alert, which are described in detail in the following sections. All the information gathering, decision-making process, and infrastructure are part of the operational monitoring center.

Types of tsunami advisories

Based on the earthquake information, the tsunami modeling scenarios and operational capability, different types of advisory products are issued to the disaster management officials and to the general public. Each advisory has a distinct meaning and clearly suggests the action to be taken. The earthquake information bulletin issued solely on the earthquake information is a qualitative advisory while all others are quantitative advisories (Figure 5) issued taking into consideration the modeling results, and real-time sea-level observations, along with the earthquake information.

1. *Earthquake information bulletin* is issued to inform the occurrence of an earthquake and to advise its potential to generate a tsunami based on the magnitude. The bulletin contains information about the earthquake parameters (origin time, latitude, longitude, focal depth, and magnitude) and the name of the geographical area where it has occurred.
2. *Warning* is the highest level of advisory where immediate action has to be taken. This advisory is issued when a potential tsunami with significant widespread inundation is imminent or expected. Warnings alert the public that widespread, dangerous coastal flooding accompanied by powerful currents is possible and may continue for several hours after arrival of the initial wave. Warnings also alert emergency management



---- TEST TEST TEST TEST -----

TSUNAMI INFORMATION BULLETIN 03
Indian Tsunami Early Warning Centre (ITEWC)
Indian National Centre for Ocean Information Services (INCOIS), Hyderabad
PHONE: 91-40-23895011 FAX: 91-40-23895012 E-MAIL: tsunami@incois.gov.in

Date and Time of Bulletin Issue: 14-Oct-2009 0715 IST

Earthquake Parameters (Revised)

Preferred Magnitude	9.2 Mw (Great)
Network Magnitude(s)	9.2 (mB), 9.2 Mwp, 9.2 Mw(mB), 6.8 (MLv), 6.9 (mb)
Earthquake Origin Date & Time	14-Oct-2009 06:30:00 (IST)
Latitude	3.4 N
Longitude	95.7 E
Focal Depth	20 km
Land/ Ocean	Ocean
Water Level Depth (if Ocean)	3217 m
Region	Northern Sumatra, Indonesia

Tsunami Information

- Sea-Level Observations: Based on Sea level readings, a tsunami has been generated. It may already have been destructive along some coasts.

Tide Gauge Name	Lat	Lon	OWAT (IST)	OWH (m)
Rangat Bay	11.66	92.48	14-Oct-2009 06:47:00	1.96
Campbell Bay	6.90	93.71	14-Oct-2009 06:49:00	2.88
Meulaboh	6.64	93.77	14-Oct-2009 06:35:00	3.63

Legend:
OWAT: Observed Wave Arrival Time (IST)
OWH: Observed Wave Height (m) measured from mean to peak

- Pre-Run Model Scenario: Based on modeled results, the following areas are under WARNING/ALERT/WATCH / NO THREAT

District/Region Name	State/Territory	ETA (IST)	EWH (m)	Status
Hut Bay, Little Andaman	Andaman & Nicobar	14-Oct-2009 06:35	12.75	Warning
North Kamorta	Andaman & Nicobar	14-Oct-2009 06:35	10.23	Warning
West Kamotra	Andaman & Nicobar	14-Oct-2009 06:35	8.08	Warning
Campbell Bay	Andaman & Nicobar	14-Oct-2009 06:35	7.31	Warning
North Car Nicobar	Andaman & Nicobar	14-Oct-2009 06:35	6.39	Warning

Tsunami Watch and Warning Centers, Figure 5 An example of quantitative advisory bulletin product. The product includes graphic of tsunami travel time, along with text indicating the forecast of wave heights and arrival time along with threat status to each administrative zone in the Indian Ocean. (Courtesy: INCOIS.)

officials to take action for the entire tsunami hazard zone. Warnings may be updated, adjusted geographically, downgraded, or canceled.

3. *Watch* is issued to alert emergency management officials and the public of an event which may later impact the watch area. Immediate public evacuation is not required, Local officials should be prepared for evacuation if it is upgraded to warning status. The watch area may be upgraded to a warning or canceled based on updated information and analysis. Watches are normally issued based on seismic information without confirmation that a destructive tsunami is underway.
4. *Alert* contains information about the earthquake and a qualitative tsunami evaluation message indicating that a tsunami is expected. As conditions warrant, the advisory will either be continued, upgraded to a watch or warning, or canceled. Alert implies public should avoid beaches since strong currents are expected. Local officials should be prepared for evacuation if it is upgraded to warning status.
5. *Cancelation* indicates the end of the threat due to tsunami. A cancelation is usually issued after an evaluation of sea-level data confirms that a destructive tsunami will not impact the warned area anymore. A cancelation will also be issued following a destructive tsunami when sea-level readings indicate that the tsunami is below destructive levels and is subsiding in most locations.

Service levels

Based on the level of integration of key component into the TWC, each TWC has the capability to deliver advisory products at certain operational level. Each higher service level represents much more detailed threat evaluation and will be directed to more specific agencies/population group for disaster mitigation with increased confidence levels. Presently, the implementation plan of IOTWS (UNESCO-IOC, 2007) classifies these product levels into three groups, based on incremental capacity development:

Service Level I: At this level of operation, the TWC disseminates only the earthquake information and a qualitative tsunami threat evaluation to other national warning centers.

Service Level II: This is a more robust service level that includes all of the elements of Service Level I but adds modeling and forecast elements so that Watches and Advisory products include: Estimated wave height(s) (EWH) and Estimated time(s) of arrival (ETA) for offshore forecast points, potential threat zones, advanced warning and notification products of threat status and zones in image format. At this service level of operation, real-time observations are included in the decision making and in advisory products from Bulletin – 3 onward.

Service Level III: This is the most advanced level of operational service. At this level, the advisory products include local risk and hazard assessment using inundation models, and they are more specifically directed to communities at risk. Service Level III operations take into account real-time inundation forecast in evaluating the threat status. A TWC operating at this service level indicates a mature early warning system capable of detecting and evaluating tsunamigenic earthquakes occurring anywhere in the global oceans.

Conclusion

Although tsunamis are known to occur rarely, the key elements integrated in a TWC become the essential and minimum requirements in dealing with other ocean-related hazards like: cyclone related storm surges, cloud bursts, and impacts of long-term coastal erosion and sea-level rise. Geospatial technology has immensely helped in the design of early warning system for tsunami, and the GIS base layers and survey data make a TWC capable of dealing with coastal threats due to multiple hazards.

Acknowledgments

The authors would like to thank the Tsunami Early Warning Team at the Indian National Centre for Ocean Information Services (INCOIS) for their support in the preparation of graphics for this manuscript, especially Dr. Uma Devi, Mr. Ch. Patanjali Kumar, and Mr. Raghavendra S. Mupparthy. The authors are thankful to the Director of INCOIS for his active cooperation and encouragement during this process.

Bibliography

Gica, E., Spillane, M. C., Titov, V. V., Chamberlin, C. D., and Newman, J. C., 2008. Development of the forecast propagation database for NOAA's short-term inundation forecast for tsunamis (SIFT). NOAA Technical Memorandum OAR PMEL139. Contribution No. 2937 from NOAA/Pacific Marine Environmental Laboratory.

Greenslade, D. J. M., and Titov, V. V., 2008. A comparison study of two numerical tsunami forecasting systems. *Pure and Applied Geophysics*, **165**, 1991–2001, doi:10.1007/s00024-008-0413-x.

Gupta, H. K. 2005. Mega-Tsunami of 26th December, 2004, 2004: Indian initiative for early warning system and mitigation of oceanogenic hazards. Episodes, March 2005.

Imamura, F., Yalciner, A. C., and Ozyurt, G., 1995. Tsunami modeling manual (TUNAMI model). April (available at: http://www.tsunami.civil.tohoku.ac.jp/hokusai3/E/projects/manual-ver-3.1.pdf)

IOC-ICG/IOTWS-V/13. 2008. Implementation Plan for Regional Tsunami Watch Providers (RTWP), Fifth Session of the Intergovernmental Coordination Group for the Indian Ocean Tsunami Warning and Mitigation System (ICG/IOTWS-V) Kuala Lumpur, Malaysia, 8–10 April 2008.

Kuwayama, T., 2006. Quantitative tsunami forecast system. JICA Tsunami Seminar – Unpublished report.

Mansinha, L. A., and Smylie, D. E., 1971. Displacement fields of inclined faults. *Bulletin of the Seismological Society of America*, **61**, 1433–1440.

Nayak, S., and Kumar, S. T., 2008a. *The First Tsunami Early Warning Center in the Indian Ocean*. Tudor Rose: Risk Wise, pp. 175–177.
Nayak, S., and Kumar, S. T., 2008b. Addressing the risk of tsunami in the Indian Ocean. *Journal of South Asian Disaster Studies*, **1**(1), 45–57.
Nayak, S., and Kumar, S. T., 2008c. Indian tsunami warning system, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII, Part B4, Beijing.
Titov, V. V., and Gonzalez, F. I., 1997. Implementation and testing of the method of splitting tsunami (MOST) model. NOAA Technical Memorandum ERL PMEL-112. Contribution No.1927 from NOAA/Pacific Marine Environmental Laboratory.
UNESCO-IOC. 2006. Tsunami Glossary. IOC Information document No.1221. Paris, UNESCO.
UNESCO-IOC. 2007. IOTWS Implementation Plan for Regional Tsunami Watch Providers, IOC Technical Series 81. Paris, UNESCO.
UNESCO-IOC, Manual and Guides, 2006. Manual on Sea-level Measurements and Interpretation, Volume IV; JCOMM Technical Report No.31; WMO/TD. No.1339: An update to 2006. Paris, Intergovernmental Oceanographic Commission of UNESCO. 78 pp.

Cross-references

Earthquakes, Early and Strong Motion Warning
Tsunami
Tsunami: Bay of Bengal
Seismicity, Subduction Zone

V

VERTICAL SEISMIC PROFILING

James W. Rector, III[1], Maria-Daphne Mangriotis[2]
[1]Department of Civil and Environmental Engineering, University of California at Berkeley, Berkeley, CA, USA
[2]Applied Geophysics, Institute of Engineering Seismology and Earthquake Engineering (ITSAK), Thessaloniki, Greece

Synonyms
VSP

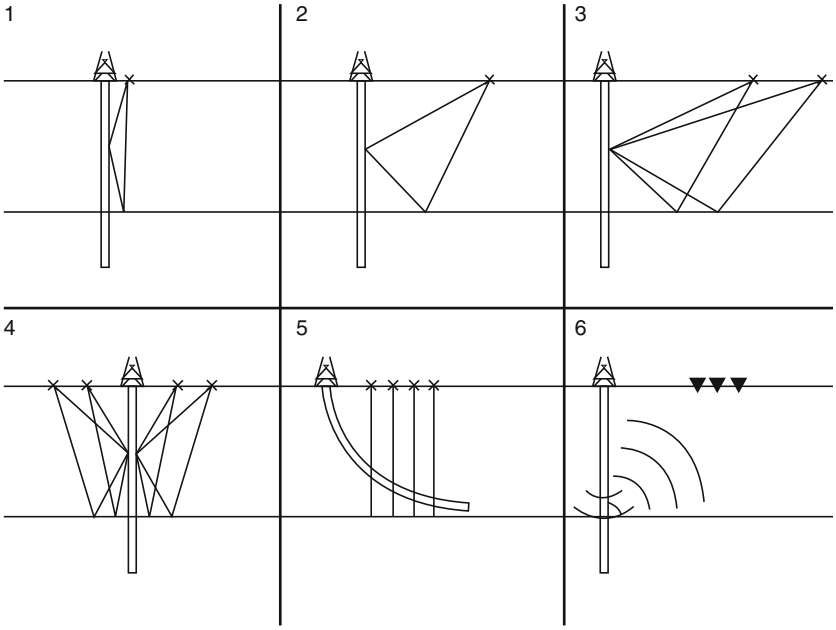
Definition
Vertical Seismic Profiling A geophysical technology for measuring the *seismic* properties in a *profile* of the earth using a set of sources and receivers, either of which are placed along the depth (*vertical*) axis.

Introduction
Vertical seismic profiling (commonly abbreviated to the name VSP) is a major technique in geophysical exploration, which involves measuring downgoing and upgoing seismic wavefields through a stratigraphic sequence. Shortly after the first attempts of developing surface reflection seismology, where both sources and receivers were at the surface, the line of research set out with the idea of burying either sources or receivers in boreholes. McCollum and LaRue (1931) proposed that local geologic structure could be determined from measuring the arrival of seismic energy from surface sources with buried receivers. This geometry forms the basic principle of determining seismic wave propagation velocities from VSP measurements even today. The technology of VSP acquisition and processing was initially developed by researchers in the former Soviet Union in the 1950s through the 1970s. It was only later, in the mid 1970s, that non-Soviet geophysicists became systematically engaged in the advance of VSP technology for exploration applications. Today, the fruits of these efforts have served not only in the improvement of velocity analysis, but have materialized numerous other applications ranging from structural imaging of complex 3D structures such as faults and salt domes in exploration studies, to directing drill operations, monitoring secondary recovery processes, and hydraulic fracturing.

Acquisition techniques
A variety of data acquisition techniques are currently employed, depending on the survey's application and financial budget. Figure 1 shows the main geometric configurations, which include (1) zero-offset VSP (ZVSP), where the source is near the well-head above the vertical receiver array, (2) offset VSP, where the source is located a distance away from the vertical borehole, (3) multi-offset VSP, for multiple distances of the source to the borehole, (4) walk-away VSP, where multiple sources are located on a line at the surface and the receivers in a vertical array, (5) walk-above VSP, where the sources are located over receivers buried in a deviated or horizontal well, and (6) drill-bit SWD (seismic while drilling) VSP, where the drill-bit acts as the seismic source and the receivers are on the surface. 3D VSP uses borehole receivers, and a 2D surface source geometry (i.e., multi-azimuth multi-offset, or walk-away VSP).

The common active seismic sources used in VSP include those used in land and marine surface studies (see *Seismic Data Acquisition and Processing*) as well as the noise of the drill-bit in the case of SWD-VSP. There is a variety of geophone tools available for use in cased or uncased boreholes, which combine three orthogonal geophones, a hydrophone and other borehole seismic tools at each recording level. The receiver spacing is set by the depth-sampling interval to avoid spatial aliasing.

Harsh K. Gupta (ed.), *Encyclopedia of Solid Earth Geophysics*, DOI 10.1007/978-90-481-8702-7,



Vertical Seismic Profiling, Figure 1 Common acquisition geometries for vertical seismic profiling (VSP): (1) zero-offset VSP (ZVSP), (2) offset VSP, (3) multi-offset VSP, (4) walk-away VSP, (5) walk-above VSP, and (6) drill-bit SWD (seismic while drilling) VSP.

To limit field acquisition time, most commonly multilevel receiver arrays are employed. The main source of noise in VSP studies are tube waves, which are generated when the fluid in the well is displaced by a seismic source, such as ground roll. Tube waves propagate up and down a well and create secondary tube waves in borehole anomalies, which may interfere with primary reflections. As a coherent noise mode, tube waves cannot be attenuated by stacking, therefore, both field procedures, such as use of attenuating heavy mud and construction of ditches to attenuate the ground roll, as well as processing techniques, such as filtering and tube wave modeling, are commonly used to minimize their effect.

VSP applications

The primary use of VSP is the integration of VSP and downhole logging to surface reflection surveys (Hardage, 1985). Contrary to surface studies, which infer subsurface conditions from measurement of the reflected wavefields with the source and receiver at the surface, a VSP configuration involves closely spaced direct physical measurements of the wavefield between the source position and the reflector.

Zero-offset VSP measurements can be correlated with sonic logs and check-shot surveys in the borehole and improve the depth positioning of reflections recorded in surface data. From a velocity perspective, zero-offset VSP measurements are denser than check-shot surveys; hence, they provide a finer interval velocity profile. On the other hand, VSP samples a larger velocity body around the well than the high frequency sonic log, which may be subject to cycle skipping and the effects of the washed-out zones. VSP also contains the signature of the near-surface, which is often absent in logging. Hence, synthetic seismograms computed from sonic logs may not accurately represent reflections recorded from surface measurements. Reflections from zero-offset VSP, however, can be matched to those from surface measurements via an appropriate static time-shift to account for the differential travel times of the two recording geometries.

Concerning seismic imaging, the recorded wavefield in a buried VSP array has higher signal to noise ratio across a larger seismic bandwidth than surface measurements for two reasons. First, surface cultural noise is suppressed in the borehole, and second, upgoing reflections do not propagate through the near-surface material, which is highly *attenuating* (see *Seismic, Viscoelastic Attenuation*). In addition, the reflections recorded in a VSP suffer less attenuation as the distance to the target is shorter. This shorter distance also implies that the *Fresnel zone* (see *Seismic, Waveform Modeling and Tomography*) is smaller in the case of a VSP, leading to higher resolution compared to surface seismic studies. Given that downgoing and upgoing wave modes are separated in the f-k space, either an f-k filter, or a trace by trace arithmetic subtraction of the estimated wave mode can isolate downgoing from upgoing waves. Estimation of the downgoing wavefield from VSP data is used in the design of deconvolution operators, which can be thereafter applied to surface data to improve the vertical resolution. On the other hand, upgoing reflections can be mapped to their origin position via either a CDP-VSP mapping, or *Seismic, Migration* algorithm, or a combination of the two. A migration step is strongly advisable in the case of

complex geologies. In recent years, VSP seismic imaging has transformed from an analysis of acoustic reflections to migrating converted-wave reflections (Grech et al., 2001), which may be more sensitive to imaging tight-sand reservoirs compared to P-P reflections (O'Brien and Harris, 2006), as well as to applications of free-surface multiples, which can increase the narrow cone of coverage of a VSP to that of a 3D CDP survey (Jiang et al., 2005). Another promising technique for seismic imaging applications is interferometric imaging, which through the creation of virtual sources (VS) in the location of the receivers eliminates the need for any knowledge of the complex overburden (Bakulin and Calvert, 2004). In addition, unlike traditional migration schemes, interferometric imaging circumvents a priori knowledge of the velocity structure in the vicinity of targets, such as salt flanks, and can be applied to imaging complex structures above or below a receiver array (Xiao et al., 2006; Vasconcelos et al., 2008).

Through the separation of downgoing and upgoing wavefields and their relative amplitudes, it is possible to extract the true *reflection coefficients* (see *Energy Partitioning of Seismic Waves*) used for lithological interpretation. Moreover, from multi-offset 3-component VSP data it is possible to determine the *AVO response* (see *Energy Partitioning of Seismic Waves*) with no assumptions to the path of the propagating wavefield, which is required in surface seismic AVO analysis.

Another benefit of using VSP measurements is their robustness in estimating *Seismic Anisotropy*, which improves lithologic interpretations as well as increases the quality of seismic imaging. Given the seismic velocity response to changes in reservoir porosity and fluid content, velocity *Seismic Anisotropy* can also be applied in monitoring of fluid-drainage patterns during production. Fracturing will also affect seismic velocity; hence, VSP velocity measurements can also be used in monitoring hydraulic fracturing.

With *attenuation* (see *Seismic, Viscoelastic Attenuation*) gaining increasing importance as a seismic parameter linked to lithological properties, VSP has proven to be a powerful tool for Q (see *Seismic, Viscoelastic Attenuation*) estimation. Aside from true amplitude processing required in stratigraphic interpretation, VSP-derived Q profiles are also used in the design of inverse Q-filters to enhance seismic resolution. Recent studies also suggest a link between anisotropic attenuation and fracturing (Maultzsch et al., 2007); hence, VSP azimuthal measurements of attenuation may potentially be applied in monitoring hydraulic fracturing in hydrocarbon and geothermal applications.

Summary

In recent years, VSP surveys are becoming increasingly popular for reducing risk in well placement and improving reservoir monitoring. Intrinsic to the complex geometry of the VSP, which poses several data acquisition and processing challenges, is the ability of the VSP to expand the spatial and time-lapse subsurface characterization potential of conventional downhole logging and surface reflection studies. Henceforth, following the success of VSP surveys and their increased benefits over their financial cost, the VSP technique is a rapidly developing area of geophysical research and may soon become a standard logging service.

Bibliography

Kirtland Grech, M. G., Lawton, D. C., and Cheadle, S., 2001. Integrated Pre-stack Depth Migration of VSP and Surface Seismic Data. CSEG, Calgary, Alberta, Canada, November Issue: 50–56.

Bakulin, A., and Calvert, R., 2004. Virtual Source: New Method for Imaging and 4D Below Complex Overburden. *Society of Exploration Geophysicists Expanded Abstracts*, **23**, 2477–2481.

Hardage, B. A., Helbig, K., and Treitel, S., 1985. Vertical Seismic Profiling, *Handbook of Geophysical Exploration, Seismic Exploration*: Geophysical Press. Elsevier Science Ltd, The Boulevard, Langford Lane, Kidlington, Oxford, OX5, GB, UK.

Jiang, Z., Yu, J., and Schuster, G. T., 2005. Migration of multiples. *The Leading Edge*, **24**, 315–318.

Maultzsch, S., Chapman, M., Liu, E., and Li, X.-Y., 2007. Modeling and analysis of attenuation anisotropy in multi-azimuth VSP data from the clair field. *Geophysical Prospecting*, **55**, 627–642.

McCollum, B., and LaRue, W. W., 1931. Utilization of existing wells in seismograph work. *Early Geophysical Papers*, **1**, 119–127.

O'Brien, J., and Harris, R., 2006. Multicomponent VSP imaging of tight-gas sands. *Geophysics*, **71**, 83–90.

Vasconcelos, I., Snieder, R., and Hornby, B., 2008. Imaging of internal muliples from subsalt VSP data – Examples of target-oriented interferometry. *Geophysics*, **73**, 157–168.

Xiao, X., Zhou, M., and Schuster, G. T., 2006. Salt-flank delineation by interferometric imaging of transmitted P- to S-waves. *Geophysics*, **71**, 197–207.

Cross-references

Energy Partitioning of Seismic Waves
Seismic Anisotropy
Seismic Data Acquisition and Processing
Seismic, Migration
Seismic, Viscoelastic Attenuation
Seismic, Waveform Modeling and Tomography

VERY LONG BASELINE INTERFEROMETRY

Helmut Wiesemeyer[1], Axel Nothnagel[2]
[1]Millimeter and Submillimeter Astronomy, Max-Planck-Institute for Radio Astronomy, Bonn, Germany
[2]VLBI Research Group, Institute of Geodesy and Geoinformation, University of Bonn, Bonn, Germany

Synonyms

VLBI

Definition

Very Long Baseline Interferometry (VLBI). correlation interferometry of celestial objects with separately

operating radio antennas equipped with frequency standards, off-line data reduction remotely by a digital correlator.
e-VLBI. VLBI with transmission of data by a high-speed network instead of storage media.

Introduction

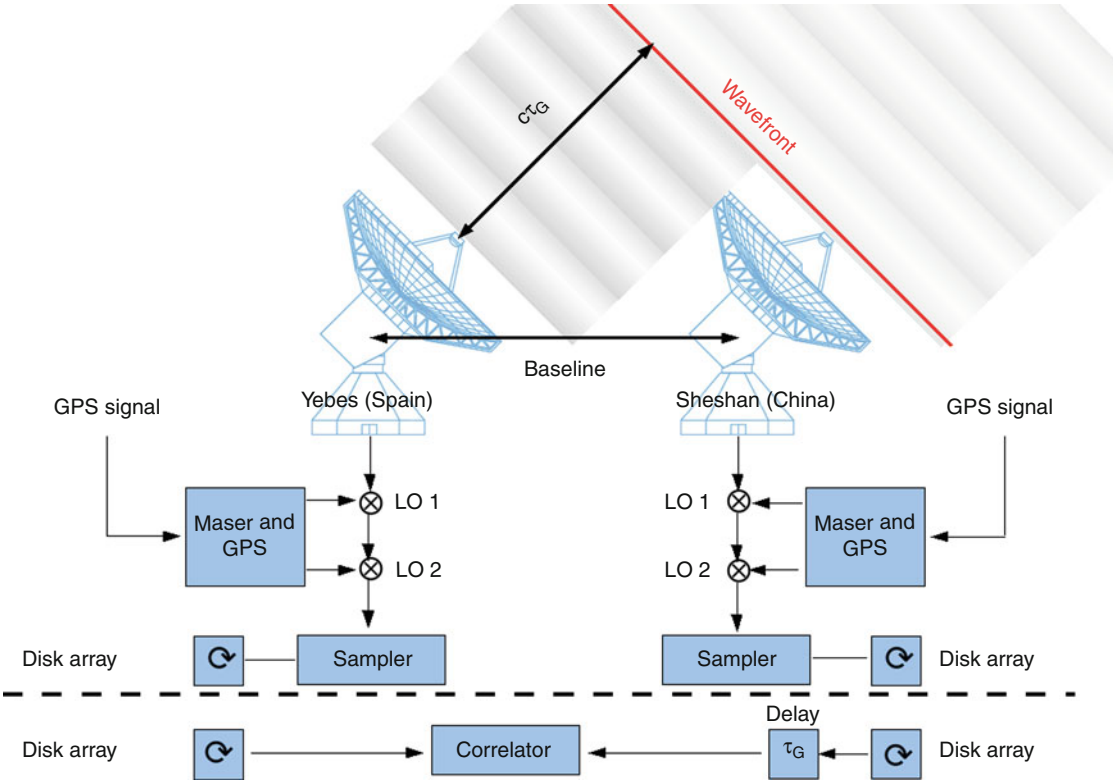
Very long baseline interferometry (VLBI) is a radio-astronomical technique to achieve high spatial resolution in astronomical imaging (of the order of milliarcseconds), involving antennas separated by thousands of kilometers (a "VLBI network"). Because the reconstruction of the image depends on the precise knowledge of the vector (hereafter: baseline) connecting the two antennas of a station pair, the a priori knowledge of the image can be used to determine the baseline vector by inverting the measurement equation, with a precision corresponding to a fraction of the wavelength. Such an a priori knowledge is naturally given if the astronomical source on the sky is spatially unresolved and therefore a "point source." Such sources – usually quasars, that is, the nuclei of bright radio galaxies – are spread over the entire sky and regularly observed in radio astronomy. Whereas VLBI in radio astronomy aims at resolving these active galactic nuclei, improper assumptions about source structure limit the accuracy of the determinations of antenna stations. This can be avoided by using appropriate source structure models (Tornatore and Charlot, 2007). The importance of VLBI techniques in Geophysics is evident if the two antennas of a baseline are located on different tectonic plates, whose relative motion and deformation can therefore be determined with high accuracy (qv *Geoid*).

A VLBI array can be thought as a fictive giant single-dish telescope whose aperture is synthesized while the baseline vector sweeps the sky thanks to the rotation of the Earth. The resulting image therefore depends on the angular momentum vector of the Earth. This dependency implies applications of VLBI in the measurements of the Earth's spin axis in space and with respect to the solid crust as well as irregularities of the *Earth Rotation*. Other, less precise methods are position determinations with the Global Positioning System (qv *GPS, Tectonic Geodesy*) and synthetic aperture radar measurements (qv *SAR Interferometry*), the latter being an interferometric method, like VLBI.

Methods and techniques in very long baseline interferometry

The idea of connecting several telescopes to an array in order to synthesize part of the surface of a single, large telescope dates back to 1946, when Ryle and Vonberg (1946) used for the first time a multi-element radio telescope for solar observations. In both optical and radio astronomy, the idea turned out to be extremely fruitful for achieving the highest spatial resolutions. For a single telescope, the resolution (in radians) is given by the ratio of wavelength to aperture diameter (up to a factor of order unity). By consequence, a radio telescope working at a wavelength of 21 cm needs an aperture of about 50 km in order to achieve the same angular resolution as, for example, an optical telescope limited by atmospheric seeing. The largest, fully steerable radio telescopes have apertures of the order of 100 m and reach, at that wavelength, an angular resolution of 9 arc minutes. As an alternative to even larger single-dish radio telescopes, several telescopes can be connected by microwave or fiber optics cables in order to form a connected-element interferometer. If the telescopes of such an interferometer are suitably distributed, a sufficiently long observation of a celestial source can be reconstructed to form an image whose angular resolution is of the order of the ratio of the wavelength to the longest among the distance vectors ("baseline," see Figure 1) between telescope pairs. The advantage of connected-element interferometers is that the signals can be correlated in real time, such that the phase of the electromagnetic wave when it arrives at the different telescopes is preserved. This is achieved by mixing the incoming signal with that from a local oscillator. In a connected-element interferometer, the local oscillators in all antennas are driven with a common phase reference. The Very Large Array (VLA, New Mexico) and the British e-Merlin array have elements connected by optical fiber cables, with baselines of up to 50 km (VLA) and 200 km (e-Merlin). Despite modern communication technology sometimes used (e.g., satellite links), the physical connection of array elements becomes impractical beyond a certain limit.

The step from a connected-element interferometer to a VLBI experiment consists of abandoning (for practical and economical reasons) the microwave or optical fiber links, registering the data, together with time marks, on a suitable storage medium, and analyzing the data later with a digital correlator. While local oscillators are necessary to convert the observed frequency band pass to one at a lower frequency (for recording the data), they cannot be locked to a common phase reference because of the large separation between the antennas. Time marks are therefore necessary. This technique became feasible with the availability of reliable frequency standards for the time marks, and magnetic tapes of sufficient storage capacity. The first VLBI observation correlated the data of a single baseline between the Hat Creek (California) and Green Bank (West Virginia) observatories (Bare et al., 1967). Since then, the major improvements in VLBI were mainly due to the increase of baseline length (i.e., spatial resolution) and collecting area (i.e., sensitivity), and due to the availability of more reliable clocks and faster digital samplers, leading to larger spectral bandwidth and thus sensitivity. Nowadays, hydrogen masers are used as frequency standards, while arrays of hard disks replace the tapes. The distinction between a connected-element interferometer and a VLBI network is somewhat imprecise. Using network connections of large bandwidth, VLBI data can now be transmitted in real time to the central correlator, while the local oscillators of the antennas participating in the



Very Long Baseline Interferometry, Figure 1 Schematic representation of an antenna pair on a baseline between Spain and China (distances not to scale). The electromagnetic wave from a distant point source (a quasar) is shown together with the position of the wave front when arriving in China (red line). The signal path corresponding to the time delay τ_G is shown, as well as the correction for the latter in the data reduction. There is no physical connection between both antennas, each operating with an independent local oscillator (labeled LO). The dashed line separates the measurement from the off-line data processing. Credits: Greve/IRAM/Rebus.

network are not synchronized. This technique has been demonstrated in 2005 with the European VLBI Network (EVN), involving a transatlantic baseline.

Physical principles

Interferometric experiments detect both amplitude and phase of electromagnetic waves (conveniently described by complex functions) and combine them in a suitable way to provide images, spectra, and polarization information. Phase-sensitive detection is possible by measuring correlations (or coherence functions) in the incoming electromagnetic field. It has been shown (Sudarshan, 1963) that for the case of two-point correlation functions, classical and quantum electrodynamics lead to equivalent results if nonlinear effects are ignored. We therefore limit our discussion to the case of ensembles of incoherent packets of quasi-monochromatic electromagnetic waves from different directions but confined to a small field of view.

The starting point of any interferometric imaging experiment is the van Cittert-Zernicke theorem (van Cittert, 1934; Zernike, 1938), stating that the correlation of the electromagnetic field at two points (two-point or mutual coherence function) is, up to a complex scaling factor, given by the Fourier transform (also called "visibility function") of the spatial intensity distribution (example: the interference pattern observed on a screen behind an illuminated slit or a grid). The complex scaling factor involves a correction due to the fact that the antenna baselines are not parallel to the plane of the sky (and, for the case of VLBI, not coplanar among them). This so-called delay correction τ_G needs to be applied because of the lag between the arrival times of the front of the electromagnetic wave at the antennas of a given baseline (Figure 1), and is equivalent to steering the fictive large single-dish telescope mentioned above to the source. The applied delay correction will differ from the theoretical one, owing to the clock error and to the inaccuracy of the assumed baseline length. In the absence of an error-free, absolute time scale, the data streams from an antenna pair rather have to be synchronized by maximizing the correlated signal, applying reasonably guessed delay corrections. A first synchronization is done by using the GPS time signal (qv *GPS, Data Acquisition and Analysis*) to which the local oscillators in the antennas are locked. A refinement needs to be applied later when the data are cross-correlated. There are actually two free parameters

per baseline pair, not only the delay correction, but also the frequency of the oscillations, due to the Earth's rotation, in the interference pattern ("fringes"). This frequency has to be fitted, too, because it depends on the clock rate. For hydrogen masers, whose radio signal at a frequency of 1.4 MHz is now commonly used as a frequency standard for VLBI experiments, the fractional frequency drift (e.g., due to temperature drifts) after 100 s is of the order of 10^{-14}. Such a high precision is required if phase information is to be extracted once the signals are correlated. There is yet another difference between a connected-element interferometer with baselines of up to 10^6 wavelengths, and a VLBI experiment: in the former case, the field of view of the interferometer is limited by the central lobe of the antenna reception pattern ("main beam"), whereas in VLBI the field of view is limited by the largest delay τ_G within which the signal from a given direction is coherent. This delay depends on the baseline length with respect to the wavelength, on the spectral bandwidth across which these signals are averaged, and on the accuracy required (e.g., astrometry or precise baseline determinations need a higher degree of coherence). Therefore, the field of view of VLBI observations is much smaller than that of observations with connected-element interferometers. However, nowadays more evolved data analysis techniques exist that permit to overcome this limitation (Garrett et al., 1999).

For each telescope of the VLBI network, the data from the receiving system are recorded on an array of hard disks, together with the time marks from the clock. Later these data are correlated with a digital correlator. Modern correlators currently used are the Mark IV and Mark V or K4 and K5 systems. They allow us to analyze VLBI data for large frequency bands (for high sensitivity to continuum signals) or high-frequency resolution (for observing spectral lines).

Data calibration and imaging

After applying the aforementioned delay and fringe rate corrections, the phases of the complex data need to be recovered. The visibility functions are first smoothed to an integration time interval, which depends on the tolerated "visibility smearing": owing to the rotation of the Earth during this time span, different spatial frequencies are measured at different times, even for a single baseline. Furthermore, at the highest frequencies, where atmospheric fluctuations become important (mm VLBI), the elementary integration times are kept below 20 s, which allows one to apply atmospheric phase corrections. Another upper limit to the integration time interval is given by the stability of the clock. However, the time constant of a hydrogen maser is longer than the two previous limitations.

Unlike connected-element interferometers, where the phase of the front of the incident electromagnetic wave is not too much distorted by atmospheric effects, the phase difference between signals received by different VLBI stations is completely lost. This is because the atmospheric fluctuations at the different antenna sites are uncorrelated, and because the local oscillators in the antennas are not locked by a common phase reference. Fortunately, only phase differences with respect to a reference antenna are important. For a triangle of VLBI antennas, the phase errors due to atmospheric and instrumental fluctuations cancel out in the sum of the phase differences between the antennas of the three baselines involved ("phase closure"). Since the number of baselines scales roughly with the square of the number of antennas, whereas the number of phase errors increases only linearly with it, a considerable part of the phase information can be recovered (e.g., 89 % for the 18 antennas of the EVN). This recovery of phase information can be improved by a "self-calibration" initially assuming a point source as model, calibrating the measured phases under this assumption, and then creating a new source model using an imaging algorithm. In the underlying theory, complex antenna gains are used, and closure relations also exist for their amplitudes.

Such imaging algorithms are needed because the VLBI stations cannot sample the visibility functions on a regular grid of spatial frequencies. Therefore, the brightness distribution resulting from a Fourier transform is convolved with the "synthesized beam," which is the Fourier transform of the sampling of spatial frequencies. Iterative algorithms exist to deconvolve the image such that the result is the brightness distribution as observed with a fictive single dish telescope whose diameter corresponds to the longest available baseline. The CLEAN algorithm (Högbom, 1974) considers the brightness distribution as a collection of spatially unresolved sources. Other deconvolution algorithms have been proposed and applied since then, depending on the nature of the observed brightness distributions. The combination of CLEAN and phase closure is known as hybrid mapping (e.g., Readhead et al., 1980). Global fringe fitting (Schwab and Cotton, 1983) extends the closure relations to time delays and fringe rates and is iteratively used together with a deconvolution algorithm like CLEAN.

For high-precision work (e.g., measurements of the Earth rotation parameters, see below), general relativistic effects need to be considered. As an indication, the gravitational time delay due to the Earth's gravity amounts to up to ~0.02 ns on a 6,000 km long baseline, and the positions of the celestial sources have to be corrected for light bending in the gravitational field of the sun (4 milliarcsecond for an incidence orthogonal to the direction from Earth to Sun, for a summary and further references, see Heinkelmann and Schuh, 2010; Fomalont et al., 2010).

VLBI networks

VLBI networks now extend all over the world. The most sensitive network is the European VLBI Network (EVN), connecting 18 radio telescopes in Europe, Puerto Rico, and South Africa, operating together for more than

3 months per year. The American Very Large Baseline Array (VLBA) works for the whole year. Other national VLBI facilities exist in the UK (the Merlin array), Japan (Japanese VLBI network JVN), and Australia (long baseline array, LBA). Occasionally, some of these networks work together to form global VLBI networks (e.g., the EVN with the VLBA or Merlin, and the Asia-Pacific telescope APT with observatories in Australia, China, New Zealand and Japan). Currently, the best angular resolutions (down to 0.1 milliarcsecond) are achieved with global VLBI at 3 mm wavelength. Future improvements of the spatial resolution scale are to be expected from VLBI below 3 mm wavelength, and from networks including satellite stations (with the Japanese HALCA antenna, in operation from 1997 to 2005, to be replaced by the VSOP 2 mission). These VLBI networks are mainly used for astrophysical research. However, VLBI is also used extensively for geodetic and geophysical applications (see below). Owing to the importance of the VLBI techniques for these disciplines, the International VLBI Service for Geodesy and Astrometry (IVS, Schlüter and Behrend, 2007) coordinates the observations of a dedicated network of VLBI stations on all continents (Figure 2), operated by national research agencies.

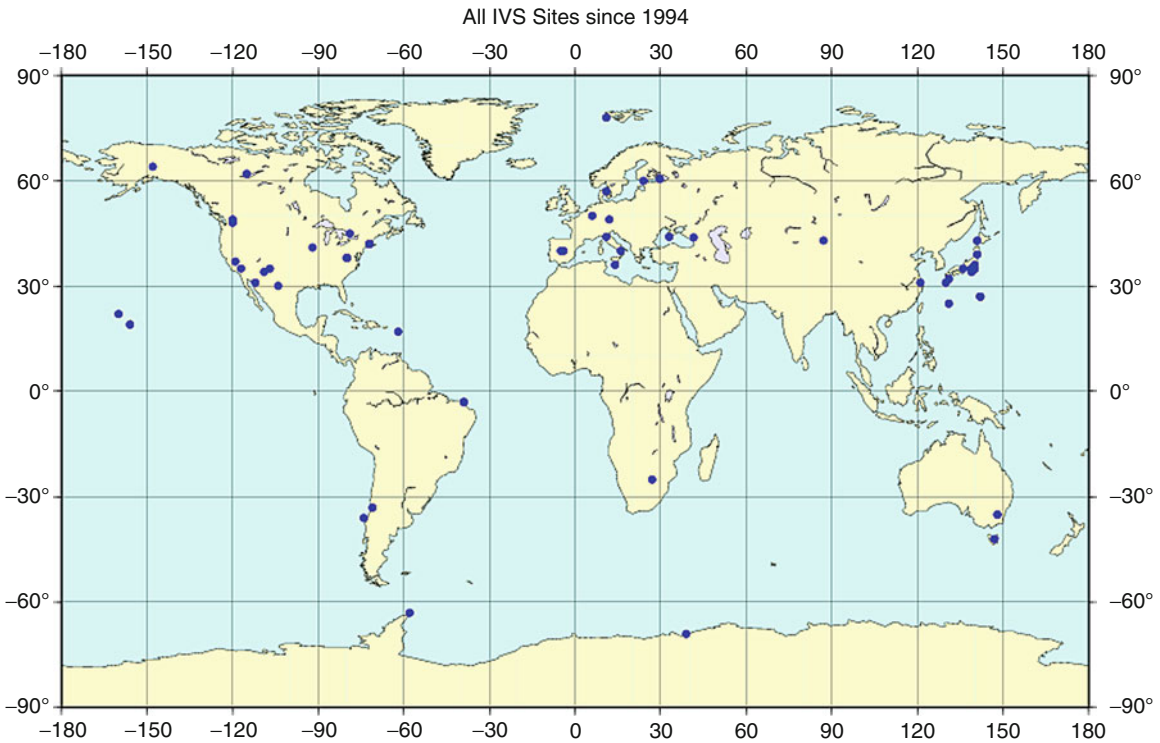
Geodetic and geophysical applications

The technique of VLBI is also used for geodetic and geophysical applications. In this context, the positions of compact extragalactic radio sources, like quasars, are considered to be known to sufficient accuracy and the relative telescope coordinates are inferred from time delay observations which reflect the baseline-source geometry. In most cases, networks of six to eight telescopes (Figure 2) are employed in series of weekly sessions organized by the IVS for the determination of Earth orientation parameters (qv *Earth Rotation*) and radio telescope coordinates. Observing between 50 and 60 radio sources in a repeated sequence of scans of 50–100 s duration over a 24-h period provides enough redundancy and geometric stability to determine telescope coordinates and Earth orientation parameters.

Reference systems and tectonics

Geodetic and geophysical applications of VLBI have to do a lot with coordinates of the radio telescopes on Earth and in space. Through sophisticated engineering and constant developments in the theoretical models, the coordinates of radio telescopes around the world can be determined with a relative accuracy of a few millimeters today. These coordinates are always referred to a very specific point of the telescope structure which lies at or close to the intersection of the axes about which the telescope is rotated to aim at the quasars.

The coordinates of geodetically used radio telescopes are determined repeatedly over the years and are fed into the database of the International Earth Rotation and



Very Long Baseline Interferometry, Figure 2 Locations of the VLBI antennas of the IVS network participating in the period of 1994 to 2009.

Reference Systems Service (IERS, qv *Earth Rotation*, see also http://www.iers.org). Together with observations of other space-geodetic techniques like the Global Navigation Satellite Systems (GNSS) or Satellite Laser Ranging (SLR, qv *Satellite Laser Ranging*), the VLBI results are merged into the International Terrestrial Reference Frame (ITRF, qv *Geodesy, Networks and Reference Systems*). The last version of the ITRF, ITRF2008, contains about 700 points on Earth which are all accurate to sub-centimeter. The importance of VLBI for this endeavor is that VLBI is the technique which is used to determine the scale of the ITRF. Here, VLBI benefits from the fact that the geodetic VLBI measurements link radio telescopes over intercontinental distances with extreme accuracy. The observations of the other techniques are then used to densify the global reference frame.

Closely linked to the coordinates of the radio telescopes are the changes of the positions which are caused by global tectonics better known as continental drift (qv *Geoid*). Geodetic VLBI was the first technique to verify Alfred Wegener's theory of continental drift through real measurements across the North Atlantic between 1979 and 1984 (Carter et al., 1985; Herring et al., 1986). Today, a full global scenario of several plates and deformation zones exists for which geodetic VLBI has helped to determine the magnitude and direction of the tectonic movements. While the distance between North America and Europe extends by about 20 mm per year, the changes in the Pacific region reach 80 mm per year (Kondo et al., 1987).

Likewise, the positions in space of the Earth and of the quasars observed by VLBI are defined in the International Celestial Reference System (ICRS). The ICRS has been realized through the International Celestial Reference Frame (ICRF, Ma et al., 1998) and its successor, the ICRF2 (Ma et al., 2009), which rely on the VLBI positions of many hundred extragalactic radio sources. Due to the large distances of the latter, the ICRF can be considered as definition of a local inertial system (or quasi-inertial system). The origin of the ICRF is the barycenter of the solar system. Systematic effects and additional random errors of the ICRF2 do not exceed 50–100 microarcseconds (Ma et al., 2009). The accuracy of the ICRF is expected to further improve in the future, thanks to the aforementioned space VLBI (Charlot, 2009).

Earth rotation

Polar motion and irregularities in the Earth's rotation: Geodetic VLBI is one of the few main techniques for the determination and monitoring of the Earth's variable rotation (qv *Earth Rotation*). Through the direct link between the terrestrial and the celestial reference frames, VLBI is the only technique which permits to determine all components of Earth rotation without any input from other techniques. While today polar motion is regularly determined precisely with GPS observations (qv *GPS, Tectonic Geodesy*) on a routine basis, VLBI uniquely contributes the Earth's phase of rotation represented through the time UT1 (Universal Time 1). UT1 is quite variable, due to movements of the atmosphere and of the oceans, and information on UT1 is needed for a number of applications including orbit transformations of Global Navigation Satellite Systems. For this purpose, the International VLBI Service for Geodesy and Astrometry (IVS) runs daily observations providing UT1 with an accuracy of about 5–10 μs.

Precession and Nutation: Looking at the Earth's axis of rotation in space, one can see that it describes a cone with an opening angle of 23.45°, which needs about 25,700 years to complete. This phenomenon is called precession. Superimposed on the precession cone are periodic movements of the Earth's axis of rotation which have a magnitude of a few arcseconds and are called nutation. For historical reasons, precession and nutation are handled independently although they are caused by the same phenomenon: The astronomical bodies of Sun, Moon, and planets exert a torque on the equatorial bulge of the Earth trying to pull it toward the ecliptic. Through its desire to conserve angular momentum, the Earth reacts with a precession effect. Through the periodic rotations of the attracting bodies, superimposed periodicities result and more than a 1,000 nutation periods can be identified covering a spectrum from 13.6 years down to a few days.

General relativistic effects in Earth Rotation: Owing to the orbital motion of the Earth, spin-orbit coupling leads to a "geodetic precession" of about 2 arcseconds per century (qv *Earth Rotation*, Klioner et al., 2010) that are measured by VLBI, which lends itself also to other tests of general relativity (Fomalont et al., 2010).

Today, geodetic VLBI is the only technique that is used to determine precession and nutation. The accuracy of the nutation offsets determined from 24 h of VLBI observations is on the order of 80–100 microarcseconds which correspond to about 3 mm when transformed to a metric displacement at the Earth's surface. The agreement between these observations and a model which is derived geophysically using an elastic Earth model is on the order of 1 milliarcsecond. The main cause of the deviations is the effect of the so-called free core nutation (FCN). This phenomenon results from the fact that the Earth's core has a slightly different ellipticity than the rest of the body resulting in different torques.

Bibliography

Bare, C., Clark, B. G., Kellermann, K. I., Cohen, M. H., and Jauncey, D. L., 1967. Interferometer Experiment with Independent Local Oscillators. *Science*, **157**, 189–191.

Carter, W. E., Robertson, D. S., and Mackay, J. R., 1985. Geodetic radio interferometric surveying: Applications and results. *Journal of Geophysical Research*, **90**, 4577–4587.

Charlot, P., 2009. Space VLBI and the radio reference frame. In Hagiwara, Y., et al. (eds.), *Approaching micro-arcsecond resolution with VSOP-2: Astrophysics and technologies*. ASP Conference Series, Vol. 402, pp. 480–485.

Fomalont, E., Kopeikin, S., Jones, D., Honma, M., and Titov, O., 2010. Recent VLBA/VERA/IVS tests of general relativity. In

Klioner, S. A., Seidelmann, P. K., and Soffel, M. H. (eds.), *Relativity in Fundamental Astronomy*. Cambridge: Cambridge University Press, pp. 291–295.
Garrett, M. A., Porcas, R. W., Pedlar, A., Muxlow, T. W. B., and Garrington, S. T., 1999. Wide-field VLBI imaging. *New Astronomy Reviews*, **43**, 519–522.
Heinkelmann, R., and Schuh, H., 2010. Very long baseline interferometry: accuracy and relativistic tests. In Klioner, S. A., Seidelmann, P. K., and Soffel, M. H. (eds.), *Relativity in Fundamental Astronomy*. Cambridge: Cambridge University Press, pp. 286–290.
Herring, T. A., Shapiro, I. I., Clark, T. A., Ma, C., and Ryan, C., 1986. Geodesy by radio interferometry: Evidence for contemporary plate motion. *Journal of Geophysical Research*, **91**, 8341–8347.
Högbom, J. A., 1974. Aperture synthesis with a non-regular distribution of interferometer baselines. *Astronomy & Astrophysics Supplement*, **15**, 417–426.
Klioner, S. A., Gerlach, E., and Soffel, M. H., 2010. Relativistic aspects of rotational motion of celestial bodies. In Klioner, S. A., Seidelmann, P. K., and Soffel, M. H. (eds.), *Relativity in Fundamental Astronomy*. Cambridge: Cambridge University Press, pp. 112–123.
Kondo, T., Heki, K., and Takahashi, Y., 1987. Pacific plate motion detected by the VLBI experiments conducted in 1984-1985. *Radio Research Laboratories Journal*, **34**, 1–14.
Ma, C., Arias, F., Bianco, G., Boboltz, D., Bolotin, S., Charlot, P., Engelhardt, G., Fey, A., Gaume, R., Gontier, A.-M., Heinkelmann, R., Jacobs, C., Kurdubov, S., Lambert, S., Malkin, Z., Nothnagel, A., Petrov, L., Skurikhina, E., Sokolova, J., Souchay, J., Sovers, O., Tesmer, V., Titov, O., Wang, G., Zharov, V., Böckmann, S., Collioud, A., Gipson, J., Gordon, D., Lytvyn, S., MacMillan, D., and Ojha, R., 2009. *The Second Realization of the International Celestial Reference Frame by Very Long Baseline Interferometry (IERS Technical Note 35)*. Frankfurt am Main, Germany: Verlag des Bundesamtes für Kartographie und Geodäsie.
Ma, C., et al., 1998. The International Celestial Reference Frame as realized by very long baseline interferometry. *Astronomical Journal*, **116**, 516–546.
Readhead, A. C. S., Walker, R. C., Pearson, T. J., and Cohen, M. H., 1980. Mapping radio sources with uncalibrated visibility data. *Nature*, **295**, 137–140.
Ryle, M., and Vonberg, D. D., 1946. Solar radiation at 175 Mc/s. *Nature*, **158**, 339–340.
Schlüter, W., and Behrend, B., 2007. The International VLBI Service for Geodesy and Astrometry (IVS): current capabilities and future prospects. *Journal of Geodesy*, **81**, 397–387.
Schwab, F. R., and Cotton, W. D., 1983. Global fringe search techniques for VLBI. *Astronomical Journal*, **88**, 688–694.
Sudarshan, E. C. G., 1963. Equivalence of semiclassical and quantum mechanical descriptions of statistical light beams. *Physical Review Letters*, **10**, 277–279.
Tornatore, V., and Charlot, P., 2007. The impact of radio source structure on European geodetic VLBI measurements. *Journal of Geodesy*, **81**, 469–478.
van Cittert, P. H., 1934. Die wahrscheinliche Schwingungsverteilung in einer von einer Lichtquelle direkt oder mittels einer Linse beleuchteten Ebene. *Physica*, **1**, 201–210.
Zernike, F., 1938. The concept of degree of coherence and its application to optical problems. *Physica*, **5**, 785–795.

Cross-references

Earth Rotation
Geodesy, Networks and Reference Systems
Geoid
GPS, Data Acquisition and Analysis
GPS, Tectonic Geodesy
SAR Interferometry
Satellite Laser Ranging

W

WAVELET ANALYSIS

Mikhail Kulesh
Information Technology, DIS AG, Mahlow, Germany

Synonyms

Time-frequency analysis; Wavelet transform; Wavelets

Definition

Time-frequency representation (TFR). A view of a time series represented over time and frequency.
Wavelet. A special function that is well localized in both time and frequency domains.
Continuous wavelet transform (CWT). TFR of the analyzed signal obtained as a set of scalar products of all dilated and translated wavelets with this signal.
Discrete wavelet transform (DWT). A wavelet transform for which the wavelet is dilated and translated discretely.

Introduction

Spectral analysis methods of the time series based on the Fourier transform are significantly limited in their ability to distinguish the waves with similar frequency content but having the different time of arrival. This limitation can be overcome using different time-frequency analysis techniques that allow us to examine how the frequency content of a given signal changes as a function of time.

A view of a time series represented over both time and frequency is called the time-frequency representation (TFR). Time-frequency analysis means analysis into a TFR. There are different TFR methods each suitable for different applications:

1. The short-time (alternatively short-term or windowed) Fourier transform that is used to determine the sinusoidal frequency and phase content of local sections of a signal as it changes over time (Allen and Mills, 2004).
2. The Gabor transform is a special case of the short-time Fourier transform that uses a Gaussian function to create a window of time from which the spectrum of the local signal values is computed (Allen and Mills, 2004).
3. The Wigner (alternatively Wigner–Ville) distribution (Wahl and Bolton, 1993) and the pseudo-Wigner–Ville distribution (Prosser et al., 1999; Andria and Savino, 1996) that can be formulated by the multiplicative comparison of a signal with itself, expanded in different directions about each point in time.
4. The S-transform (Stockwell et al., 1996; Pinnegar and Mansinha, 2003) that is similar to the short-time Fourier transform, but with a Gaussian window whose width scales inversely, and whose height scales linearly, with the frequency.
5. The continuous wavelet transform (CWT) of a real or complex signal with respect to a real or a complex analyzing wavelet that is a set of scalar products of all dilated and translated wavelets with this signal (Grossmann and Morlet, 1984; Heil and Walnut, 1989; Chui, 1992; Daubechies, 1992; Holschneider, 1995; Mallat, 1999; Antoine, 2004). A wavelet is a special function that is well localized in both time and frequency domains.
6. The discrete wavelet transform (DWT) that is a wavelet transform for which the analyzing wavelet is dilated and translated discretely (Heil and Walnut, 1989; Holschneider, 1995).

The short-time Fourier transform has a fixed resolution dependent on the windowing function width. A wide window gives better frequency resolution (frequency components close to each other can be separated), but poor time resolution. A narrower window gives good time resolution

Harsh K. Gupta (ed.), *Encyclopedia of Solid Earth Geophysics*, DOI 10.1007/978-90-481-8702-7,

(the time at which frequencies change), but poor frequency resolution. Among all possible window functions, there is one signal in particular that shows the best performance regarding time and frequency localization: the Gaussian. Thus, the Gabor transform and the S-transform are the short-time Fourier transforms with the best joint time-frequency resolution.

Since the Wigner distribution is a correlation of the signal with a time and frequency translation of itself, there is no limitation in the time-frequency resolution. However, its distribution may show negative energy levels and exhibit interferences that do not have any physical meaning. To minimize the effect of these interferences, the smoothed pseudo-Wigner–Ville distribution can be used. Unfortunately, the use of the smoothing operator introduces time-frequency smearing in the resulting TFR and therefore downgrades its advantage over the Gabor or wavelet transform, namely, the exact localization over time and frequency.

The CWT and DWT are computed by correlating the signal with families of time-frequency atoms (wavelets). Therefore, the time-frequency resolution of this method is limited by the uncertainty relation bound to the simultaneous time-frequency resolution of the analyzing wavelet.

All of these above-mentioned methods are widely used in the geophysical applications. However, the wavelet transform gives a more general framework for the time-frequency analysis of geophysical data; this approach is powerful and elegant and has several remarkable properties that can be directly applied to the polarization and dispersion analysis of seismic waves in the time-frequency domain.

In the next sections, we briefly introduce the direct and inverse wavelet transforms and some of their properties. Then, we will demonstrate the applications of the CWT in the geophysical time-series analysis.

Continuous wavelet transform

Let us introduce the following notations:

1. $s(t)$ is a square-integrable (finite-energy) real or complex function over time from $L^2(\mathbf{R})$ space: $s(t) \in L^2(\mathbf{R}) \Leftrightarrow \int_{-\infty}^{+\infty} |s(t)|^2 dt < \infty$.
2. $\omega = 2\pi f$ is the angular frequency (in radian/s) and f is the physical frequency (in Hz).
3. $\hat{s}(\omega) = \int_{-\infty}^{+\infty} s(t) e^{-i\omega t} dt, \quad s(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{s}(\omega) e^{i\omega t} d\omega$ are the direct and inverse Fourier transforms.
4. $g(t) \in L^2(\mathbf{R})$ is a real or complex function called analyzing (mother) wavelet (Grossmann and Morlet, 1984).
5. $T_t : g(\tau) \mapsto g(\tau - t)$ is a shifting operator for the analyzing wavelet.
6. $D_a : g(\tau) \mapsto g(\tau/a)/a$ is a dilation operator for the analyzing wavelet.
7. $W_g s(t, a)$ are the complex wavelet coefficients dependent on the dimensionless scaling factor $a \in \mathbf{R}$ and dimensional time factor $t \in \mathbf{R}$ (in s).
8. The asterisk superscript means the complex conjugate.

The wavelet transform $W_g s(t, a)$ of the signal $s(t)$ with respect to an analyzing wavelet $g(t)$ is the set of L^2-scalar products $\langle T_t D_a g, s \rangle$ of the signal with all dilated and translated wavelets:

$$W_g s(t, a) = \langle T_t D_a g, s \rangle = \int_{-\infty}^{+\infty} \frac{1}{a} g^*\left(\frac{\tau - t}{a}\right) s(\tau) d\tau \,. \tag{1}$$

Note that there are different possibilities to define the dilation operator D_a. We use the factor $1/a$ in this operator. For this factor, the amplitude of the wavelet spectrum is proportional to the amplitude of the signal for pure frequencies. Alternatively, the dilation $D_a : g(\tau) \mapsto g(\tau/a)/\sqrt{a}$ with corresponding adaptation of the transform (1) can be also used (Heil and Walnut, 1989), but the relation between the amplitude of the spectrum and the amplitude of the signal will be in this case frequency (scale) dependent.

Using the properties of the Fourier transform, we can also define the shifting and the dilation operators as $T_t : \hat{g}(\omega) \mapsto e^{-i\omega t} \hat{g}(\omega)$ and $D_a : \hat{g}(\omega) \mapsto \hat{g}(a\omega)$ that allows us to rewrite the transform (1) as being used for Fourier-spectra of the analyzed signal and the analyzing wavelet (Holschneider, 1995):

$$\begin{aligned} W_g s(t, a) &= \frac{1}{2\pi} \langle T_t D_a \hat{g}, \hat{s} \rangle \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{i\omega t} \hat{g}^*(a\omega) \hat{s}(\omega) d\omega. \end{aligned} \tag{2}$$

The last equation can be used for the fast computation methods to calculate the CWT of the discrete signals $s(t_n)$ with the signal length equal to 2^N points. In this way, a fast Fourier transform algorithm can be employed in practice to compute the integral in the Equation 2.

The inversion formula

The wavelet transform preserves all information of the transformed signal since there is an explicit formula that allows us to recover the signal $s(t)$ from its wavelet spectrum (Holschneider, 1995):

$$\begin{aligned} s(t) &= \mathrm{M}_h W_g s(t, a) \\ &= \frac{1}{C_{g,h}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{1}{a^2} h\left(\frac{t - \tau}{a}\right) W_g s(\tau, a) d\tau da, \end{aligned} \tag{3}$$

where $h(t) \in L^2(\mathbf{R})$ is a reconstruction wavelet used for the inverse transform; it can differ from the analyzing

wavelet that has been used to calculate $W_g s(t,a)$. The constant $C_{g,h}$ is the normalization coefficient having the same meaning as the coefficient $1/2\pi$ in the inverse Fourier transform. However, this coefficient depends on which pair of wavelets has been taken for direct and inverse transformation:

$$C_{g,h} = \int_0^{+\infty} \frac{1}{\omega}\left(\hat{g}^*(\omega)\hat{h}(\omega) + \hat{g}^*(-\omega)\hat{h}(-\omega)\right)d\omega. \quad (4)$$

Use of the physical frequency

In the practical applications, it is convenient to choose as the scaling parameter the physical frequency f instead of the dimensionless factor a: $f = f_0/a$, where f_0 is the characteristic (central) frequency of the analyzing wavelet. In this case, we can rewrite the direct and inverse transforms (1)–(3) as follows:

$$\begin{aligned} W_g s(t,f) &= \int_{-\infty}^{+\infty} \frac{f}{f_0} g^*\left(\frac{(\tau - t)f}{f_0}\right) s(\tau)d\tau \\ &= \int_{-\infty}^{+\infty} e^{2\pi i t\varphi}\hat{g}^*\left(\frac{f_0\varphi}{f}\right)\hat{s}(\varphi)d\varphi, \\ s(t) &= \mathrm{M}_h W_g s(t,f) \\ &= \frac{1}{C_{g,h}}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{1}{f_0} h\left(\frac{(t-\tau)f}{f_0}\right) W_g s(\tau,f)d\tau df. \end{aligned} \quad (5)$$

Wavelets

The analyzing wavelet, that is, wavelet used for direct wavelet transform, must satisfy the following conditions (Antoine, 2004):

1. The wavelet must be well localized both in the time and in the frequency domain.
2. The wavelet must satisfy the admissibility condition, which guaranties the invertibility of the wavelet transform: $\int_{-\infty}^{+\infty} \frac{|\hat{g}(\omega)|^2}{|\omega|}d\omega < \infty$, see also Equation 4.
3. In the most cases, the previous condition may be reduced to the requirement that the 0-th moment of the wavelet should vanish: $\int_{-\infty}^{+\infty} g(t)dt = 0$.
4. The previous property can be generalized by requiring that some consecutive moments vanish, which improves the efficiency of the wavelet at detecting singularities in the signal: $\int_{-\infty}^{+\infty} t^n g(t)dt = 0$.

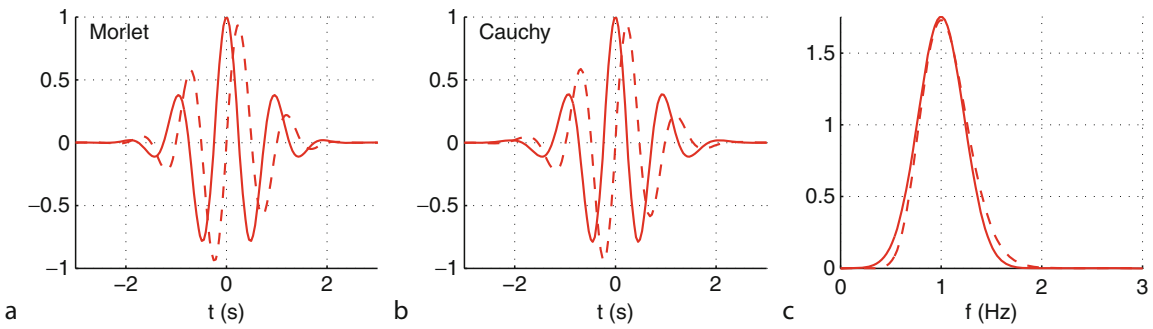
The choice of the analyzing wavelet depends on what typical oscillation modes are present in the signal. For example, to analyze rectangular oscillations, the real Haar wavelet proposed in (Haar, 1910) or a similar but symmetric French hat (FHAT) wavelet is most suitable. The most convenient for the seismic signals are real wavelets based on different-order derivatives of the Gaussian or complex progressive wavelets. If a two-component signal presented in a complex form is analyzed, or the separation into prograde and retrograde wavelet spectrum is needed, the choice of a complex progressive wavelet is preferable.

A special set of progressive wavelets is the set of functions for which the Fourier transform is real valued. Let us consider two such wavelets:

1. The Morlet wavelet is represented in Figure 1a. The solid line corresponds to the real part, and the dashed line to the imaginary part of the wavelet. The Fourier spectrum is plotted in Figure 1c as the solid line. This wavelet can be written with its Fourier transform as:

$$g(t) = e^{i\omega_0 t}e^{-t^2/(2\sigma^2)}, \hat{g}(\omega) = \sigma e^{-(\omega-\omega_0)^2\sigma^2/2}, \quad (6)$$

where the parameter σ describes the variance of the wavelet, and $\omega_0 = 2\pi f_0$ is the central frequency. Strictly speaking, this function is not a wavelet because it is not of zero mean. However, the function values at negative frequencies are small compared to the values at positive frequencies that allow us to use it as



Wavelet Analysis, Figure 1 (**a**) Morlet wavelet in the time domain, (**b**) Cauchy wavelet in the time domain, (**c**) Fourier-spectra of the Morlet wavelet (*solid line*) and Cauchy wavelet (*dashed line*).

a progressive wavelet with necessary precision in almost all practical applications.

2. The Cauchy wavelet is shown in Figure 1b and can be written as:

$$g(t) = \left(1 - \frac{2\pi i t}{p-1}\right)^{-p},$$
$$\hat{g}(\omega) = \frac{(p-1)^p}{(p-1)!}\left(\frac{\omega}{2\pi}\right)^{p-1} e^{-(p-1)\omega/2\pi}. \quad (7)$$

This wavelet has the unit central frequency, the parameter p describes the variance of the wavelet, and the Fourier spectrum is nonsymmetrical (Figure 1c, dashed line). The wavelet transform with respect to this wavelet is closely connected to the analysis of analytical functions over the half plane (Holschneider, 1995).

One can choose the translated δ-function $\delta_t = \delta(\tau - t)$ as the reconstruction wavelet $h(t)$ for the inverse wavelet transform, which gives us a rather simple and fast reconstruction formula (Holschneider, 1995):

$$s(t) = \mathrm{M}_\delta W_g s(t,f) = \frac{1}{C_{g,\delta}} \int_{-\infty}^{+\infty} \frac{1}{f} W_g s(t,f) df. \quad (8)$$

However, the δ-function is not a wavelet and its choice as the analyzing wavelet $g(t)$ for the direct CWT is not possible.

Wavelet spectrum amplitude and phase

Because the wavelet coefficients are complex, when using a complex wavelet, we can write the spectrum $W_g s(t,f)$ as a function of wavelet spectrum modulus $\left|W_g s(t,f)\right|$ and phase $\arg W_g s(t,f)$:

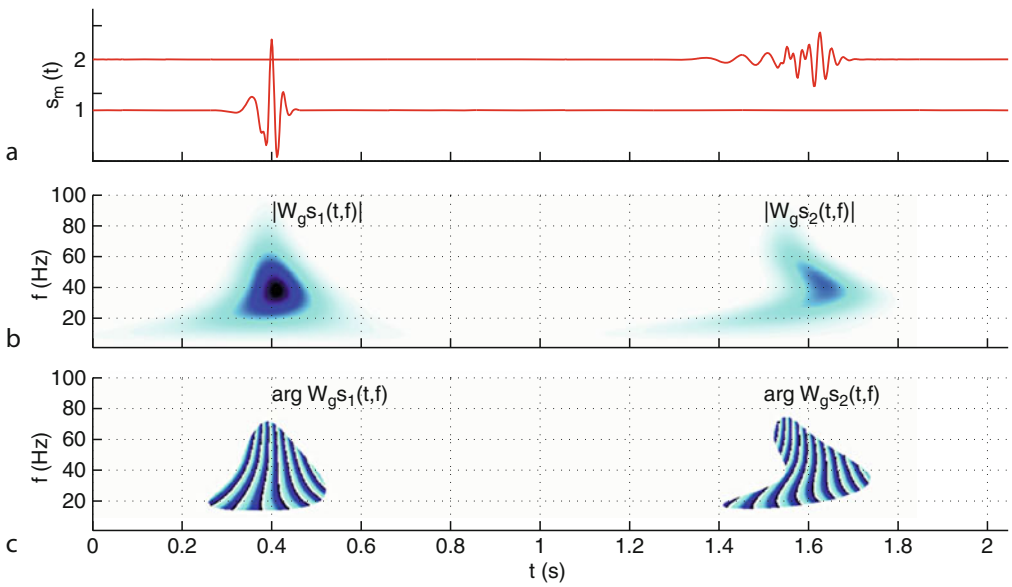
$$W_g s(t,f) = \left|W_g s(t,f)\right| e^{i \arg W_g s(t,f)},$$
$$\arg W_g s(t,f) = arctg\left(\frac{Im W_g s(t,f)}{Re W_g s(t,f)}\right), \quad (9)$$

where $arctg(x) \in [-\pi, \pi]$, which is implemented in most programming languages as atan2().

Thus, the complex wavelet spectrum contains information about the time-frequency distributions of the signal's amplitude and phase, and this amplitude and phase information can be analyzed separately as shown in the Figure 2. In this example, we consider a wave propagating in the medium with dispersion and attenuation (see *Seismic, Viscoelastic Attenuation*). This wave is represented by two signals, $s_1(t)$ and $s_2(t)$, which can be registered from two geophones (see *Seismic Instrumentation*) separated by a distance Δx (Figure 2a).

The wavelet spectrum modulus $\left|W_g s(t,f)\right|$ corresponds to the instantaneous amplitude. We can visualize it, for example, as contour line plot, where the horizontal axis corresponds to the time, and the vertical axis represents the frequency (Figure 2b). The complex wavelets Morlet and Cauchy have a phase shift between the real and imaginary components that prevent here the high-frequency modulus oscillations.

At the same time, a contour line plot of the wavelet spectrum phase $\arg W_g s(t,f)$ represents the instantaneous phase on the different frequencies. Because of practical reasons, the phases must be only plotted for time-frequency areas where the modulus $\left|W_g s(t,f)\right|$ is not infinitesimal (Figure 2c).



Wavelet Analysis, Figure 2 Wavelet spectrum amplitudes and phases for two signals.

To analyze the dynamic behavior of multivariate signals using the CWT, one may be interested in investigating of a diffeomorphic deformation of the wavelet space $W_g s_1(t,f) \mapsto W_g s_2(t,f)$ as shown in Figure 2. Such a deformation corresponds to a pseudodifferential operator acting on signal's wavelet spectrum:

$$D : s(t) \mapsto \mathrm{M}_h D W_g s(t,f). \tag{10}$$

In this context, we will consider in the next sections how the wavelet transforms of the source and the propagated signal are related through a transformation operator that explicitly incorporates the phase and group velocities as well as the attenuation factor of the medium.

Prograde and retrograde wavelet spectrum

Both Morlet and Cauchy wavelets are progressive, that is, their Fourier coefficients for negative frequencies are zero. This feature allows us to separate the wavelet spectrum into prograde $W_g^+ s(t,f)$ and retrograde $W_g^- s(t,f)$ components:

$$\begin{aligned} W_g s(t,f) &= W_g^+ s(t,f) + W_g^- s(t,f), \\ W_g^+ s(t,f) &= \begin{cases} W_g s(t,f), f \geq 0 \\ 0, f < 0 \end{cases}, \\ W_g^- s(t,f) &= \begin{cases} 0, f \geq 0 \\ W_g s(t,f), f < 0. \end{cases} \end{aligned} \tag{11}$$

To perform such a separation for nonprogressive (e.g., real) wavelets, one must first calculate the Hilbert transform of the source signal, which makes the numerical procedure more complex.

An example of the prograde and retrograde wavelet spectrum is shown in Figure 3, where we consider a two-component synthetic seismogram related to a Rayleigh wave (see *Surface Waves*). From two signals, $u_x(t)$ and $u_z(t)$ that correspond to the orthogonal components of the record, we construct a complex signal $z(t) = u_x(t) + iu_z(t)$, as shown in Figure 3a, where the solid line corresponds to the real part, and the dashed line to the imaginary part of this complex signal.

For this complex signal, we perform CWT using Equation 5 with respect to Cauchy wavelet (7) and plot only the absolute values of the complex wavelet coefficients as contour line plots for the prograde and the retrograde components (11) separately (Figure 3b). The analysis of the difference between these two spectra is exactly the idea to find out polarization properties of a signal as introduced in the next sections.

Discrete wavelet transform

As the CWT, the discrete transform also has a huge number of applications in science, engineering, mathematics, and computer science. In its basic form, the DWT employs a dyadic grid (integer power of two) in the timescale plane:

$$\begin{aligned} &T_{jk} : g(\tau) \mapsto g\left(\tau - k/2^j\right), \quad D_j : g(\tau) \mapsto 2^{j/2} g\left(2^j \tau\right), \\ &\psi_{jk}(\tau) = T_{jk} D_j g(\tau), j, k \in \mathbf{N}. \end{aligned}$$

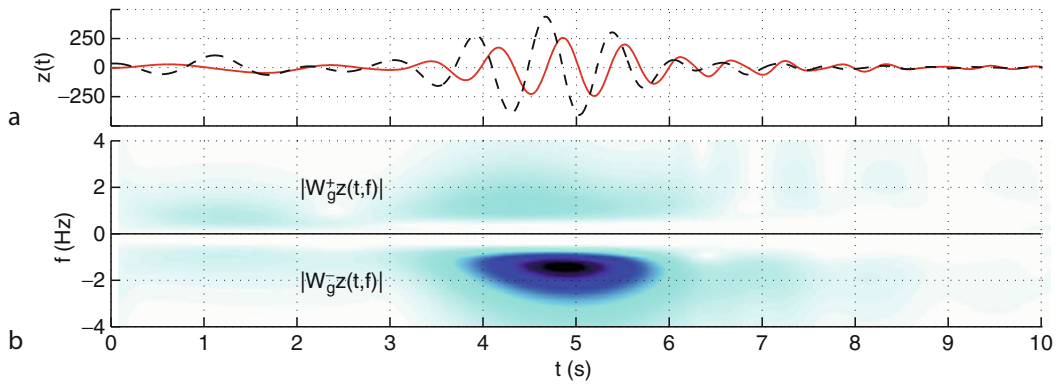
Actually, the transform integral remains continuous for the DWT, but is determined only on a discretised grid of frequencies and time points (Heil and Walnut, 1989):

$$W_g s_{jk} = \left\langle \psi_{jk}, s \right\rangle = \int_{-\infty}^{+\infty} 2^{j/2} g^*\left(2^j \tau - k\right) s(\tau) d\tau. \tag{12}$$

For the DWT, the orthonormal wavelets are usually used. The wavelet $g(t)$ is called orthonormal, when the functions family $\left\{\psi_{jk}\right\}$ in the Equation 12 corresponds to an orthonormal basis in the $L^2(\mathbf{R})$ space, that is, $\left\langle \psi_{jk}, \psi_{nm} \right\rangle = \delta_{jn}\delta_{km}$,where δ_{jn} is the Kronecker's delta. In this case, each function $s(t) \in L^2(\mathbf{R})$ can be represented as a sum of a convergent series:

$$s(t) = \sum_{j,k} W_g s_{jk} \psi_{jk}(t). \tag{13}$$

The simplest possible orthonormal wavelet is the Haar wavelet:



Wavelet Analysis, Figure 3 Prograde and retrograde wavelet transform of a Rayleigh wave arrival.

$$g(t) = \begin{cases} 1, 0 \le t < 1/2, \\ -1, 1/2 \le t < 1, \\ 0, t < 0, t \ge 1. \end{cases} \quad (14)$$

This function was proposed by Haar (1910) to give an example of a countable orthonormal system for the space of square-integrable functions on the real line. The technical disadvantage of the Haar wavelet is that it is not continuous, and therefore not differentiable.

Use of wavelets in geophysical time-series analysis

Instantaneous parameters

The concept of instantaneous parameters has been widely used in electric engineering and geophysics. Parameters such as instantaneous amplitude $|W_g s(t,f)|$, instantaneous phase $\arg W_g s(t,f)$, and instantaneous frequency $\Omega(t,f)$ are directly related to geometry and physical property variations of the medium through which the radar signal propagates. The instantaneous frequency, for example, responds to both wave propagation effects and depositional characteristics, hence it is a physical attribute and can be used as a hydrocarbon indicator by low frequency anomaly, a fracture zone indicator, or a bed thickness indicator. Using the wavelet spectrum, the instantaneous frequency is defined as follows: $\Omega(t,f) = \frac{\partial}{\partial t} \arg W_g s(t,f)$.

To avoid the numerical derivatives in the above expression, it is possible to derive an exact analytical formula for the computation of $\Omega(t,f)$ that involves the wavelet transforms of the signal with respect to the analyzing wavelet $g(t)$ and its derivative $g'(t)$. This approach is widely used for instantaneous frequency estimate with Gabor transform (Borum and Jensen, 1999) and can be adapted to the case of wavelet transform as follows:

$$\Omega(t,f) = -\mathrm{Im}\left(\frac{W_{g'} s(t,f)}{W_g s(t,f)}\right). \quad (15)$$

Polarization analysis

Perhaps the first successful attempt to combine the time-frequency analysis with the polarization analysis of multicomponent signals was done by (Keilis-Borok et al., 1989). Afterwards, a great number of the different polarization methods working in the time-frequency domain has been proposed. Some of these methods are based on:

1. The DWT with respect to the orthogonal wavelets and the singular value decomposition (Lilly and Park, 1995; Claassen, 2001).
2. The S-transform and elliptical motion approximation (Pinnegar, 2006; Schimmel and Gallart, 2007).
3. The CWT with respect to the orthogonal Mayer's wavelets and the complex covariation method (Soma et al., 2002), the singular value decomposition method (Roueff et al., 2006), the adaptive covariance method (Kulesh et al., 2007), as well as using analytical signals (Burjanek et al., 2009).
4. The wavelet packet decomposition of three-component seismic traces (Lyubushin, 2006).

As an example let us consider a simple polarization method based on the complex trace analysis. This method can be applied for the two-component signals. First, we construct a complex signal $z(t) = u_x(t) + iu_z(t)$ from two orthogonal components $u_x(t)$ and $u_z(t)$ of a record to be analyzed. Next, we perform CWT and split the obtained wavelet spectrum into prograde $W_g^+ z(t,f)$ and retrograde $W_g^- z(t,f)$ components using Equation 11. Then, we can compute the phase difference $\Delta\phi(t,f)$, the major $R(t,f)$ and minor $r(t,f)$ half-axes, and the tilt angle $\theta(t,f)$ of the polarization ellipse in the time-frequency domain as follows

$$\begin{aligned} \Delta\phi(t,f) &= \arg\left(\frac{W_g^+ z(t,f) + W_g^- z(t,f)^*}{W_g^+ z(t,f) - W_g^- z(t,f)^*}\right) + \pi/2, \\ R(t,f) &= |W_g^+ z(t,f)| + |W_g^- z(t,f)|, \\ r(t,f) &= |W_g^+ z(t,f)| - |W_g^- z(t,f)|, \\ \theta(t,f) &= \arg(W_g^+ z(t,f)\, W_g^- z(t,f))/2. \end{aligned} \quad (16)$$

Thus, the particle motions can be represented by instantaneous attributes that show distinct features for waves of different polarization characteristics. These attributes are explicitly related to the wavelet transform coefficients of the analyzed signal that can be used to perform the complete wave mode analysis (Pacor et al., 2007) and separation/filtering (Diallo et al., 2006) of two-component seismograms in the wavelet domain.

Dispersion analysis

Surface wave propagation in heterogeneous media can provide a valuable source of information about the subsurface structure and its elastic properties. For example, surface waves can be used to obtain subsurface rigidity through inversion of the shear wave velocity. The processing of experimental seismic data sets related to the surface waves is computationally expensive and requires sophisticated techniques in order to infer the physical properties and structure of the subsurface from the bulk of available information.

One of these techniques is the calculation and interpretation of the phase and group velocities from the experimental seismograms. The time-frequency representation is today commonly used for the phase and group velocities computation (Levshin et al., 1972; Abbate et al., 1995; Prosser et al., 1999; Niethammer et al., 2001; Pedersen et al., 2003). Most of the methods are essentially based on the study of the square modulus of the representation, which can be interpreted as the energy density in the time-frequency plane. These methods provide only partial information that cannot be directly used to characterize the phase velocities and the different propagation modes.

In order to analyze the dynamical behavior of multi-arrived signals using the model (10), it is also possible to develop a wavelet transformation operator that approximates the wave propagation in the dispersive medium and explicitly incorporates the phase and group velocities as well as the attenuation function (Kulesh et al., 2005):

$$W_g s_m(t,f) = e^{-\alpha(f)D_{mj}} \left| W_g s_j\left(t - \frac{D_{mj}}{C_g(f)}, f\right)\right| \times e^{i\cdot \arg W_g s_j\left(t - D_{mj}/C_p(f), f\right)}. \quad (17)$$

In this operator, $C_p(f) = f/k(f)$ is the phase velocity, $C_g(f) = 1/k'(f)$ is the group velocity, and $\alpha(f)$ is the attenuation function. The signals $s_m(t)$ and $s_j(t)$ correspond to records obtained from two geophones placed with distance $D_{mj} = D_m - D_j$ apart, and $k(f)$ is the frequency-dependent wave-number function.

The relationship (17) has the following interpretation. The group velocity is a function that "deforms" the image of the absolute value of the source signal's wavelet spectrum, the phase velocity "deforms" the image of the wavelet spectrum phase, and the attenuation function determines the frequency-dependent real coefficient by which the spectrum is multiplied (see also Figure 2b and c). Based on this property, one can construct the optimization (Holschneider et al., 2005) and correlation (Kulesh et al., 2008; Waltisberg and Raišutis, 2008) methods that allow us to extract both phase and group velocities from the wavelet spectra of multi-arrived signals containing waves with several wave modes.

Summary

In contrast to the Fourier transform, the CWT provides a two-dimensional spectrum of the analyzed signal, where the frequency and the time are treated as independent variables. As a result, we can investigate the signal in the time domain and in the frequency domain at the same time. The fact that the time-frequency-dependent polarization attributes are explicitly related to the wavelet transform of the signals enables us to perform the wave mode separation in the wavelet domain. It is also possible to construct the wavelet-based dispersion operators that can be used to model wave propagation in a dispersive and attenuating medium.

Bibliography

Abbate, A., Frankel, J., and Das, P., 1995. Wavelet transform signal processing for dispersion analysis of ultrasonic signals. *IEEE Ultrasonic Symposium*, **1**, 751–755.

Allen, R. L., and Mills, D. W., 2004. *Signal Analysis: Time, Frequency, Scale, and Structure*. New Jersey: IEEE Press (Wiley-Interscience).

Andria, G., and Savino, M., 1996. Interpolated smoothed pseudo Wigner-Ville distribution for accurate spectrum analysis. *IEEE Transactions on Instrumentation and Measurement*, **45**(4), 818–823.

Antoine, J.-P., 2004. Wavelet analysis: a new tool in physics. In van den Berg, J. C. (ed.), *Wavelets in Physics*. Wageningen: Agricultural University.

Borum, S., and Jensen, K., 1999. Additive analysis/synthesis using analytically derived windows. In *Proceedings of the 2nd COST G-6 Workshop on Digital Audio Effects (DAFx99)*. NTNU, Trondheim, December 9–11, 1999, pp. 125–128.

Burjanek, J., Gassner-Stamm, G., Poggi, V., Moore, J. R., and Faeh, D., 2009. Ambient vibration analysis of an unstable mountain slope. *Geophysical Journal International*, **180**(2), 820–828.

Chui, C. K., 1992. *An Introduction to Wavelets*. San Diego: Academic.

Claassen, J., 2001. Robust bearing estimation for three-component stations. *Pure and Applied Geophysics*, **158**(1), 349–374.

Daubechies, I., 1992. *Ten Lectures on Wavelets*. Philadelphia: Society for Industrial and Applied Mathematics.

Diallo, M. S., Kulesh, M., Holschneider, M., Scherbaum, F., and Adler, F., 2006. Characterization of polarization attributes of seismic waves using continuous wavelet transforms. *Geophysics*, **71**(3), V67–V77.

Grossmann, A., and Morlet, J., 1984. Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM Journal on Mathematical Analysis*, **15**(4), 723–736.

Haar, A., 1910. Zur Theorie der orthogonalen Funktionensysteme (German). *Mathematische Annalen*, **69**, 331–371.

Heil, C., and Walnut, D. F., 1989. Continuous and discrete wavelet transform. *SIAM Rev*, **31**(4), 628–666.

Holschneider, M., 1995. *Wavelets: an Analysis Tool*. Oxford: Oxford University Press.

Holschneider, M., Diallo, M. S., Kulesh, M., Ohrnberger, M., Lück, E., and Scherbaum, F., 2005. Characterization of dispersive surface waves using continuous wavelet transforms. *Geophysical Journal International*, **163**(2), 463–478.

Keilis-Borok, V. I., Levshin, A. L., and Yanovskaya, T., 1989. *Seismic Surface Waves in a Laterally Inhomogeneous Earth*. Boston: Kluwer Academic.

Kulesh, M., Holschneider, M., Diallo, M. S., Xie, Q., and Scherbaum, F., 2005. Modeling of wave dispersion using continuous wavelet transforms. *Pure and Applied Geophysics*, **162**(5), 843–855.

Kulesh, M., Diallo, M. S., Holschneider, M., Kurennaya, K., Krüger, F., Ohrnberger, M., and Scherbaum, F., 2007. Polarization analysis in the wavelet domain based on the adaptive covariance method. *Geophysical Journal International*, **170**(2), 667–678.

Kulesh, M., Holschneider, M., Ohrnberger, M., and Lück, E., 2008. Modeling of wave dispersion using continuous wavelet transforms II: Wavelet based frequency-velocity analysis. *Pure and Applied Geophysics*, **165**(2), 255–270.

Levshin, A. L., Pisarenko, V. F., and Pogrebinsky, G. A., 1972. On a frequency-time analysis of oscillations. *Annales Geophysicae*, **28**, 211–218.

Lilly, J. M., and Park, J., 1995. Multiwavelet spectral and polarization analyses of seismic records. *Geophysical Journal International*, **122**(3), 1001–1021.

Lyubushin, A. A., 2006. Wavelet packet polarization method for an automatic detection of P and S wave arrivals. *Izvestiya Physics of the Solid Earth*, **42**(4), 30–39.

Mallat, S., 1999. *A Wavelets Tour of Signal Processing*, 2nd edn. London: Academic.

Niethammer, M., Jacobs, L. J., Qu, J., and Jarzynski, J., 2001. Time-frequency representations of Lamb waves. *The Journal of the Acoustical Society of America*, **109**(5), 1841–1847.

Pacor, F., Bindi, D., Luzi, L., Parolai, S., Marzorati, S., and Monachesi, G., 2007. Characteristics of strong ground motion data recorded in the Gubbio sedimentary basin (Central Italy). *Bulletin of Earthquake Engineering*, **5**(1), 27–43.

Pedersen, H. A., Mars, J. I., and Amblard, P.-O., 2003. Improving surface-wave group velocity measurements by energy reassignment. *Geophysics*, **68**(2), 677–684.
Pinnegar, C. R., 2006. Polarization analysis and polarization filtering of three-component signals with the time-frequency S-transform. *Geophysical Journal International*, **165**(2), 596–606.
Pinnegar, C. R., and Mansinha, L., 2003. The S-transform with windows of arbitrary and varying shape. *Geophysics*, **68**(1), 381–385.
Prosser, W. H., Seale, M. D., and Smith, B. T., 1999. Time-frequency analysis of the dispersion of Lamb modes. *The Journal of the Acoustical Society of America*, **105**(5), 2669–2676.
Roueff, A., Chanussot, J., and Mars, J. I., 2006. Estimation of polarization parameters using time-frequency representations and its application to waves separation. *Signal Processing*, **86**(12), 3714–3731.
Schimmel, M., and Gallart, J., 2007. Frequency-dependent phase coherence for noise suppression in seismic array data. *Journal of Geophysical Research*, **112**, B04303.
Soma, N., Niitsuma, H., and Baria, R., 2002. Reflection technique in time-frequency domain using multicomponent acoustic emission signals and application to geothermal reservoirs. *Geophysics*, **67**(3), 928–938.
Stockwell, R. G., Mansinha, L., and Lowe, R. P., 1996. Localization of the complex spectrum: the S-transform. *IEEE Transactions on Signal Processing*, **44**(4), 998–1001.
Wahl, T. J., and Bolton, J. S., 1993. The application of the Wigner distribution to the identification of structure-borne noise components. *Journal of Sound and Vibration*, **163**, 101–122.
Waltisberg, D., and Raišutis, R., 2008. Group velocity estimation of Lamb waves based on the wavelet transform. *Ultragarsas* (*Ultrasound*), **63**, 35–50.

Cross-references

Seismic Instrumentation
Seismic, Viscoelastic Attenuation
Surface Waves

Author Index

Subject Index